1-1-2010

Evolutionary Dynamics in Molecular Populations of Ligase Ribozymes

Carolina Diaz Arenas
Portland State University

Evolutionary Dynamics in Molecular Populations of Ligase Ribozymes

by

Carolina Diaz Arenas

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

Doctor of Philosophy
in
Biology

Dissertation Committee:
Niles Lehman, Chair
Susan Masta
Suzanne Estes
Kenneth Stedman
James McNames

Portland State University

Abstract

The emergence of life depended on the ability of the first biopolymer populations to thrive and approach larger population sizes and longer sequences that could store enough information, as required for a cellular type of life. The evolution of these populations very likely occurred under circumstances under which Muller's Ratchet in synergism with random drift could have caused large genetic deterioration of the biopolymers. The genetic deterioration of the molecules caused by the accumulation of mutations occurred during the copying process, can drive the populations to extinction unless there is a mechanism to counteract it. To test the effect of the mutation rate and the effective population size on the time to extinction, we used clonal populations of B16-19 ligase ribozymes, evolved with the continuous evolution *in vitro* system. The experiments were done using populations of 100, 300, 600 and/or 3000 molecules, and at low and high mutation rates. The error-prone Moloney Murine Leukemia virus reverse transcriptase was used with and without the addition of Mn(II). Populations evolved without Mn(II) were of four effective sizes. The times to extinction for those populations were found to be directly related to the effective size of the population. The small populations approached extinction at an average of 24.3 cycles; while the large populations did so at an average of 44.5 cycles. Genotypic characterization of the populations showed the presence of deleterious mutations in the small populations, which are the likely cause of their genetic deterioration and

extinction via mutational meltdown. These deleterious mutations were not observed in the large populations; in contrast an advantageous mutant was present. Populations of 100 and 3000 molecules were evolved with Mn(II). None of the populations showed signs of genetic deterioration nor did they become extinct. Genotypic characterization of the 100-molecule population indicated the presence of a cloud of mutants forming a “quasispecies” structure. The high error rate used generated an extended class of closely genetically-related mutants, as indicated by their Hamming distance. The close connectedness of the mutants facilitates the recovery of one from another in the event of being removed from the population by random genetic drift. Thus, quasispecies shift the target of selection from the individual to the group and through cooperative behavior the populations stay extant. The fitness of the six most abundant molecules evolved was measured. The total fitness of the molecules was measured by identifying the fitness component of the system that affect the ligase replication cycles: the ligation, the reverse transcription and the transcription reactions. It was found that the strength of the three components of fitness varied in different chemical environments, and each has a differential effect in the total absolute fitness of the ligases. The ligase molecules evolved have different total absolute fitness values, and ranged above and below the fitness of B16-19.

Dedication

To my family

To you, for being an active searcher with the effort it takes to be educated in a world of growing distractions and noise.

"Ignorance is the root and stem of every evil"
Plato

"Education is the most powerful weapon which you can use to change the world"
N. Mandela

Acknowledgments

Dr. Niles Lehman

Dr. Kenneth Stedman

Dr. Susan Masta

Dr. Suzanne Estes

Dr. James Mc Names

Lehman laboratory members: Aaron Burton, David Gofreed, Orin Holland, Brian Larson, Nilesh Vaidya, and former members Paul Cernak, Eric Hayden, and Steven Soll.

Biology and Chemistry peers in Dr. Iwata-Reuyl and Dr. Stedman laboratories

The Chemistry and Biology Departments at Portland State University

The International Student Support and the Human Resources Offices

Family and Friends, here and there

Scientists who have inspired me with their great work

You

Me

Table of Content

List of Tables

List of Figures

CHAPTER ONE

Introduction

(Adapted from Díaz Arenas and Lehman. 2009. *Int. J. Biochem. Cell Biol*, 41:266-273)

How was life created on Earth and how did it evolve in ancient times? These are questions that one cannot pretend to solve in a single study, but rather by the contribution of many particular studies focused on different aspects of the topic. Although life's tape cannot be replayed (Gould, 1989), research can be done to follow what seems to have plausibly happened at the origins of life.

One important theory about the origins of life is that of the RNA World (Gilbert, 1986). This theory refers to a hypothetical timeframe in early life evolution in which the genes were naked RNA replicating molecules. The information was stored and transferred from RNA molecules to other RNAs without the aid of any protein molecule. Two pieces of evidence provide strong support to this hypothetical primordial world: (1) Catalytic RNA molecules (ribozymes) naturally exist and catalyze a variety of reactions *in vivo* and *in vitro* (e.g., Tarasow, *et al*., 1997; Zhang and Cech, 1997; Shabarova and Bogdanov, 1994) and thus, they may be molecular fossils of the predominant catalytic activity of early life (Watson, *et al*., 1987; Gilbert, 1986; Gesteland, *et al*., 2006); (2) the flow of information from RNA is bidirectional, going to proteins

as well as to DNA by means of reverse transcriptase and ribosome enzymes (Baltimore, 1970; Temin, 1970).

Now, if one puts oneself in the RNA World timeframe, it is instructive to ask how those molecules could have ensured their survival through time. At that time, populations of biopolymers were probably small and the molecules were of short length. These two characteristics have strong implications in the accumulation of mutations and furthermore in the survival of the populations, if the replication fidelity is low. Thus, the emergence of life required the first information-bearing biopolymers to approach larger population sizes and/or longer sequences that allow information storage for more sophisticated functions (e.g., enhanced catalytic activity and/or more efficient folding) to thrive and evolve.

In order to understand how small populations of RNA molecules could have survived and evolved under mutational pressure, one can use an *in vitro* evolution technique (Joyce, 2007). The evolutionary processes, as imagined by Darwin 150 years ago, are evident not only in the wild but also in the test tube; thus a variety of evolutionary dynamics can be observed in molecular populations in relatively short periods of laboratory experimentation (Díaz Arenas and Lehman, 2009).

The Continuous Evolution *in vitro* (CE) technique, developed by Wright and Joyce (1997) allow us to model evolutionary phenomena because it allows a population of molecules to evolve continuously following the dynamics dictated by the mere interaction among the individual molecules in a relatively constant environment (Schmitt and Lehman, 1999; McGineness, *et al*., 2002; Johns and Joyce, 2005; Voytek and Joyce, 2007; Joyce, 2007; Paegel and Joyce, 2008).

We use the CE system with catalytic RNA molecules (ribozymes) to model evolutionary phenomena for two reasons: (1) these molecules have both an evolvable genotype and a distinct phenotype. This characteristic allows the ribozymes to behave as 'organisms' in term of selective and evolutionary forces (Joyce, 1989; Lehman, *et al*., 2000; Langhammer, 2003; Kun, *et al*., 2005). The genotype of ribozymes (Figure 1.1A) is constituted by a sequence of nucleotides and the phenotype (Figure 1.1B) is constituted by the catalytic function encoded in the sequence (Cech, 1987), and (2) the sequences of the ribozyme used is short, in our case about 150 nucleotides; thus complete sequence analysis of their "genomes" is possible. This allows a meaningful search in the "genomes" of these molecules for the causes that may underlie the observed phenotypic changes during the course of the evolution of the ribozymes.

The CE system (Figure 1.2) consists of a series of catalytic events and selective amplification cycles. A cycle is initiated by the ligation reaction, during which the ribozyme catalyses the reaction initiated by attack of the 3'-end hydroxyl group of a *trans* substrate onto the α–phosphate of the 5'-end of the ribozyme itself, with concomitant formation of the phosphodiester bond (Figure 1.1B). A key characteristic of the system is that the *trans* substrate carries the promoter sequence for the later transcription to regenerate RNA from DNA. Therefore, ribozymes that are unable to catalyze the reaction will not be ligated to the substrate and consequently later they will not be reproduced by action of the RNA polymerase.

The second step in the cycle is the reverse transcription of the ribozymes. Because the reverse transcriptase (RT) is already present in the reaction vessel, DNA copies of reacted and unreacted ribozymes are made. RT is an error-prone enzyme, and thus mutations are likely introduced at this step of the cycle. The third step of the cycle, the transcription of RNA from this DNA, is initiated once the cDNA has been produced. The T7 RNA polymerase recognizes the promoter sequence located in the substrate-ligase cDNA complex and transcribes it. At this step, the effect of selection can be seen because only catalytically active ribozymes can pass to the next generation to undergo subsequent rounds of amplification. This selection event implies that if the number of non-reactive ribozymes increases as a result of the

accumulation of mutations, the population can experience a reduction in size and become at risk of extinction, as tested during this study (Chapter 2).

Each completion of the cycle is a “generation”, and leads to an approximately 10-fold amplification of the “fittest” RNA molecules. Each generation is accomplished in about 10 minutes or less, so in principle, hundreds of generations can be completed in a single day. When the raw materials (such as nucleotides and protein enzymes) have been exhausted, typically in about three generations, a small fraction of the RNA population can be transferred to a new test tube with fresh reagents for another set of generations. We call each transfer a "burst" because it results in a burst of RNA amplification on the order of 1000-fold (Wright and Joyce, 1997; Schmitt and Lehman, 1999). In practice, it is easy to run several lineages in parallel, either in absolute replicate or with variation of single experimental variables (Johns and Joyce, 2005).

The goal of this study is to understand the interplay of the mutational rate of the replication process and of the effective population size of the RNA populations in the time to extinction. To accomplish this, clonal populations of ligase ribozyme B16-19, were evolved with the CE technique. B16-19 is an artificial ribozyme with a high catalytic rate (Schmitt and Lehman, 1999). Therefore mutations that accumulate in its sequence tend to have a

deleterious impact on its fitness and in the population, unless there is a mechanism to mitigate them.

This document is divided into chapters that describe the distinctive evolutionary phenomena manifested in the evolved populations. In chapter 2, we study the effect of the effective population (N_e) size on the time to extinction, due to accumulation of mutations. We found that mutations accumulated in the ligase structure generate a mutational load that leads the small population to extinction. Populations of different sizes (100, 300, 600, and 3000 molecules) were tested, and as a consequence of a lack of a mechanism that counteracted the mutational load, the small populations (100 and 300 molecules) experienced a Muller's Ratchet phenomena, which in synergism with random genetic drift drove them to extinction via mutational meltdown. The time to extinction was found to be directly proportional to the N_e.

In chapter 3, we studied the effect of a high mutation rate on the time to extinction of the populations. We used populations of 100 molecules and 3000 molecules. None of the populations went extinct. To find the reason for the extended time to extinction observed in the populations of 100 molecules, we did an extensive genetic characterization of these populations. We found a quasispecies structure in all of the bursts inspected from two lineages. The

quasispecies is a population structure of closely connected mutants, which can easily regenerate one another in the event of one being removed from the population by random drift. Their close connection in genotypic landscape implies (given the secondary structure properties of RNA) that these mutants have relative similar fitness values. Thus, no one in particular is essential for the survival of the population. A Muller's ratchet phenomena cannot be accommodated by this population structure.

In chapter 4, we studied the fitness of the ligases evolved during the evolution experiments. We selected the five most abundant mutants and B16-19 (the "wildtype"). We studied the fitness components of the CE cycles and calculated total absolute and total relative fitness values for each ligase. We found that the incidence of each component of fitness in the total absolute fitness of each ligase is different, and that the incidence of a fitness component on the total fitness of the ligases varies under different chemical environments.

This study is an important contribution to the understanding of the mechanisms by which RNA populations can become extinct or develop a mechanism that allows them to escape extinction. Most of the work has been published in peer-review journals (Soll, *et al*., 2007; Díaz Arenas and Lehman, 2009; Díaz Arenas and Lehman, 2010a and 2010b).

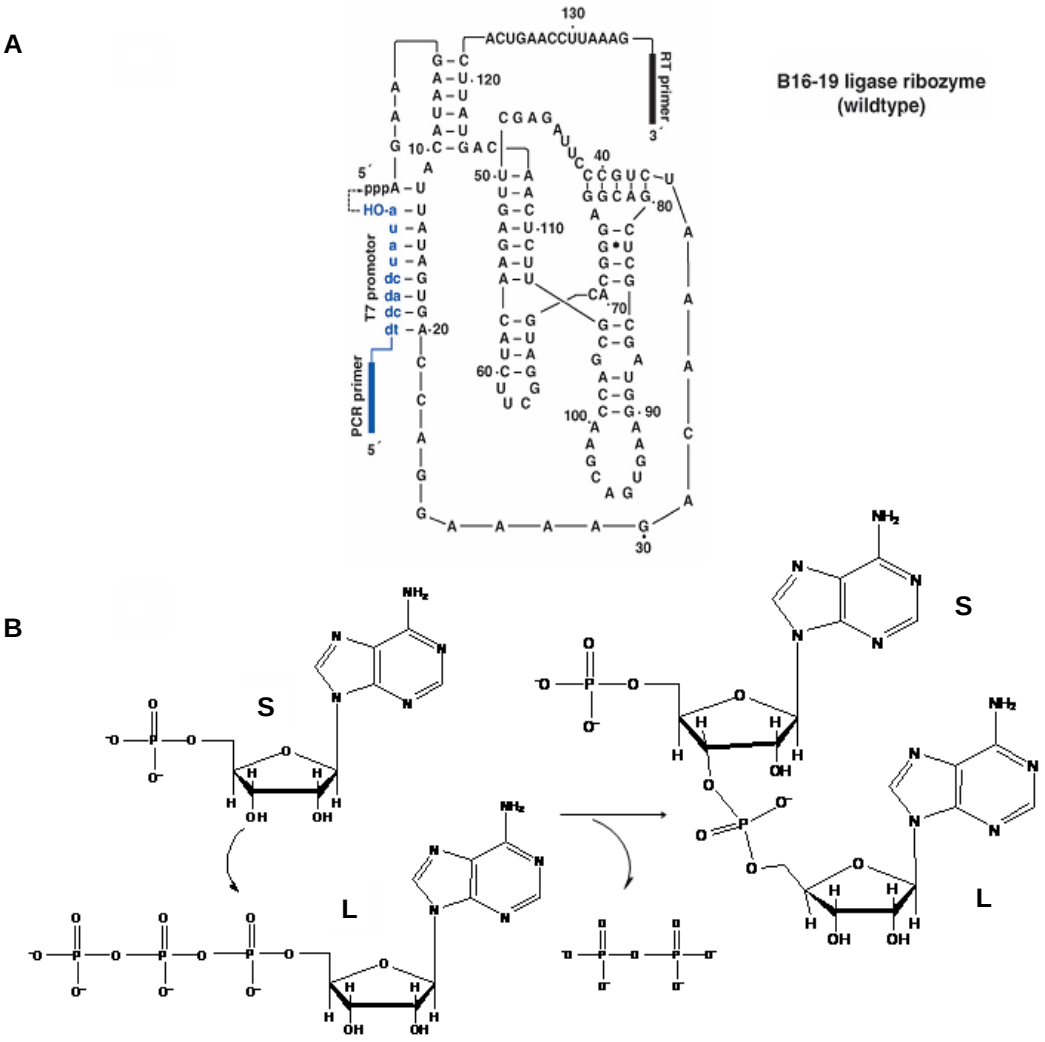


Figure 1.1. Secondary structure of B16-19 ligase ribozyme and ligation reaction. (A) The secondary structure (black letters) of the "wildtype" ligase B16-19 (Schmitt and Lehman, 1999), showing the *trans* substrate (blue, lowercase letter) and the reaction site between the substrate and the ligase (dashed arrow). (B) Detail of the chemical reaction catalyzed by the ligase ribozyme, showing attack of the 3'-OH of the substrate onto the α-phosphate of the 5'-end of the ligase. The reaction generates a phosphodiester bond and a release of pyrophosphate. L stands for ligase and S stands for substrate. The chemical reaction was drawn with Ultra ChemDraw v. 10 (2005). The structure of the ligase was taken from Soll, *et al*., (2007).

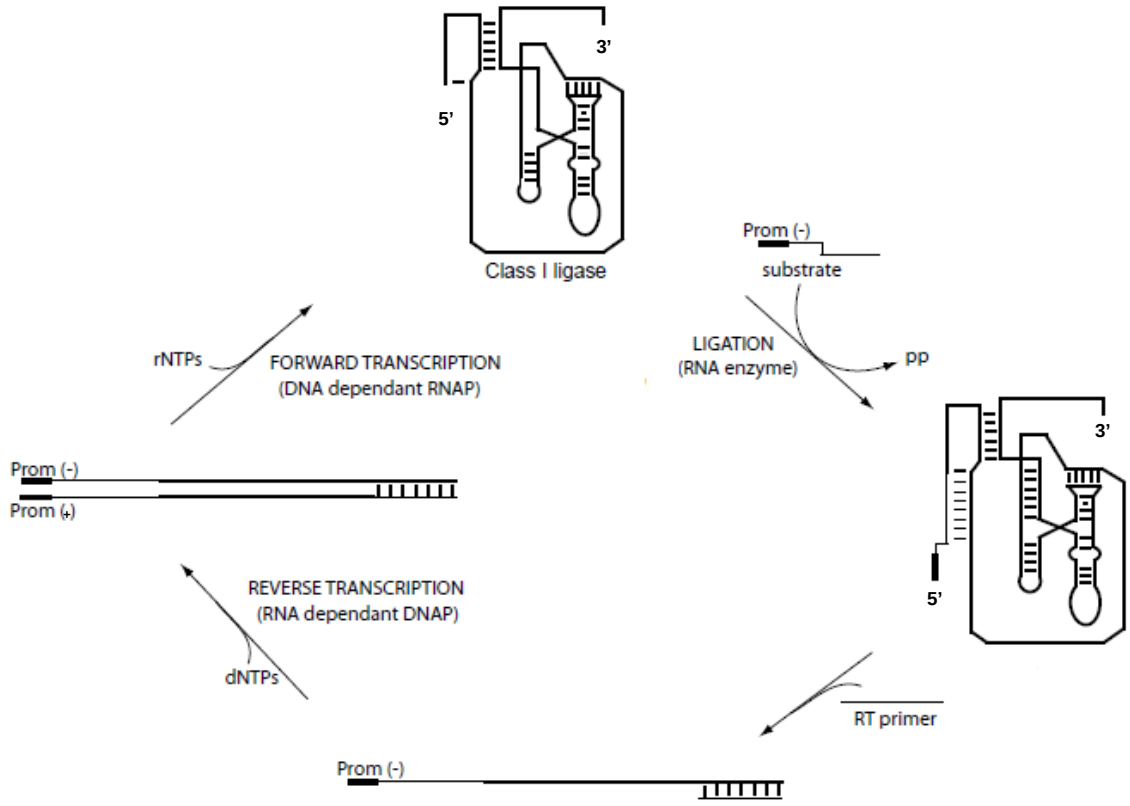


Figure 1.2. The Continuous Evolution *in vitro* (CE) model. The CE cycles start with a ligase ribozyme (cartoon at the top) that ligates an exogenous substrate to its 5'-end. The substrate carries the promoter sequence for T7 RNA polymerase. Upon binding of the RT primer to the ligase, the RNA dependant DNAP make cDNA copies of all the ligases (reacted and unreacted ones). This is an error prone enzyme; therefore mutations are likely introduced during this step of the evolution cycle. Once DNA copies have been made, the RNA polymerase transcribes new ligases. Note that only active reacted ligases are recognized by the T7 RNA polymerase, and are selected to undergo further amplification process.

CHAPTER TWO

Accumulation of Deleterious Mutations in Small Abiotic Populations of RNA

(Soll, Arenas, Lehman. 2007. *Genetics*, 175:267-275)

Background

As early as 1937, it was noted by J. B. S. Haldane that mutations with a negative effect on the average fitness of individuals could accumulate in a population (Haldane, 1937) leading to what was later called a mutational load by Muller (1950). This prediction has been borne out empirically and experimentally, in a wide variety of wild and laboratory organisms (Wallace, 1987; Lynch, *et al*., 1999). In fact, it has led to a vigorous debate over the origins and advantages of sexual reproduction. The argument is often made that sexuality provides an escape from Muller's ratchet because even occasional blending of genotypes can produce offspring with a lowered mutational load—an option not available to strictly asexual lineages. Another issue of great interest is the relationship between mutational load and population size. It has been predicted that these two factors can act synergistically, in that as the load increases, the population size should decrease, leading to a higher probability of fixing new deleterious mutations (Lynch and Gabriel, 1990; Gabriel, *et al*., 1993; Lynch, *et al*., 1993). Eventually a threshold is crossed, and the population spirals into extinction via

a "mutational meltdown", as can be seen in ciliated protozoans for example (Smith and Pereira-Smith, 1977; Tagaki and Yoshida, 1980).

One of the goals of the current study was to achieve the first mutation accumulation (MA) experiment with evolving populations of RNA molecules, with the advantage that a detailed genotypic characterization would be within reach. Another goal was to use such a study to observe both the accumulation of slightly deleterious mutations and the mutational meltdown in a very simple and tightly controlled genetic system that was essentially free of confounding factors such as pleiotropy. By using an abiotic milieu of catalytic RNAs (ribozymes) evolving *in vitro*, we endeavored to test the hypotheses that (i) the mutational load has a clear biochemical origin, and (ii) that smaller asexual populations are at a greater risk for mutational meltdown than larger ones. At the same time we would be able to examine the influence of the accumulation of deleterious mutations during the origins of life on Earth, another subject on which Haldane provided pioneering insight (Haldane, 1929). Our intent to track RNA genotypes and phenotypes over time is directly relevant to the RNA world hypothesis, in that life may very well have passed through an RNA stage en route to its current DNA/protein-based existence (Gilbert, 1986; Gesteland, *et al*., 2005).

We employed the continuous evolution *in vitro* (CE) system (Wright and Joyce, 1997) with ligase ribozymes as a means of observing mutational load in RNA populations. In this test-tube setting, ribozymes are challenged to perform a catalytic ligation on an exogenous RNA substrate, and only those that succeed can be replicated by the sequential action of two protein enzymes: reverse transcriptase and RNA polymerase. This system has highly beneficial features for MA experiments. It starts with a molecule that is catalytically proficient: the B16-19 ligase ribozyme, which has been described to sit atop of a fitness landscape, thus mutations are more like to have a deleterious effect on fitness, and their accumulation can eventually drive the molecule into a fitness valley (Lehman, 2004). Also, the populations under study are repeatedly subjected to bottlenecking, and because the size is kept under strict experimental control, variations are mainly due to evolutionary phenomena such as Muller's Ratchet (Muller, 1950) and eventual mutational meltdown (Lynch, *et al*., 1993). In the experiments described here, we chose bottleneck population sizes (Table 2.1) ranging from 100 molecules (167 ymol) to 3000 molecules (5 zmol).

Mutations in the CE system are generated by the protein enzymes. The MMLV reverse transcriptase used here is particularly error prone *in vitro*, with mutation rates estimated at 2×10^{-5} mutations/nucleotide/replication pass (De Angioletti, *et al*., 2002). Strictly on the basis of this rate, on average ~1% of the RNA molecules in each lineage would be expected to suffer a mutation each

“burst”. The T7 RNA polymerase may also contribute to the net mutation rate. Note that although the CE system uses contemporary protein enzymes to accomplish replication—enzymes that would not have been available in a prebiological RNA world (Gilbert, 1986)—these enzymes serve as convenient surrogates for RNA replicase ribozymes postulated to have been a crucial feature of the origins of life, despite having far higher mutation rates than modern protein polymerases (Johnston, *et al*., 2001). The combined use of these enzymes, a 22-min burst time with three generations per burst, and parallel treatments of lineages, meant that we could accomplish 25 MA lineages of 50–150 generations each with a strong mutational pressure in a few weeks' time.

Materials and Methods

RNA preparation:

The starting B16-19 RNA was obtained by run-off transcription of PCR DNA obtained from a cloned genotype arising in a previous *in vitro* evolution experiment (Schmitt and Lehman, 1999) and was gel purified to length homogeneity (152 nucleotides) prior to use. The concentration was measured by UV spectrometry at 260 nm and carefully diluted by a serial dilution from 10.0-μM stocks into several separate aliquots of 100 molecules/8.20 μL (2.03×10^{-8} nM), 300 molecules/8.20 μL (6.11×10^{-8} nM), 600 molecules/8.20 μL (1.22×10^{-7} nM), or 3000 molecules/8.20 μL (6.11×10^{-7} nM).

Continuous evolution *in vitro*:

The CE protocol was followed essentially as described previously (Wright and Joyce, 1997; Schmitt and Lehman, 1999; Lehman, 2004) except that vastly smaller input RNA population sizes were used. Briefly, 8.2 µL of a diluted RNA stock was incubated with 64 pmol S-163 DNA/RNA substrate (5'-CTTGACGTCAGCCTGGAC**TAATACGACTCAC<u>UAUA</u>**-3', with the T7 promoter sequence in bold and the ribonucleotides underlined), 50 pmol RT primer (5'-GCTGAGCCTGCGATTGG-3'), 250 units M-MLV reverse transcriptase (United States Biochemicals,Cleveland), 50 units T7 RNA polymerase (Ambion, Austin, TX), 5 nmol each dNTP, 50 nmol each rNTP, and 25 mM $MgCl_2$ in reaction buffer [50 mM KCl, 30 mM 4-(2-hydroxyethyl)piperazine-1-propanesulfonic acid (EPPS), pH 8.3] in a total volume of 25 µL for 22 min at 37°C. At the end of the incubation period, 3 µL were removed and diluted into 981 µL of water. An 8.2-µL aliquot of this dilution was used to seed the next 25-µL burst, resulting in an overall 1000-fold dilution from one burst to the next. In the second and all subsequent bursts, the diluted mixture from the previous burst was incubated with fresh amounts of substrate, primer, protein enzymes, nucleotides, and buffer in the quantities described above. To ensure that the dilution factor was matching the amplification factor for each burst, some lineages were run as above but additionally in the presence of 3.75 µCi [α-^{32}P] ATP. In these cases, an

additional 3 µL were removed after 22 min, quenched in acrylamide gel-loading buffer (0.05% bromophenol blue, 40% sucrose), and subjected to electrophoresis through 6% polyacrylamide/8 M urea gels and phosphorimaging. After overnight exposure to the phosphor screen, failure to detect the appearance of a 152-nt RNA species after >10 bursts, despite the appearance of strong 187-bp PCR products (see below), was indicative that the RNA population was not growing because the net dilution over this time would be as high as 1030-fold (Wright and Joyce, 1997; Johns and Joyce, 2005).

Genotypic monitoring:

A total of 25 lineages were maintained, 6 each of 100, 300, and 3000 molecules, and 7 of 600 molecules. The status of each lineage was monitored by amplification of 2.75 µL of the 981-µL post burst dilutions using the RT primer and a second primer (5'-CTTGACGTCAGCCTGGA-3') matching a portion of the S-163 sequence. Amplification of the B16-19 genotype, or of point mutations of this genotype, generates a 187-bp product. These products were digested with *Taq*αI to detect the CUGAACCUUA(123–132) → AAUCG mutation (which shortens the PCR product to 182 bp, generating a *Taq*αI restriction site and fragments of 160 and 22 bp) and with *Xmn*I to detect the U62 → A mutation (which destroys a *Xmn*I restriction site in the B16-19 sequence that would cut the PCR product essentially in half). Products from

selected bursts were also cloned via the TOPO-TA cloning kit (Invitrogen, San Diego). DNA extracts from single bacterial colonies from these clones were amplified with the same primers as above. Selected burst PCR pools and individual cloned amplification products were both subjected to sequence analysis on an ABI 3100 Prism using Big Dye (v.3) chemistry.

Results and Discussion

We began each MA lineage with a genotypically pure population of a high-fitness ligase ribozyme genotype, denoted B16-19 (Figure 1.1). This sequence has been selected repeatedly from randomized populations under a variety of experimental conditions (Schmitt and Lehman, 1999; Lehman, 2004); and as mentioned above, it's a strong competitor in the CE environment because of its high catalytic rate proficiency. Therefore, mutations that accumulate in its sequence likely have a deleterious impact in its catalytic rate. Of course many mutations would be lethal, either destroying the ligase activity of the ribozyme or rendering it unable to fold properly in the 22-min burst time, but those types of mutations are not assayed by MA experiments. Only mutations with small effects that can be fixed in small populations through the sampling error of genetic drift and can persist for a measurable length of time are assayed. In CE, this drift is manifest because one one-thousandth of the population is transferred to a new reaction vessel each burst, resulting in effective

population sizes small enough to allow less-fit genotypes to increase in frequency by chance.

We monitored the progress of CE via PCR amplification of the cDNA that is made during the reaction cycle. Samples of a CE lineage were taken every burst and amplified with primers specific to the ligation substrate and to the reverse transcriptase primer-binding site such that all fit ligase ribozymes should be amplified. While some lineages produced robust PCR products of the expected size (187 bp), others faded out over time, typically developing a high-molecular-weight smear and eventually losing the 187-bp band (Figure 2.3). We equated complete loss of the main band with lineage death, as this meant that the amplification of the wild-type length sequences was not strong enough to keep up with the 1/1000-fold dilution each burst. Note that the loss of a PCR product should trail a few bursts behind the loss of the actual RNA population until the residual cDNA from the last RNA survivors gets diluted below the PCR detection threshold. We also monitored the RNA population itself, in a few cases, by the use of [α-^{32}P] ATP nucleotides in the reaction milieu and polyacrylamide gel electrophoresis to ensure that the RNA population was not substantially growing and outpacing the dilution factor. In fact, the 22-min burst time was chosen to maintain this balance in the sampled lineages.

We continued each lineage until it died or survived to 50 bursts (~150 generations), whichever came first. We ran six or seven replicates of each starting population size (Table 2.1). Strikingly, the average time to extinction is negatively correlated with population size (Figure 2.4). The 100-molecule (bottleneck population size) lineages never survived past 34 bursts, while two each of the 600- and 3000-molecule lineages survived to 50 bursts, and a third 3000-molecule lineage died at burst 49. The average times to extinction were calculated using a value of 50 for those four lineages that survived to burst 50, even though they could have persisted for much longer (Figure 2.5A). Thus the averages for the 600- and 3000-molecule populations are conservative underestimates. Nevertheless, using these averages, there is a statistical difference between time to extinction when the 100-molecule lineages are compared to the 600- or 3000-molecule lineages (P = 0.050 and 0.0060 respectively) and when the 300-molecule lineages are compared to the 3000-molecule lineages (P = 0.010). All tests were made using a model I ANOVA with multiple (k = 6) planned comparisons of means (T' method).

Thus we observed clear evidence of mutational meltdown in these abiotic populations. To determine the underlying mutational events leading to these lineage deaths, we genotyped the PCR populations resulting from selected bursts. This was done both by RFLP analysis of the PCR DNA from each and every burst in our study and by direct nucleotide sequence analysis of

selected bursts. For the direct sequence analysis, we performed the analysis from at least one burst from all lineages, typically within 10 bursts of extinction. Primarily, however, we focused on two lineages, one 100-molecule lineage that died the earliest (4V; Table 2.1) and one 600-molecule lineage that survived to burst 50 (5A; Table 2.1). We performed sequence analysis of the burst PCR products on 10 bursts of lineage 4V (6–16 except 11) and on two bursts of lineage 5A (6 and 9). In addition, we cloned bursts 5, 9, 13, and 14 from lineage 4V and bursts 22, 34, and 47 from lineage 5A. From each of these cloned populations, we obtained complete bidirectional sequence data from between 6 and 10 individual molecules.

From these genotypic data, four types of mutation were evident (Figure 2.6):

The load (polyAdenylation): First, the lineages that developed smears above the 187-bp PCR product became increasingly dominated over time by molecules possessing polyAdenylation (polyA) tracts near the 3' end of the RNA. The mutated RNAs in these populations contained between 1 and ≥ 1000 additional adenosine residues in the region immediately preceding the primer-binding site for reverse transcriptase. In the starting B16-19 RNA, the last nucleotides prior to the primer are AAAG, and this set of three A's is where the polyA expansion takes place. These mutations appear in lineages that are destined for extinction, which often, but not always, evolved by the

following sequence of events: appearance of a smear above the 187-bp band, intensifying and lengthening of the smear, loss of the 187-bp band entirely, and then eventual fading away of the population as a whole.

The floodgate (G135 → A): In some mutants, the terminal guanosine prior to the reverse transcriptase primer-binding site was missing. All of these mutants possessed long polyA tracts as well, and although the reverse case is not necessarily true, we conclude that floodgate was clearly associated with the existence of long polyA tracts.

The immunity: While the first two classes of mutations were associated with lineages destined for extinction, two types of advantageous mutations were occasionally observed. The first is the conversion of the 10-nt sequence CUGAACCUUA (from positions 123 to 132) to the 5-nt sequence AAUCG. This mutation appeared in all four lineages that survived to 50 bursts. Because this mutation results in the creation of a unique *Taq*αI restriction site in the PCR DNA, it was possible to survey its frequency in any given burst, and we did so in all 25 lineages at all bursts prior either to death or to the appearance of a polyA smear. It was detected only once in the 21 lineages that died prior to burst 50, and that was in the last few bursts of the 3000-molecule lineage that died at burst 49. On the other hand, it became fixed between bursts 15 and 20 in the two surviving 600-molecule lineages and

fixed between bursts 5 and 10 in the two surviving 3000-molecule lineages. Thus, this mutation appears to provide immunity against polyA tract formation by an unknown mechanism.

The insurance (U62 → A): The final common mutation that we encountered was U62 → A. This mutation was found in all six 3000-molecule lineages, in three 600-molecule lineages, in two 300-molecule lineages, and in one 100-molecule lineage. Its appearance was related to that of the immunity mutation. It became fixed in all four surviving lineages and in the one 3000-molecule lineage that survived to burst 49, but in only one other lineage. In the other six lineages where it appeared, it was present at a low frequency in a given population (i.e., <10%) and often persisted for only a few bursts before disappearing. When it appeared in the four lineages that survived to 50 bursts, the U62 → A mutation appeared after the establishment of the immunity mutation (Figure 2.5C).

The mutations observed here are examples, at the raw molecular level, of mutational loads and of epistatic responses to counteract them. The polyadenylation happens gradually, as the lengths of the smears increase with generational time (Figure 2.3). The cause of this process is most likely slippage by the reverse transcriptase as it attempts to copy the three adenosines immediately past its primer. As each additional adenosine is

incorporated, the chances for further slippage increase, and the accumulation of adenosines accelerates. The polyA tracts do not completely inhibit ligase activity, but they can contribute to a genetic load. As the molecule gets longer, more time is needed for both reverse and forward transcription, and even small increments in these times can lead to lower fitness in the CE environment. Moreover, should the number of adenosines in the polyA tract greatly exceed the size of the rest of the molecule, two additional negative fitness consequences could arise. First, the enlarged 3'-end of the molecule may interfere with proper folding. Second, the large numbers of adenines in the RNAs can deplete the dTTP pool for reverse transcription and the ATP pool for forward transcription, lowering the reproduction rates for all members of the population.

An important consideration in this study was to guard against the possibility that the results were not simply the result of stochastic sampling failures. If sheer numbers of RNA molecules were being lost by sampling dilute solutions each burst regardless of genotype, then one might expect premature deaths for the smaller populations. We have at least two strong pieces of evidence that this was not happening. First, the lineages do not go extinct suddenly, as would be predicted if RNA was simply not being transferred from burst to burst. The lineages tend to die out gradually, as the cDNA in their composite populations becomes less concentrated and has

a less robust input for the PCR, akin to the process of qPCR. Second, and more importantly, the specific mutations we observe in the lineages are correlated either to lineage survival or to extinction. If transfer failures were the ultimate cause of lineage extinction, then the appearance of, say, the immunity mutation in all four surviving lineages (and only once elsewhere) would not be expected.

Our data demonstrate that RNA populations can accumulate a genetic load in an abiotic environment. The observed mutations can be ascribed to concrete biochemical events that affect fitness. These data reiterate empirical evidence that small population sizes and population bottlenecks magnify the efficacy of random genetic drift, mechanisms proposed to spark significant evolutionary transitions within lineages (Lynch and Conery, 2003). New genotypes arise via high mutation rates imposed in a strictly "asexual" (here, nonrecombining) mode of reproduction. There is little chance of recombination by template jumping by reverse transcriptase (Hu and Temin, 1990; Negroni and Buc, 2001) because RNA concentrations are so low in these CE experiments (e.g., 5 zmol in 25 μL = 2×10^{-10} μM). It remains to be seen whether the loads in these populations can be ameliorated by recombination, which could in theory be deliberately introduced in vitro (Lehman and Unrau, 2005).

While mutation accumulation and mutational meltdown have been documented for extant populations of organisms, the conditions that promote genetic loads and mutational meltdowns would have been especially prevalent during an RNA world, before the advent of cellular life. Naked RNA molecules evolving on the primitive Earth would have suffered high mutation rates for replication, initially low population sizes, and extremely rugged fitness landscapes, all of which would have exacerbated the accumulation of deleterious mutations. Although computer simulations show that compensatory and phenotypically neutral mutations can relax the error threshold for larger populations (tens of thousands) of catalytic RNAs (Kun, *et al*., 2005), our experimental results show that small populations would still be at risk for accumulation of sublethal mutations. It is likely, then, that recombination would have been needed early in the history of life—even before the advent of cellular life and true sexual reproduction—as a means to maintain high-fitness genotypes.

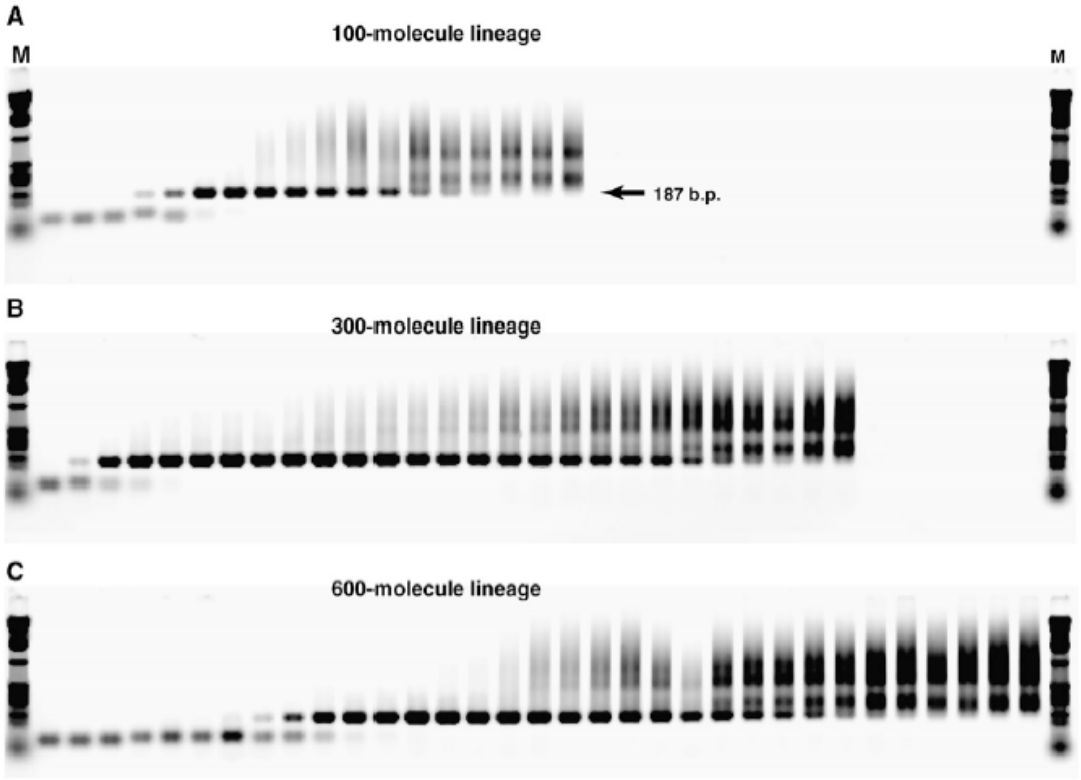


Figure 2.3. PCR of dying populations of 100, 300 and 600 molecules. Samples from consecutive bursts of three replicate lineages were PCR amplified and shown here. In (A) the death of a 100-molecule lineage (4V) can be observed by the complete loss of the 187-bp band, occurring at burst 18 bursts 1–18 approximately. Burst bursts 1–18 are shown. In (B) the death of a 300-molecule lineage (3J) can be observed at burst 27. Bursts 1–27 are shown. In (C) the death of a 600-molecule lineage (3X) was observed at burst 37. Bursts 1–33 are shown.

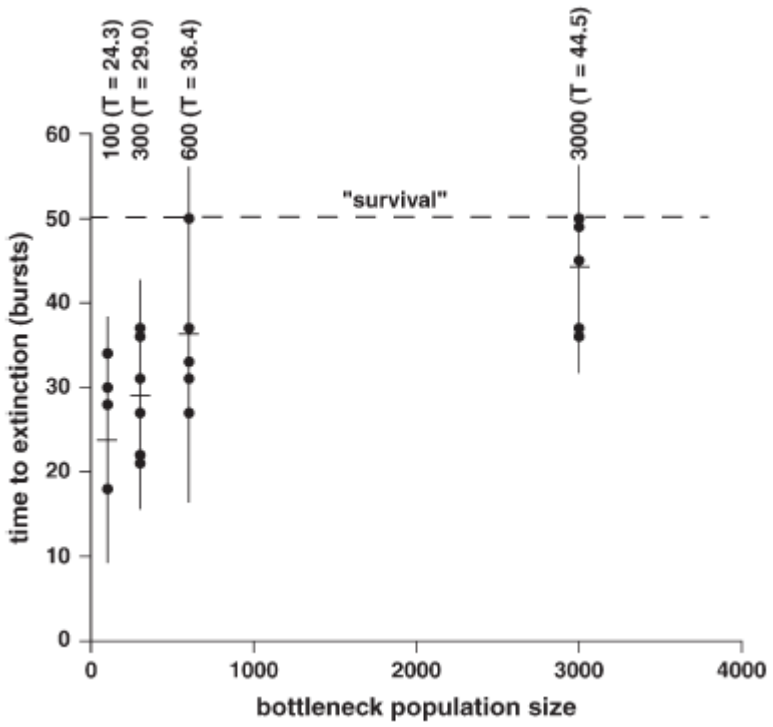


Figure 2.4. Times to extinction according to the effective population size. The dots represent replicate lineages of the same population size. The crosshair in a vertical line indicates the mean time to extinction (*T*) ± 2 SD, for each of the four population sizes. Values of *T* are significantly different for each population size, as explained in the text. A (priori) designation of 50 bursts was made as being the threshold for a surviving lineage. Lineages were not assayed further even if they did not show a sign of extinction. This threshold is indicated by the dashed line.

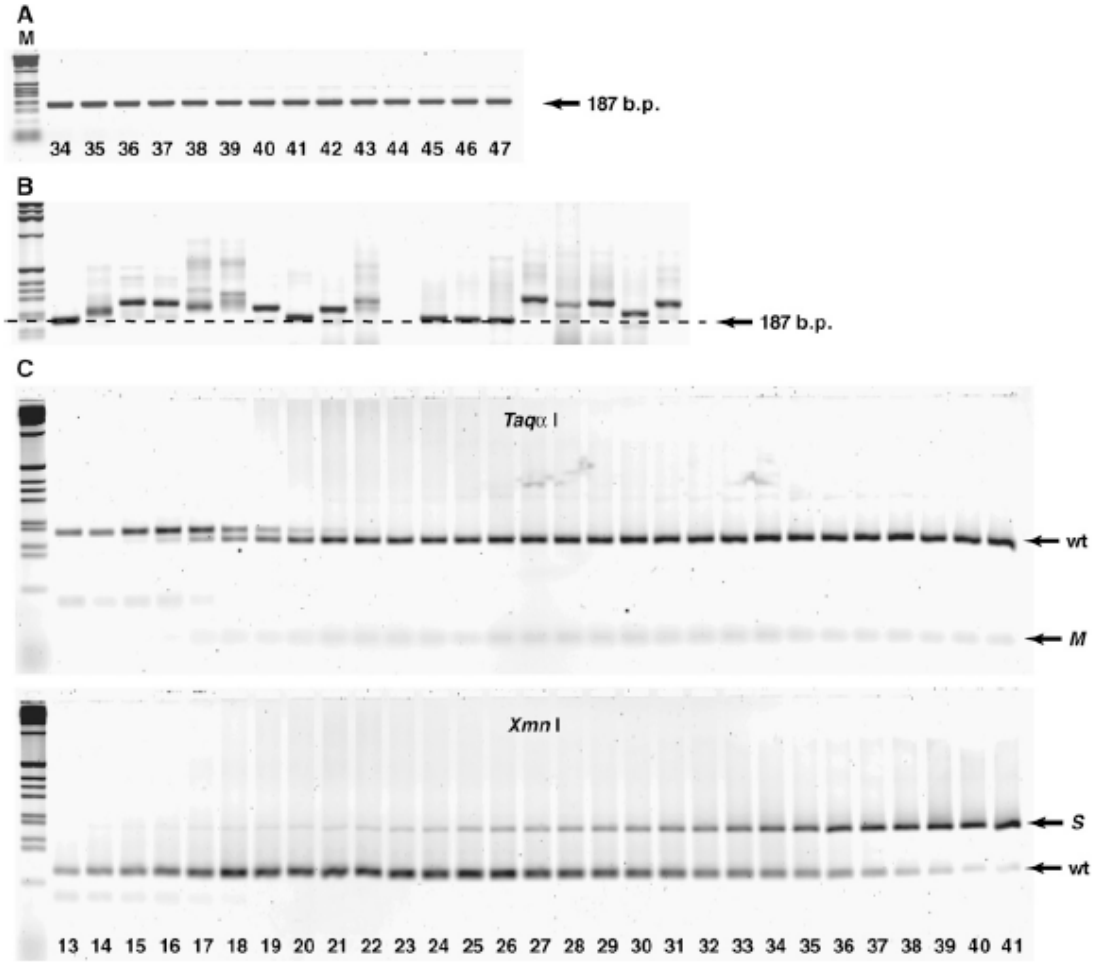


Figure 2.5. RFLP of a survival line, and colony PCR of a death line. (A) Few bursts (34–47) of a surviving lineage of 600-molecule (5A), showing a clean band of the wild-type size (187-nt). This lineage does not show polyA mutant (seen as a smear in the gel image) that caused mutational load in the observed extinct lineages. Immunity and insurance mutations evolved in this lineage as shown in C. The size marker (M) used is a 1-kb ladder. (B) PCR amplicons of 20 cloned molecules from a 100-molecule lineage (burst 13) that went extinct (burst 18). The difference in size in the bands is due to the variable lengthening of the ribozymes caused by polyadenylations. (C) RFLP analysis of bursts 13–41 of 600-molecule lineage (5A) that survived for 50 burst (as shown in A). The enzyme *Taq*αI is used to distinguish between the wild-type B16-19 and the immunity (M) mutation. The enzyme *Xmn*I is used to differentiate B16-19 from the insurance (S) mutant. The immunity mutation arises near burst 15 and become fixed by burst 21, the insurance mutation arises around burst 20 and becomes fixed near burst 43 (not shown in the image).

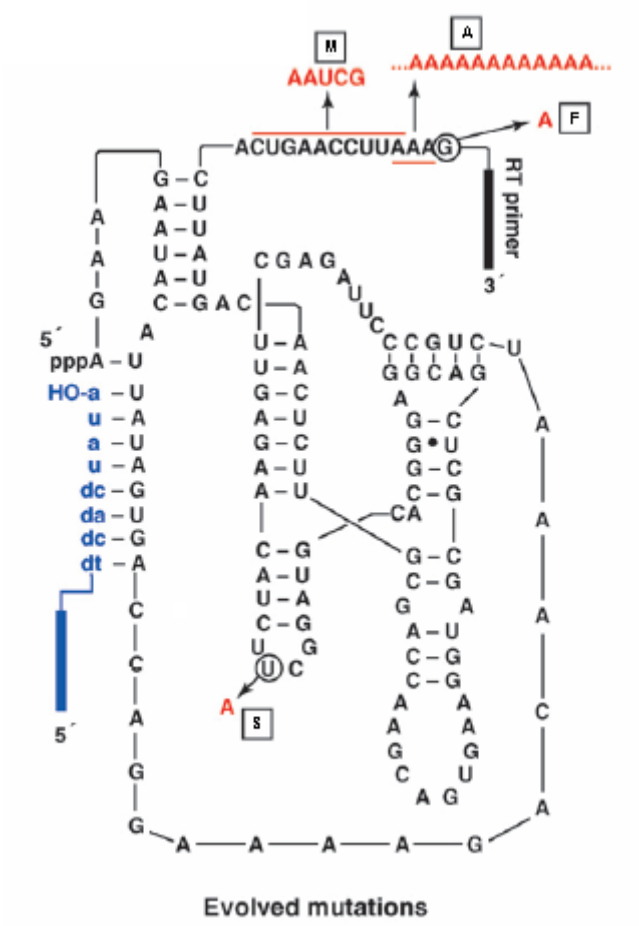


Figure 2.6. Mutations shown in the secondary structure of the ligase B16-19. Mutations (red) were detected in various lineages evolved with the CE. A, polyadenylation of the 3'-end of the ligase (mutational load); F, the G135/A substitution that facilitates polyA tails (floodgate); M, the short sequence CUGAACCUUA replacement by 5' (123) AAU CG (132) 3' called immunity; and S, the base substitution U62/A called insurance. Wild type sequence is shown in black letters and the substrate (DNA/RNA chimera) in blue lowercase letters.

Lineage designation	Bottleneck population size	Extinction at burst no.	Mutations observed[a] (burst appeared, burst fixed)
4V[b]	*100*	*18*	*A (10, 18), F*
4R	100	18	A (11, 18), F
4Z	100	18	A (13, 18), F
4Y	100	28	A (18, 28)
4W	100	30	A (22, 30)
4U	100	34	A (21, 34), S (ND, ND)
3K	300	21	A (15, 21), F
3O	300	22	A (16, 22), F
3J	300	27	A (22, 27), F
3N	300	31	A (21, 31)
3W	300	36	A (30, 36), S (ND, 28)
3X	300	37	A (29, 37), S (ND, ND)
3A	600	27	A (20, 27), F
3V	600	27	A (20, 27), F
3S	600	31	A (19, 31), F
3E	600	33	A (25, 33)
3Q	600	37	A (26, 37)
5A	*600*	*50+*	*M (15, 21); S (20, 43)*
3D	600	50+	M (ND, NF); S (ND, 32)
2I	3000	36	A (18); S (7, gone by 25)
2K	3000	37	A (22); S (3, 30)
2A	3000	45	A (31); S (8, gone by 10)
2J	3000	49	A (23, NF); M (40, NF); S (6, 30)
1A	3000	50+	M (5, 10); S (6, 10)
1C	3000	50+	M (5, 6); S (6, 6)

Table 2.1. Summary of the evolutionary history of the ribozyme lineages. The superscripts are as follows: [a]The letters represent A, polyadenylations; F, floodgate; M, immunity; S, insurance; NF, never fixed (always polymorphic); ND, not determined. [b]Lineages in italics were subjected to first RFLP and then to direct sequence analyses.

CHAPTER THREE

Quasispecies Behavior Observed in RNA Populations Evolving in a Test Tube

(Díaz Arenas and Lehman. 2010. *BMC Evol Biol*)

Background

In 1971, Manfred Eigen introduced the concept of the quasispecies to understand the dynamics of genotypes in populations of infinite size (Eigen, 1971; Eigen and Schuster, 1979). Later, Eigen (2000) did a linear approximation of the non-linear differential equations used to explain the behavior of RNA or RNA-like molecules in error-prone environments in order to adapt the model to finite population sizes. A quasispecies is basically an equilibrium population of mutant molecules distributed around a central parental sequence; the so-called master sequence. This population structure occurs at high mutational rates, such that the progeny of an individual RNA sequence (the mutant cloud) can be rapidly produced, a key feature for the quasispecies formation. Different sequences (e.g., genotypes) can form mutant clouds of various sizes that can compete for survivorship during the evolution of the population. This can generate a fluctuating dynamic as clouds of mutants are replaced by other ones at the interplay of selection and random drift.

Quasispecies behavior posited in naked RNA populations has great relevance to the study of the origins of life on Earth. During this period, the replication of genotypes did not have the benefit of an error correction process that required more sophisticated machinery (e.g., editing replicases) and larger genomes (Szathmary and Maynard-Smith, 1997). Also, the sizes of the populations were likely very small and thus the effect of random drift was strong (Wilke, 2005). These two characteristics would have imposed challenges for the survival of the nascent populations. In particular, genetic deterioration was a risk by means of the accumulation of deleterious mutations and the continual removal of the fittest class by random drift, a process called Muller's Ratchet (Felsenstein, 1974).

Mutants in a quasispecies cloud are characterized by short Hamming distances (Eigen, 1993), meaning that genotypes can be regenerated from closely related ones relatively easily. Additionally, because ribozymes have a phenotypic plasticity that allows more than one genotype to code for the same phenotype, there is a wider spectrum of mutations that have a neutral or slightly neutral effect on fitness (Nimwegen, *et al*., 1999; Lehman, *et al*., 2000). Consequently, information that is relevant for the survival of populations can potentially be stored in a quasispecies and not in individual genotypes. This confers mutational robustness to the population (Wilke, 2005), and serves as a route to escape Muller's Ratchet.

Quasispecies behavior is a higher-order-effect phenomenon because of the multiple and constant interactions among genotypes in the populations, making it difficult to demonstrate experimentally (Wilke, 2005). To date, it has been documented mostly in viral populations, but also recently in cellular automata, in plant viroids, and in animal RNA viruses associated with the "survival-of-the-flattest" phenomenon (Sardanyés, *et al*., 2008; Sanjuán, *et al*., 2007; Comas, *et al*., 2005). However, quasispecies behavior has never been demonstrated using test tube experimentation with empirical populations of naked genes that may resemble those existing during biogenesis on Earth.

Most of the *in vitro* experiments performed to date have either one of the following two outcomes: (1) There is an inexorable reduction in the genetic variability of the populations with the passage of successive generations. This outcome can be either the consequence of Darwinian selection favoring some genotypes, usually the fittest class over others, or the consequence of artificial selection applied to the system in order to obtain molecules with specific properties. Examples of this type of experimental outcome have been reviewed (Lehman and Joyce, 1993; Schlosser and Li, 2005; Schlosser, *et al*., 2009; Tuerk and Gold, 1990; Ellington and Szostak, 1990; Carothers and Szostak, 2006; Jhaveri and Ellington, 2002; Joyce, 2004). (2) There is recurrent outcomes of general motifs or even specific genotypes, as in the

case of the hammerhead motif (Salehi-Ashtiani and Szostak, 2001), the isoleucine aptamer (Lozupone, *et al*., 2003), or the replicability and recurrence of both group I ribozymes (Hanczyc and Dorit, 2000), and class I ligases (Lehman, 2004), to mention a few.

In the current study, continuous evolution (CE) *in vitro* (Wright and Joyce, 1997) was used to track the evolutionary dynamics of small populations of catalytic RNA under a high mutational pressure imposed by alteration of the chemical environment. As the CE experiments involve a reduced intervention of the experimenter and no directed selective pressure applied during the experimentation, the outcome is a true mimic of what happens to molecular dynamics *in vivo* (Joyce, 2004). Presented here is the first report of quasispecies behavior observed in ribozyme populations evolving in a test tube.

Material and Methods

RNA Preparation

B16-19 ligase ribozymes were freshly prepared by transcription of B16-19 clones obtained in a previous *in vitro* evolution experiment (Schmitt and Lehman, 1999). The transcripts were purified by 8% polyacrylamide/8 M urea gel electrophoresis. The concentration of the transcripts obtained was measured with a UV spectrophotometer at 260nm. A dilution series was then

performed to obtain the desired concentration of 100 molecules in the 8.20 µl aliquot used to seed the evolution experiments.

Continuous *in vitro* evolution

Ligase ribozyme populations were evolved using the continuous *in vitro* evolution methodology (Lehman, 2004; Wright and Joyce, 1997; Schmitt and Lehman, 1999). To summarize, 2.03×10^{-8} nM B16-19 ligase (100 molecules) were incubated with 64 pmol of oligonucleotide S-163 (5′-CTTGACGTCAGCCTGGAC**TAATACGACTCACUAUA**-3′ = the chimeric substrate, with ribonucleotides underlined, and the T7 promoter in bold face letters), 50 pmol of RT primer (5′-GCTGAGCCTGCGATTGG-3′), 240 U of MMLV reverse transcriptase (United States Biochemicals, Cleveland), 60 U T7 RNA polymerase (Ambion, Austin, TX), 5 nmol each dNTP, 50 nmol each rNTP, 25 mM $MgCl_2$, and 40 µM $MnCl_2$, in a reaction buffer with 50 mM KCl, 30 mM 4-(2-hydroxyethyl)piperazine-1-propanesulfonic acid (EPPS), pH 8.3. The 25 µL reaction mix was incubated at 37°C for 22 minutes. At the completion of this time, a 3µL aliquot was taken from the reaction tube and mixed with 981 µL of DEPC-treated and/or RNase-free water in the dilution tube, to stop the reaction. These amounts ensure the preservation of a constant harmonic mean of the population size (Appendix). An aliquot taken from the dilution tube was used to seed the next reaction cycle of the continuous evolution event.

Population assessments

(1) The survival of the evolved populations was surveyed through PCR amplifications of all of the bursts used to seed a reaction cycle. The PCR products were run through 2% agarose gel electrophoresis containing ethidium bromide. Visualization of the gels by trans-illumination allowed the identification of the correct band size (187 bp) when the population is alive.

(2) Preliminary genetic variability was evaluated by RFLP test using the restriction enzymes used previously (Soll, *et al*., 2007). In the populations where genotypic variability was detected, a more extensive genotypic characterization was done.

Genotypic characterization

Specific bursts of the populations with genotypic diversity were cloned using the CloneJet^{TM} PCR Cloning Kit (Fermentas, Maryland) and *E. coli* competent cells (Invitrogen, San Diego). Colony PCR was used to extract the insert from single clones and further sequencing was done with BDT V3.1 chemistry.

The sequences were aligned with ClustalX 2.0.11 software; the alignments were edited with BioEdit sequence alignment editor v7.0.9.0 (Tom Hall, Ibis biosciences, Carlsbad) and the chromatogram viewer FinchTV v1.4 (Geospiza

Inc, Washington). To estimate if the number of clones sampled in each selected burst contained all the unique genotypes present in the population, rarefaction plots were produced by constructing tables of random numbers (random.org) in MS Office Excel (2003) from which a non-linear curve fitting following the procedure in Lehman and Wayne (1991) was then produced using Origin Pro v8.0 software (OriginLab Corp, Massachusetts).

Phylogenetic network mapping

For each quasispecies found, all the genetic variants were aligned using DNA alignment v1.3.0.1 (Fluxus Technology Ltd.), and plotted together using the median-joining method (Bandelt *et al*., 1999) implemented in NETWORK v4.5.1.0 software (Fluxus Technology).

Results and Discussion

CE experiments

We evolved four clonal 100-molecule populations of B16-19 ligase ribozymes using the continuous *in vitro* evolution (CE) method (Wright and Joyce, 1997). The CE protocol is a means to induce the rapid evolution of ligase ribozymes using the relatively error-prone Moloney Murine Leukemia Virus Reverse Transcriptase (MMLV-RT) and T7 RNA polymerase to sustain RNA populations through sequential serial transfers (Wright and Joyce, 1997; Voytek and Joyce, 2007). Each serial transfer involves roughly three cycles of

amplification that produces a rapid proliferation of RNA molecules, and hence is termed “burst” in this study. The experimental conditions used were the same as in earlier experiments (Wright and Joyce, 1997, Schmitt and Lehman, 1999; Soll, *et al*., 2007), with the exception that we added $MnCl_2$ to the reaction vessel to increase the error rate of the protein enzymes. Both *in vivo* and *in vitro*, Mn^{2+} ions lower the substrate specificity of reverse transcriptase, resulting in a significantly higher error rate (El-Deiry, *et al*., 1984; Lazcano, *et al*., 1992; Vartanian, *et al*., 1999).

Populations evolved under these high mutation rate conditions did not show a shortened extinction time (24.3 bursts), as it was observed in the previous experiments that used only a weak mutational pressure of no added $MnCl_2$ (Figure 2.4 and Soll, *et al*., 2007). We were able to evolve all four lineages for 50 bursts without observing the loss of viability caused by the Muller’s Ratchet effect, and thus a mutational meltdown was never observed (Figure 3.7). These results are a consequence of quasispecies behavior, as the sequencing data shows (see below).

Genotypic characterization

To investigate the cause of the observed extended time to extinction, we performed a preliminary inspection of the population genetic variability using RFLP. The cDNA in a population at any burst can be amplified via PCR and

then genotyped by either RFLP or direct nucleotide sequence analysis. Through our preliminary inspection, we find mutant forms that have been nearly fixed in all the bursts we selected from the different replicate lineages. Based on RFLP assays (Figure 3.8A) we selected two lineages (6H and 6L) for a more extensive characterization of the genotypic variability, and from each we studied three and four bursts, respectively (Figure 3.8B). These bursts were cloned and sequenced. We constructed rarefaction plots from the sequencing data to calculate the number of clones in each burst that needed to be genotyped in order to sample the bulk of unique genotypes present in the population. We found that in most cases the sample gathered was representative of the population diversity. For cases in which our sample was not representative, we performed more cloning and sequencing until a good representation of population diversity was obtained.

We found that the number of clones required for inspection was similar when comparing bursts within the two lineages, but different when comparing the two lineages themselves (Table 3.2). This result is not surprising as the quasispecies is in a fairly stable equilibrium and the environmental conditions are nearly constant during the CE experiments. Consequently, the populations can experience different equilibrium dynamics from the same starting point but once found remain nearly steady (Bull, *et al*., 2005).

Network analysis

Alignment of the nucleotide sequences data showed a trend in the population dynamics in which a majority of the clones have the same genotype, while a minority has slightly different ones. This observation was the first indication that quasispecies behavior was present in these lineages. We drew phylogenetic networks (Figures 3.9 and 3.10) to find the genetic relationship among the mutants and the structures of each putative quasispecies (Fernandez, *et al*., 2007). The structures of the networks show a dynamic characterized by a dominant sequence that is present at the highest frequency, the master sequence around which the other less frequent mutant sequences are located. This population structure is characteristic of a quasispecies in which the dominant sequence is called the “master sequence” and the surrounding mutants form the mutant cloud (Eigen and Schuster, 1977).

Quasispecies behavior has never been demonstrated before during *in vitro* evolution experiments with catalytic RNA. Other *in vitro* experiments have shown either convergence on a phenotype or recurrence of a genotype or motif, but not the type of dynamic of quasispecies that we are documenting here. For example, Yingfu Li and colleagues studied how the composition of a population of RNA-cleaving DNAzymes changed over time in response to selective pressures acting on the phenotype (Schlosser and Li, 2005;

Schlosser, *et al*., 2009). Similarly to the findings reported here, they found a dynamic fluctuation in the structure of the population. Many sequence classes peaked in frequency at different rounds of selection, but in that case one class appeared to consistently maintain a high frequency. It will be interesting to explore the population structure that these DNAzymes would adopt if the mutational rate were increased. Perhaps mutational coupling would arise in these molecular populations as well. Another study of relevance to the phenomenon is the evolution of the RNA variant V2 of the Qβ virus performed by Orgel and co-workers (Orgel, 1979). In this case, the mutagen ethidium bromide (EtBr) was added to the reaction vessel during the serial transfers. RNAs resistant to EtBr evolved and adapted to increasing EtBr concentrations. In this case however, the CE/Mn^{2+} data contrasts with Orgel as the mutagen has a direct effect on the RNA and therefore selection favored variants that caused mutations in the EtBr-binding sites and a single “winner” emerged.

We calculated the genetic (Hamming) distances between the mutants and the most abundant master sequence (Figure 3.11), based on the network diagrams (Figures 3.9 and 3.10), and found that the quasispecies they formed are characterized by relatively close connections (Figure 3.12) between the mutants (mean, 6; mode, 1; min, 1; max, 18). The close connectivity between these mutants may confer mutational robustness to the population and explain the extended time to extinction observed. Mutational robustness can be

phenotypic or genotypic. Here we refer to mutational robustness as the ability of the system to persist and evolve after mutations occur in its parts (Wagner, 2008). Mutational robustness can be observed at different levels. Examples include; protein tolerance to amino acid substitutions (Bowie, *et al*., 1990), the genetic robustness of microRNAs (Borenstein and Ruppin, 2006), the error tolerance of complex biological networks (Albert, *et al*., 2000), and in RNA viruses experimentally evolved in low and in high coinfection regimes (Montville, *et al*., 2005).

The quasispecies clouds studied not only confer mutational robustness to the populations, but also further evolvability (Wagner, 2008) as indicated by the fluctuating dynamic observed in the populations. This includes changes in the shape of the clouds — and gross amount of mutant sequences —from one burst to another (Figures 3.9D and 3.10E). It is likely that the high mutation rate used by addition of Mn^{2+} to the reaction vessel increases the mutational rate of the replication process and consequently alters the equilibrium distribution of the population (Bull *et al*., 2005). The quasispecies behavior sustains this shift in equilibrium distribution because of the mutational coupling among its mutants. Our populations explore variable alternatives of sequence space over the course of their evolutionary history (50 bursts).

The shift in the equilibrium distribution is possible because of two main reasons:

(1) Sequences such as ligase ribozymes posses the property of buffering mutations through epistatic interactions between secondary structure arrangements. These arrangements strongly stabilize the structure and thus a broader range of mutations will have a neutrally selective effect, hence relaxing the error threshold (Kun, *et al*., 2005; Holmes, 2005).

(2) The fitness of each genotype in the population is normalized with the total number of genotypes in the system (assuming single locus theory applies). Thus, the proportional contribution of each genotype to the total fitness decreases as the number of genotypes increases (Kauffman, 1993).

Many point mutations in the ribozyme genotypes are neutral in phenotype due to the more than one genotype-to-phenotype ratio (Nimwegen, *et al.*, 1999; Lehman, *et al*, 2000). However, the mutational buffering of secondary structure epistatic interactions mostly favors the fitness of the lower class mutants (e.g., low Hamming distance values). The genetic load generated in higher class mutants will likely disrupt secondary interactions and the stability of the individual ligases. Oddly enough, an increase in the mutational rate does not cause a proportional increase in the genetic load, and therefore the

population does not become extinct at a faster pace. What must be happening in this case is that mutants of lower class emerge quickly, generating a large low-mutant class in the early evolutionary pathway of the population. These mutants have short Hamming distances, and thus, probably similar fitness values. Wilke (2003) observed that the first couple of replication cycles mostly determine fixation or extinction for an invading sequence. This could perhaps be a group of closely connected sequences, such as the quasispecies mutant cloud. The major contribution to fixation probability comes from the connection matrix of the local genetic neighborhood of the invading sequence (or mutant class); sequences farther away on the neutral network that are less related become relatively unimportant, and may be drawn out of the population by genetic drift and mutation selection balance.

The level of connection in the matrix is determined by the Hamming distance values of the mutants in the network. Genotypes that are closely connected by short Hamming distances are closely related in the sense that they can rapidly (e.g., in a few generations) be regenerated from one another in the eventual case of being removed from the population by random drift. In contrast, poorly connected genotypes (e.g., only a few relatives) will have a slow recovery into the population, if at all. In this scenario, because mutants with short Hamming distance may have similar fitness values, individual sequences are not essential for the survival of the population but rather the group of close-

connected individuals with mutational robustness (Kimura, 1983; Wilke, *et al.*, 2001; Codoñer, *et al*., 2006). Therefore, the quasispecies cloud itself is the target of selection (Bull, *et al*., 2005) and not the individual sequences. This process is analogous to the manner in which kin selection operates in animal societies (Maynard Smith, 1964; Maynard Smith and Szathmáry, 1999). Ribozyme populations therefore can — by means of indirect reproduction effects — evolve a mutational robustness; a behavior that empowers selection with an advantage relative to other evolutionary forces (e.g., the strength that random drift has in populations of small effective sizes). This strengthening in selection allows the population to: (a) overcome Muller's Ratchet, (b) avoid a mutational meltdown, and (c) stay extant.

In lineage 6H, an early time point (burst 5 out of 50) shows that a quasispecies cloud is already formed (Figure 3.9A), in which the master sequence is still the wildtype B16-19, with a frequency of 76%. This quasispecies cloud was constructed from a sample of 102 clones, which contained 16 different genotypes. It should be noted that, although the population size was ostensibly kept constant at 100 molecules throughout each lineage, the cloning procedure involves PCR amplification. Therefore, the sampling of genotypes (e.g., 102 clones) from the population is effectively sampling with replacement of the total diversity. Nevertheless, the observed sample diversity was estimated at 1.49, as measured by Shannon Index (1).

(1) $$H' = - \sum_{i=1}^{S} (p_i * \ln p_i)$$

Where *S* is the number of species, p_i is the relative abundance of each species as given by n_i/N, n_i is the number of species *i*, and N is the total number of individuals.

Deeper examination of these bursts in this lineage (Figures 3.9B, and 3.9C) revealed a change in the master sequence identity, frequency, number of genotypes and Shannon diversity values. In general, the quasispecies formed in each burst had different identities from each other, but their characteristics are fairly similar (Table 3.3). This similarity is perhaps the result of the fact that the sequence space available for exploration by the populations is bounded by a unique starting point; they are all genotypically identical at the beginning of the experiment. Therefore, the area of sequence space that can be explored in 50 serial transfers would be relatively small and the lineages may be not very different in their diversity values.

Genotypes that are present at a higher frequency in one burst can become a master sequence at a later burst. Conversely, a master sequence that, having once been displaced, was never observed to come back to high frequency in the population. For example, in lineage 6H, the following transition can be

observed in the master sequence identity and frequency: burst 5, B16-19 (86%); burst 35, MS1 (91%); burst 50, MS2 (48%). This dynamic of master sequences being displaced by one another resembles that of clonal interference in which advantageous mutants have to compete for resources and some get displaced (Fisher, 1930; Muller, 1932; Hill and Robertson, 1966). A network drawn by combining the sequences of all bursts inspected in lineage 6H (Figure 3.9D) shows the fluctuation of the equilibrium dynamic of the quasispecies of lineage 6H over the course of its evolutionary history.

We found similar results in the replicate lineage 6L. In this lineage we studied four bursts (Figure 3.10). The dynamic of the lineage is similar to that of lineage 6H in that different quasispecies clouds emerge and evolve through time. Two successive bursts (Figure 3.8B) reveal that this fluctuation can occur in relatively short time (Figures 3.10A and 3.10B). Interestingly, some of the master sequences that appeared during this lineage (Figures 3.10B, 3.10C, and 3.10D) are the same master sequence that were observed in lineage 6H (Figures 3.9B and 3.9C). These results suggest that some quasispecies may develop a stronger mutant coupling than others that enables them to recurrently out-compete other quasispecies present during the evolution of the lineages. Similar to the pattern in lineage 6H, the characteristics of the quasispecies changed over the course of the evolutionary history as indicated by the master sequence identity, frequency,

number of genotypes and calculated diversity values (Table 3.4). This dynamics of staggered dominant genotypes that fluctuate as the population evolves (Figure 3.10E) may be a reflection of the interplay of Darwinian selection and random genetic drift acting on the quasispecies.

In general, from all the networks drawn, it is observed that the number of mutations that have appeared in more than one burst and/or lineage constitutes 61% of the total number of sequences explored, and the great majority of these mutants have become part of a master sequences during the lineages evolutionary history (green, purple and blue spheres in the quasispecies networks of Figures 3.9 and 3.10). The fact that most of the mutants that have evolved in these lineages have been able to persist in time could be a consequence of the mutational robustness of the quasispecies owed to the short Hamming distance values (Figure 3.12). Most of these recurrent mutations belong to the master sequences (Figure 3.11 and 3.13).

In lineage 6H, the initial change in master sequences from B16-19 to MS1 implies ten changes, and the further change from MS1 to MS2 implies one change (Figure 3.13, inset). Similarly, in lineage 6L, the initial change from B16-19 to MS3 implies sixteen changes; but further changes in the master sequence transitions are seventeen and one (Figure 3.13, inset). The Hamming distance values are generally low (in the range of tens), with

eighteen being the maximum value. The distance between the two most recurrent master sequences is actually only one, and these master sequences (MS1 and MS2) are the most representative in time and sequence space (54% of the total). The Hamming distance values, in addition to the high frequency of recurrent mutations, support the idea of mutational robustness evolved in the system through a quasispecies behavior in ligase populations evolved *in vitro*.

These results — of ligase RNA molecules forming population structures in which cooperation-like dynamics are more beneficial than competition — suggest that an altruistic behavior (e.g., cooperation) is an advantageous feature to ensure survival of populations during the RNA world (Hayden and Lehman, 2006), when the population size was small, when the mutational rate was high, and when random genetic drift had strong effect; conditions that probably prevailed on the prebiotic Earth (Kimura, 1983; Santos, *et al*., 2004). Additionally, quasispecies have an organizational structure with the properties proposed by Kaufmann (1993) to be necessary for the origin and preservation of genetic information. In this structure, the closely connected cloud of the quasispecies can serve as an information-preserving core and distantly surrounding genotypes can be targets for random genetic drift without affecting the information relevant to the survival of the population (stored in the core). According to Kauffman (1991), organized systems may have arisen as a

consequence of the property of some elements to establish different levels of connectivity among each other. The highly interconnected elements can create organizational cores able to preserve the information relevant to survival of the system (e.g., autocatalytic function). In contrast, less interconnected elements can serve as a reservoir of mutations without a detrimental effect on this information. Thus, during the ancient acellular times at biogenesis on the Earth, the assemblage of information cores, perhaps in the form of quasispecies clouds, may have provided the necessary route to increase population sizes and allow enough time for information to mature into more sophisticated functions necessary for cellular life.

To verify that we have characterized a quasispecies structure in the ligase ribozyme populations evolved at high mutation rate, we sequenced a number of clones from a lineage (3D) evolved at low mutation rate (Table 2.1). This lineage is the smallest that survived 50 bursts under no added Mn(II), 33 and 48 clones from burst 10 and 50 respectively were sequenced and networks were drawn for each. This lineage (Figure 3.14) does not show the main characteristics of the quasispecies: a master sequence surrounded by a mutant cloud of low connectivity (Hamming distance). The mutant clouds observed have relatively similar size and are spread over a relatively large sequence space. Burst 10 (Figure 3.14A) shows a sequence also present in burst 50 (Figure 3.14B), but the sequence never reaches more than 50%

abundance, as exemplified by the master sequence in 6L42 (Figure 3.10D) that represents 62% of the 45 clones sequenced for that burst. The master sequences of the quasispecies evolved under high mutagen experiments have a clear dominant presence in the burst in which they appeared, as expected in a quasispecies (Eigen, 1971). In contrast, this lineage (3D) does not have such a feature; hence the plot shows no more than a diversity distribution in the lineage.

In summary, our results indicate that the quasispecies formed in the high mutation rate ligase populations allow them to persist. The mutants in the mutant cloud stay closely connected and thus they can be easily regenerated from one another even if lost from the population through random genetic drift. This behavior empowers selection relative to random drift, as the information relevant to the survival of the population is stored in a close-knit network of mutants and not in the individuals *per se*. It is possible that such a population structure, if available, would have greatly benefited primordial pools of nascent RNA molecules on the early Earth. Instead of relying on the fortuitous advent of specific self-replicating genotypes, the RNA World would have the luxury of swarms of quasispecies evolving over time, buffered against the extinction by informational decay, as theorized by Eigen (1971).

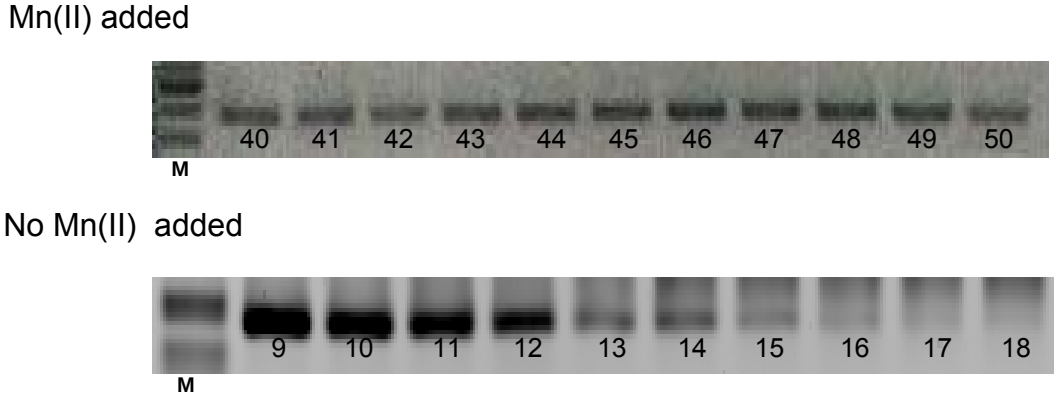


Figure 3.7. PCR of small populations evolved at high and low mutation rates. Populations evolved at a high mutation rate (top) have [40 μM] $MnCl_2$ added into the reaction vessel during the CE experiments. M indicates marker and the numbers below the PCR bands represent bursts. Notice that the lineage evolved with Mn(II) added, persisted to burst 50 (the declared viable threshold) without signs of fading out. In contrast to populations evolved at a low mutation rate (bottom) with no Mn(II) added (See chapter II, Soll, *et al.*, 2007). The persistence of the PCR band indicates the population is able to sustain itself through time, whereas a fading band is a sign that the population is being reduce in size and may be approaching extinction as a consequence of mutational meltdown (e.g. burst 18, bottom image).

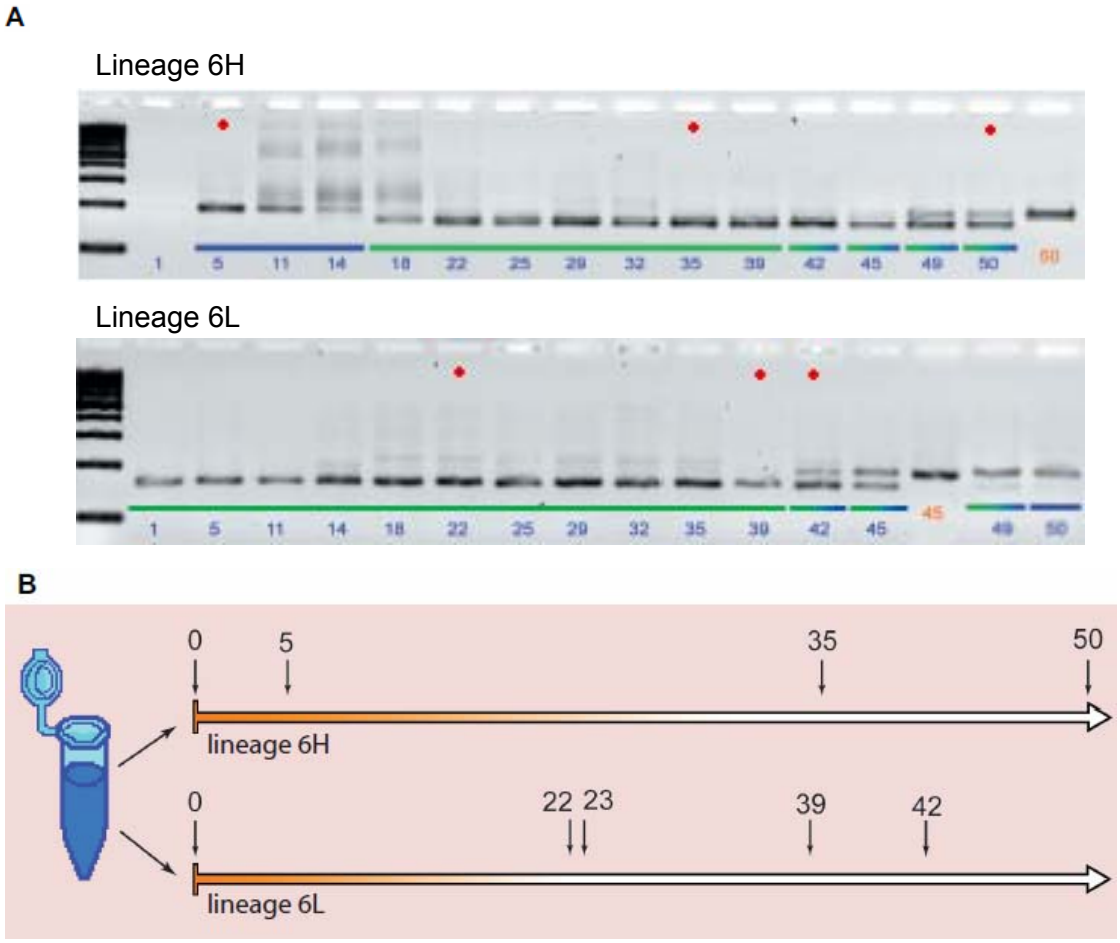


Figure 3.8. Lineages selected for genotypic analysis. A. PCR gel images of the RFLP assay, with *Taq*αI, of two populations evolved at high mutation rate, showing genetic diversity. Each lane represents a burst, as numbered (in blue), and each gel correspond to two replicate lineages (6H and 6L). A negative control (no restriction enzyme added) is shown (burst 50 in lineage 6H and burst 45 in lineage 6L). The upper PCR band in 6H and 6L gels corresponds to the wild type ligase B16-19. The lower PCR band corresponds to a mutant sequence that acquired the restriction size for the enzyme *Taq*αI. For some lineages it takes more than one burst to be above the detection limit of the PCR amplification (e.g. burst 1 in lineage 6H). The blue, green and gradient lines at the bottom of the PCR bands indicate the presence or absence of polymorphism. The red dots indicate bursts that were characterized by cloning and nucleotide sequencing (burst 23 of 6L is not shown). B. Scheme showing the lineages and bursts selected for sequencing (numbers above the lineages) to study the genetic diversity. Both lineages are represented with an arrow because they survived to 50 bursts and they were not continued.

.

Quasispecies	Number of sequenced individuals	Expected number of unique genotypes
6H-5	102	21
6H-35	98	27
6H-50	108	54
6L-22	41	13
6L-23	48	13
6L-39	54	24
6L-42	45	24

Table 3.2. Data from rarefaction plots and sequencing data. The number of individuals that needed to be sequenced to detect all the expected unique genotypes was obtained for all the bursts in the two lineages studied (first column). Genotypic diversity is inferred based on the asymptotes of rarefaction plots (last column). The number of clones sequenced (middle column) in each burst exceeds the expected number of unique genotypes calculated (last column). This indicates that the samples size taken from these lineages is a good representation of the actual number of unique genotypes present in the population.

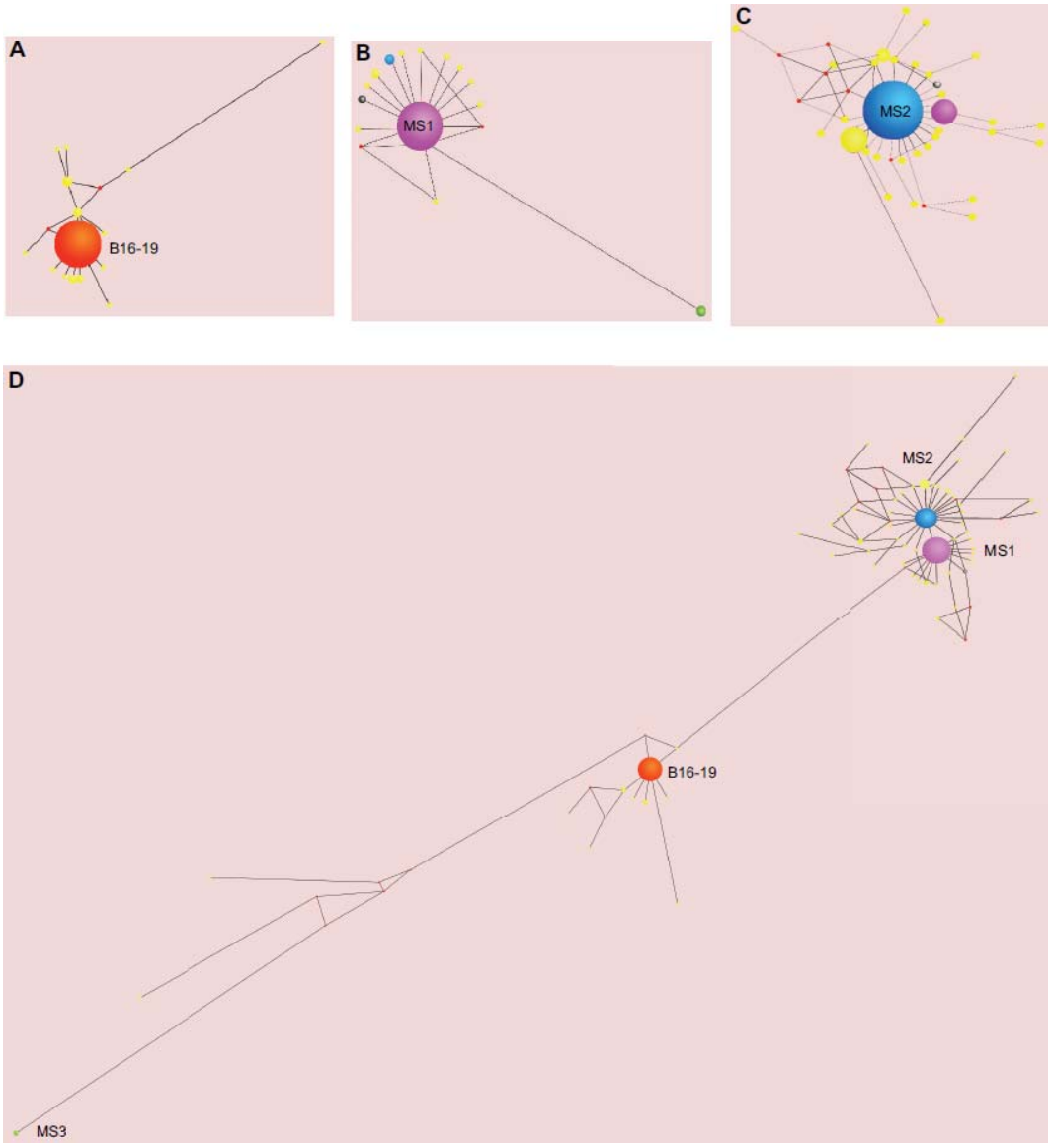


Figure 3.9. Quasispecies clouds formed in lineage 6H. The bursts evaluated (Figure 3.8) shows a population structure of a quasispecies, where the mutant dynamic change over time (A, B, and C). The spheres represent mutants distributed as clouds, with a frequency of mutants proportional to its size. The mutant clouds are connected by a hamming distance (nucleotide differences). The most frequent cloud is the master sequence (MS) which is color coded and changes from wildtype at burst 5 (A) to MS1 at burst 35 (B), and to MS2 in burst 50 (C). The relationship between all the quasispecies is shown in (D).

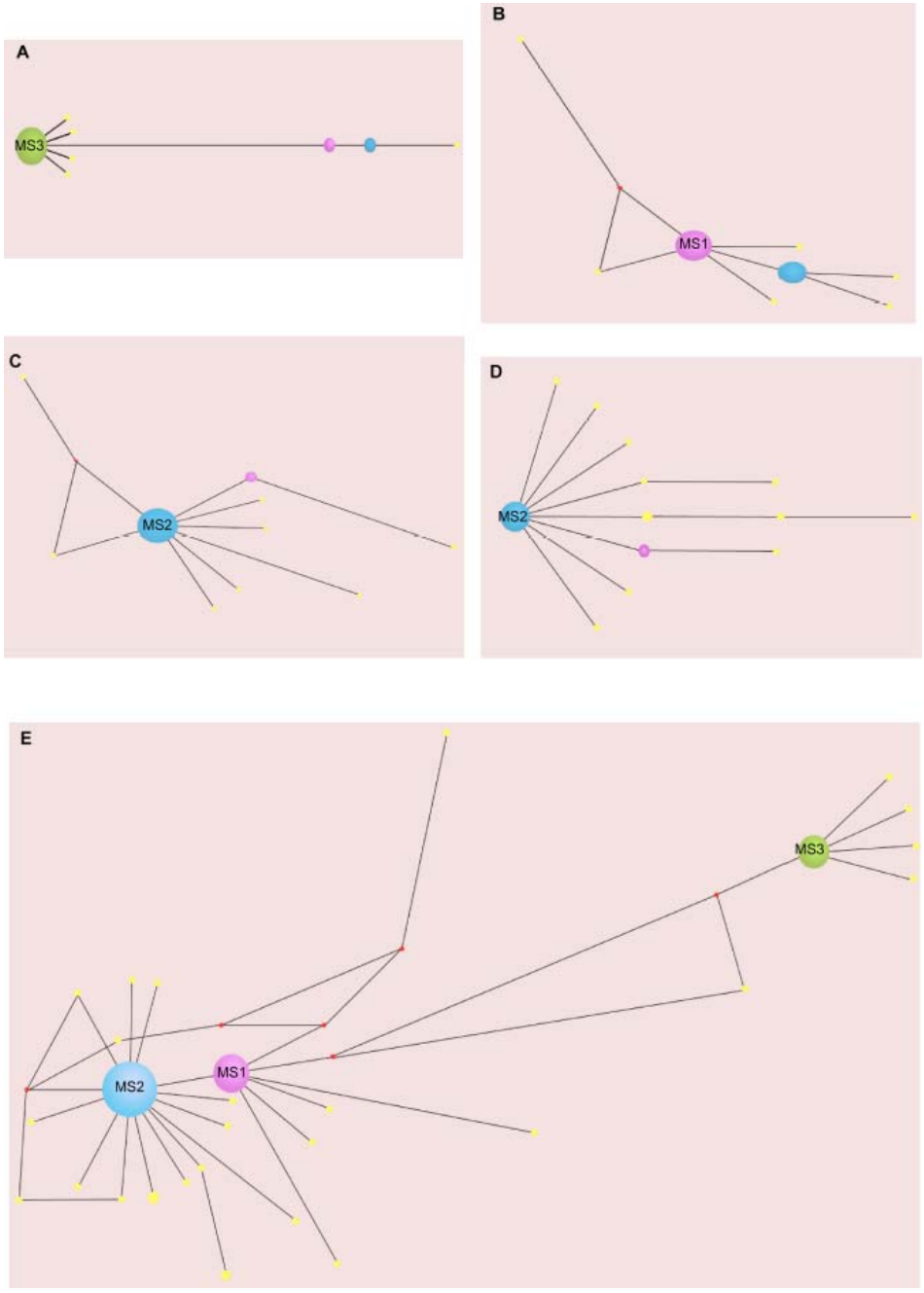


Figure 3.10. Quasispecies clouds formed in lineage 6L. Evaluated bursts (see Figure 3.8) show a quasispecies population structure, where the mutant dynamic changes over time (A, B, C and D). The dynamic is detected by the change in the master sequence (MS) from MS3 at burst 22 (A) to MS1 at burst 23 (B), and to MS2 in burst 39 (C) and 50 (D). The relationship between all the quasispecies is shown in (E). Notice MS1 and MS2 were also present in the previous lineage (6H).

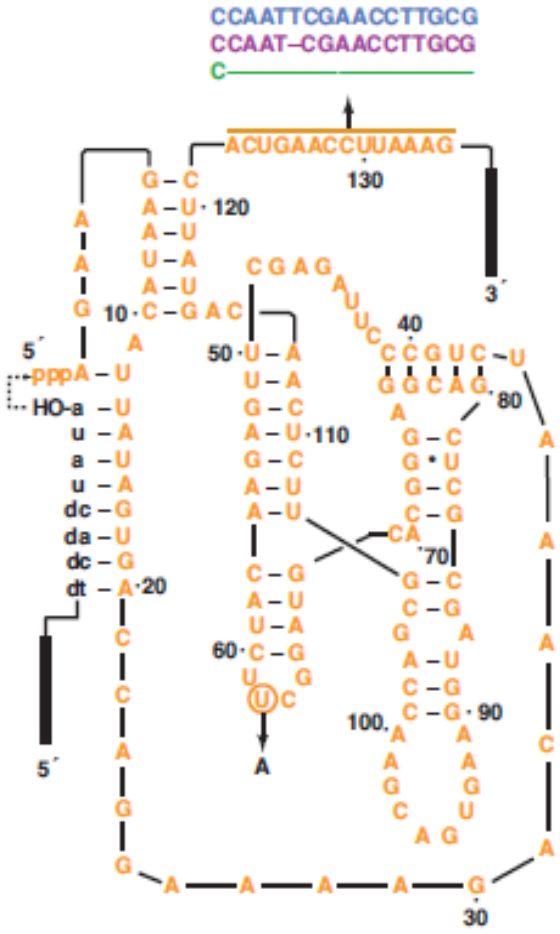


Figure 3.11. Mutants evolved during the high mutation rate experiments. (A) The structure of B16-19 wildtype is shown in orange letters; with the mutations of the master sequences indicated in black (fixed in the population), green (MS3), purple (MS1) and blue (MS2) capital letters. The substrate of the ligase is shown in black lower case letters, with the ligation reaction site shown by a dashed line arrow.

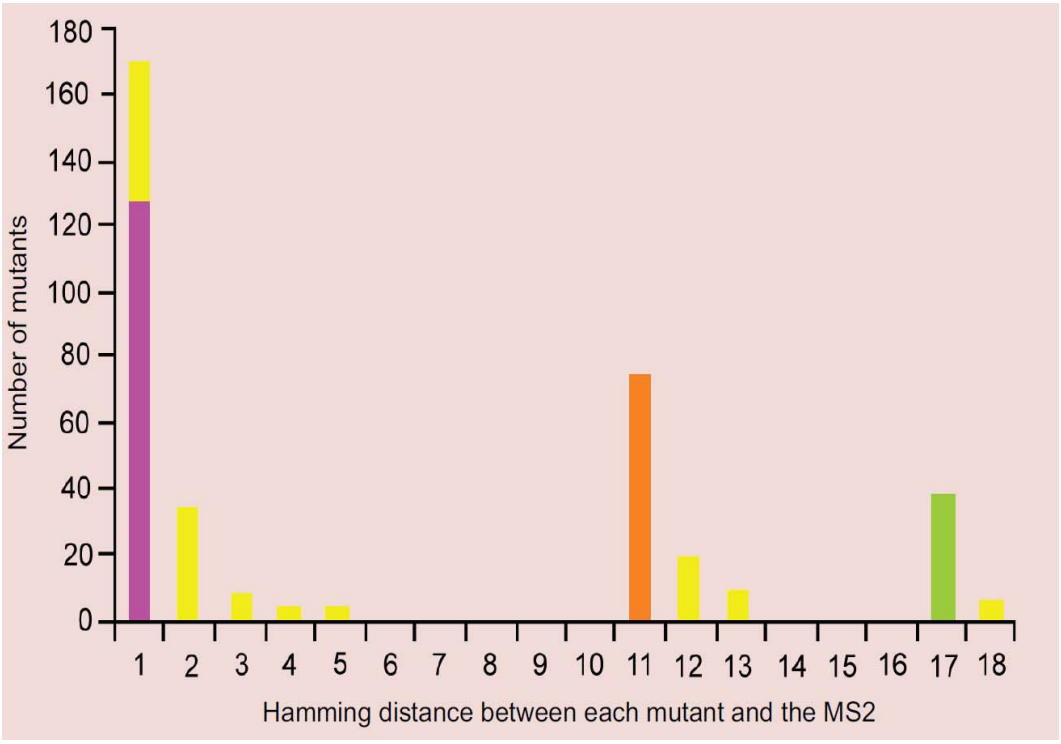


Figure 3.12. Hamming distance between the mutants and MS2. The *x*-axis shows the Hamming distance of the mutants relative to that of MS2, the most abundant sequence. The *y*-axis indicates the number of mutants that have the observed Hamming distance. Note that most of the mutants are separated by a short Hamming distance. A distance of one is observed between most of the mutations generated, meaning that the mutants are close together in sequence space. The colors of the bars follow the color code on the quasispecies networks diagrams.

Quasispecies	6H5	6H35	6H50
Master sequence (MS)	Wt	MS1	MS2
Frequency of MS (%)	77	80	45
Sample size (sequences)	102	98	108
Number of unique genotypes	16	14	36
Shannon Diversity Index	1.49	1.32	2.06

Table 3.3. Summary data of the quasispecies observed in lineage 6H. The quasispecies name is given on the top row and the name and frequency of the master sequence (MS) in the following rows. Note that the frequency of the master sequence is in general over 50%, and in the case of 6H50 where the frequency is slightly lower it is due to a transition between MS1 to MS2 (see figure 3.10C). The sample size exceeds the calculated number of expected unique genotypes and thus the diversity inference (Shannon index, last row) made from the sample are accurate representation of the population. The Shannon diversity index fluctuates considerably with time (min.= 1.32; max.= 2.06) most likely as a result of the fluctuation in master sequence abundance and identity.

Quasispecies	B22	B23	B39	B42
Master sequence (MS)	MS3	MS1	MS2	MS2
Frequency of MS (%)	68	56	77	62
Sample size (sequences)	41	48	53	42
Number of unique genotypes	8	8	10	13
Shannon diversity Index	1.15	0.74	0.46	0.73

Table 3.4. Summary data of the quasispecies observed in lineage 6L. The quasispecies name is given on the top row and the name and frequency of the master sequence (MS) in the following rows. Note that the frequency of the master sequence never drops below 50%. The sample size exceeds the calculated number of expected unique genotypes and thus the diversity inferences (Shannon index, last row) made from the sample is an accurate representation of the population. The Shannon diversity index calculated indicate that the diversity of the population fluctuates considerably with time (max.= 1.15 and min.= 0.46), as expected based on the change in master sequence abundance and identity throughout the evolution of the lineage.

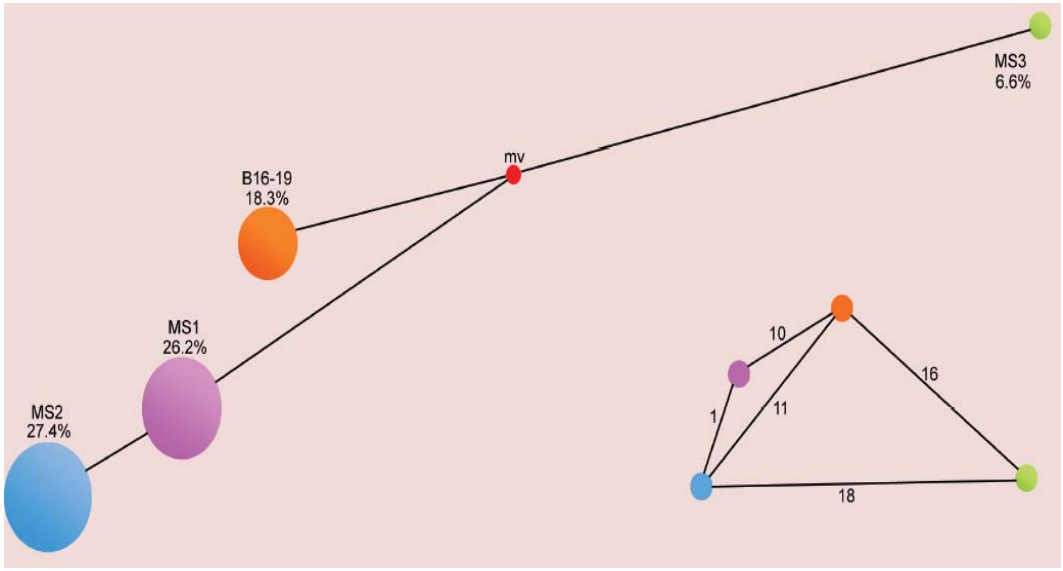


Figure 3.13. Relationship between the master sequences (MS) observed. The master sequences are shown following the same and color code. They are separated by their hamming distance and their frequency is indicated. The inset shows the hamming distance among all the master sequences, to represent their level of connectedness.

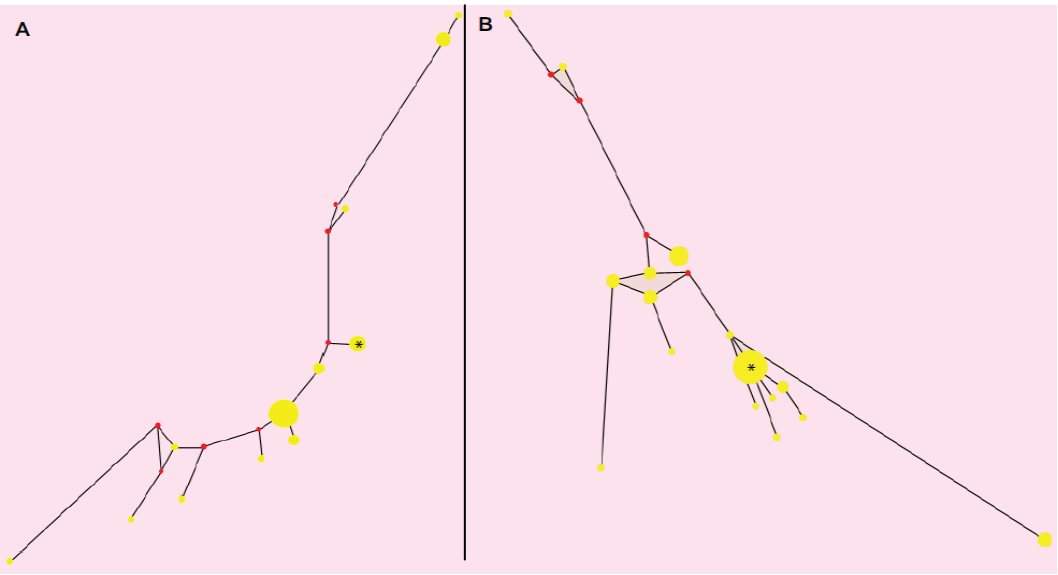


Figure 3.14. Network diagram of a non-quasispecies lineage. This lineage (3D, table 2.1) is the smallest that survived to 50 bursts, without added $MgCl_2$ (Soll, *et al*., 2007). Sequences from bursts 10 and 50 were used to draw the networks diagrams presented in (A) and (B) respectively. The general characteristics of quasispecies presented in lineages 6H and 6L (Figures 3.10 and 3.11), are not present in this network. Although, different clouds are formed in each burst, there is no formation of dominant master species. The genotype indicated by the asterisk (*) is the same in both bursts.

CHAPTER FOUR

FITNESS COMPONENTS IN RNA POPULATIONS EVOLVED *IN VITRO*

(In preparation for *Evolution*)

Background

Differences in selection along the evolutionary time course of a population can cause changes in the mean value of a phenotypic trait; thus selection differentials can be used to estimate the relationship between fitness and phenotypic traits (Falconer, 1981). The mean change in a phenotypic trait depends on the effects of selection directly on that trait and in correlated traits (for example as a consequence of pleiotropic effects), or to linkage disequilibrium (Pearson, 1903). Also, the magnitude and effect of selection can change from one part of the life-cycle to another, making them difficult to measure in wild populations. Thus, to measure selection corresponding to separate stages in the life-cycle, a multiple-level analysis of fitness variance can be used (Arnold and Wade, 1984). Examples of how a partial regression of the relative fitness of a particular trait provides an estimate of the effect of selection on that trait can be found in Lande (1979) and Lande and Arnold (1983).

Variance in the fitness values of phenotypic traits is necessary for selection to occur, and this effectively allows evolution to happen. To calculate the change in the mean phenotypic value of a trait after *k* episodes of selection, it is

necessary to know the change in the frequency distribution of the phenotypic value and in the fitness of the trait with each round of selection. If W_k is the absolute fitness of an individual at the k^{th} round of selection, the total absolute fitness (W) of an individual can be equated to the multiplicative effect of the absolute fitness of all the episodes (Arnold and Wade, 1984), as follows (2):

(2) $$W = \prod_{k=1}^{m} W_k$$

The separation of the total fitness of an individual into different fitness components allows a better assessment of which selective agents have a more direct impact on the success of an individual. The components of fitness that are important for an individual are chosen according to the particular case; what is biologically meaningful for one organism may not be for other. Darwin may have first separated fitness into fitness components when he studied sexual dimorphism in the plumage of birds. The plumage, for example, serves as an attractive means to females and thus it enhances the potential of mating. At the same time, colorful plumage carries the trade-off that the males are more easily spotted by predators. (Conner and Hartl, 2004).

Although measurements of the various fitness components and their multiplicative effect on the total life time fitness of organisms are commonly used, this approach has rarely, if ever, been applied to molecular populations.

Few molecular evolution experimental techniques allow the molecules to evolve “naturally”, with minimal addition of external selection pressures. In continually evolving populations of molecules, which mimic the evolution of organisms in the wild, the total lifetime fitness can be separated into components as well. In the continuous evolution *in vitro* (CE) system, for example, the total fitness of the individual molecules is affected by the rate of all the individual enzymatic reactions participating in the amplification process of the molecules.

The CE system (Wright and Joyce, 1997) uses catalytic RNA molecules as the subject of study because the effects of mutations acquired during the evolution cycles can be directly measured in their phenotype (Kun, *et al*., 2005; Langhammer, 2003; Lehman, *et al*., 2000; Joyce, 1989). As cycles of evolution occur, mutations arise and accumulate, having differential effects on the phenotypic traits of the ribozymes. Some mutant ribozymes eventually become abundant enough to allow their detection by direct sequencing, rendering *a posteriori* measurement of the change in fitness value of the evolved individuals possible. Mutations arise in the CE system as a result of enzymatic mistakes made during the amplification process of the molecules. The CE system (Figure 1.2) is composed of three different enzymatic steps:

1) The first is the ligation step performed by an RNA enzyme. The ligase ribozyme catalyzes the reaction that occurs by attack of the 3'-OH of a *trans* substrate onto the 5'-phosphate of the ligase itself with consequential formation of the respective phosphordiesther linkage. The ligase-substrate complex formed carries a promoter for the T7 RNA polymerase.

2) The second is the reverse transcription, performed by MMLV reverse transcriptase (RT), an RNA-dependant DNA polymerase. This enzyme uses the RNA ligase as a substrate to synthesize complementary DNA copies. Copies of ligases reacted (ligase-substrate complex) and unreacted (ligase *per se*) are both made by the RT. This enzyme is error-prone because it lacks a 3'-5' proof-reading ability. Therefore as cycles of evolution occur, mutations accumulate … unless there is a mechanism to alleviate them.

3) The third is transcription, performed by T7 RNA polymerase (RNAP), a DNA-dependant RNA polymerase. This enzyme synthesizes RNA copies that are complementary to the cDNA strand, when it recognizes the promoter sequence. As previously mentioned (Chapter 1, and Appendix I) the promoter sequence is located in the 3'-end of the exogenous substrate, consequentially, only active ribozymes posses this promoter necessary for the RNAP to initiate transcription. Inactive ligases are not recognized by this enzyme and they represent the end of their genotype's evolution (analogous to death).

To quantify the fitness of the ligases evolved during the CE cycles, we estimated the rates of reaction of the aforementioned three enzymatic steps. For the RNA ligation step we used the Eadie-Hofstee linearization method (Bisswanger, 2008) to measure k_{cat}, and for the RT and RNAP enzymes, we measured product formation over time. Each step constitutes a component of the total lifetime fitness of the ligase evolution. We now present the first systematic study of fitness components at the molecular level.

Materials and Methods

The rates of enzymatic reactions can be measured in the laboratory in different ways. One practical way is to measure the product formation over time. A plot of product formation over time is drawn and the rate can be inferred from the slope of the curve in simple cases. If the rate of the reaction is not linear, there are several methods that can be used to convert the data into its linear representation to gain a more clear representation of the true enzymatic rate. Also, many physical variables play important roles in the rates of reaction of enzymes. For this reason we, to the best of our abilities, measured the reaction rates under the same conditions that are used during the CE experiments. In the case of fast enzymatic reactions (e.g., RNA-

catalyzed RNA ligation) some variables such as temperature or pH can be varied in order to decelerate the reaction (Lewis, 1974).

We measured all the reactions using the CE buffer prepared as in Díaz Arenas and Lehman (2010a, see Chapter 2, and Appendix I). The buffer contains 25 mM $MgCl_2$ and a pH of 8.3. All enzymatic reactions were stopped using a quench solution consisting of 2X acrylamide gel dye (0.05% bromophenol blue and 40% sucrose) and 25 mM Na_2EDTA (2,2',2'',2'''-(ethane-1,2-diyldinitrilo) tetraacetic acid).

The ligases were prepared differently depending on which enzymatic rate was to be measured, but the protein enzymes were employed at the same concentration as they were during the CE experiments. This was 9.6U/μL for MMLV-RT (USB Corp., Cleveland) and 2.4U/μL for T7 RNAP (Ambion, Austin). The following are the details for measuring the kinetics of each enzymatic step:

Ligation kinetics

Preparation of the ligase

DNA products of all ligase ribozymes evolved during the CE were transcribed *in vitro* using Ambion T7 RNA polymerase 2U/uL, 5X rNTPs (2mM each at 1X), 1X transcription buffer (100 mM $MgCl_2$, 100 mM spermidine, 100 mM

DTT, and 100 mM Tris pH 7.5). The transcription products were electrophoresed on 8% polyacrylamide/8 M urea gels, and the bands were excised following a "Dip and Dot" procedure (Burton, *et al*., 2009). Ligase transcripts were adjusted to 10 μM using UV spectrometry at 260 nm by careful dilution. From this "ligase stock" a mixture (10 μL) was prepared for each ligase concentration that was going to be used for the kinetics assays. The amount of 3.33 X CE buffer was held constant (3 μL) while the amount of stock (1.0 μM) ligase and of DEPC-treated and/or RNase-free water (Ambion, Inc.) varied as the desired final concentration of ligase varied. Nine concentrations, from 0.05 μM to 1.0 μM were prepared for each of the six ligases: 0.050, 0.075, 0.100, 0.150, 0.200, 0.250, 0.500, 0.750, and 1.000 μM.

Preparation of the substrate

The substrate for the ligase ribozyme, S-163, is a DNA/RNA chimera that has a T7 promoter (5'-CTTGACGTCAGCCTGGACTAATACGACTCACUAUA-3'). For the ligase kinetics, 1μM substrate was radiolabeled at the 5'-end using 1.26 μM ^{32}P-γ-ATP (2,280Ci/mmol) with 1X OptiKinase reaction buffer from USB Corp. (0.5M Tris-HCl (pH 7.5), 10mM $MgCl_2$, 5mM dithiothreitol DTT), and 0.5 U/μL OptiKinase^{TM} enzyme (USB Corp., Cleveland) The reaction was incubated for 1h at 37 ºC. Radiolabeled substrate was run on a 15% polyacrylamide/8 M urea gel for 2000 V*h and the bands were excised following standard "Dip-and-Dot" procedure. The product was calibrated for

concentration with a UV spectrophotometer at 260 nm. This stock substrate was used to prepare the substrate mix for the ligase kinetics, containing 10 nM radiolabeled S-163 and 1X CE buffer.

Ligation reactions

The speed of the ligation reaction between each ligase concentration and the substrate was measured by performing a series of six time point reaction samples from 5 sec to 2 min in a V-bottomed 96-well plate. The substrate mix was placed in row number 1 of the plate (the “reaction-well”) in all columns, being each one (from A to L) used for a different ribozyme concentration. One plate per enzyme was used. Quench solution was added to the “quench-wells” from row numbers 3 to 8. The timer was set up for 2 min and 10 sec. When the timer reached 2 min, the ligase mixture was added to the “reaction-well” in a 1:1 volume ratio with the substrate mixture. This ensures a [$MgCl_2$] of 25 mM and at least 10-fold enzyme concentration relative to the substrate. For each time point, an aliquot of the ligase-substrate mixture was removed from the reaction-well and was added to the quench-well in a ratio of 1:1, to get a [EDTA] of 25 mM.

The reaction products were electrophoresed on 8% polyacrylamide/8 M urea gels for 500V*h to separate the reacted ligases (187-nt fragments) from the

residual substrate (35-mer) fragments. Each time point was loaded in a single well of the gel, and all the time points for the same enzyme concentration were loaded in the same gel to account for potential gel irregularities that may affect the readings. The gels were exposed to a phosphor screen overnight and scanned in a phosphorimager. The fraction of product was quantified using ImageQuant software. Plots of the fraction reacted vs. time were drawn using KaleidaGraph v. 4.1 (2009, Synergy Software). The data were fitted following the equation $y=A(1-e^{-kt})$, where A is the asymptote and k is the slope of the curve. The k_{obs} obtained for each enzyme concentration were plotted following a modified Eadie-Hofstee method (Ordoukhanian and Joyce, 1999). The k_{obs}/[ligase] vs. k_{obs} plots were drawn using Microsoft Excel v. 11.5612.5606 and the k_{cat} and K_m values were obtained from the y-intercept and the slope of line, respectively.

MMLV-RT kinetics

RNA preparation

Ligases were transcribed *in vitro* as described above under the ligase kinetics section. The transcripts were diluted for approximately the same concentration. The reverse transcription primer was radiolabeled following the same procedure that was used to radiolabel the substrate S-163. The radiolabeled primer was run in a 20% polyacrylamide/8 M urea for 2400V*h, and the band was excised following a standard Dip and Dot procedure

(Burton, *et al*., 2009). The concentration of the homogenized primer product was measured with UV photometry at 260 nm. The concentration of the radiolabeled primer was calibrated by careful serial dilution to 0.3 μM.

Reverse transcription rate measurement

The clean RNA ligases were mixed with a dNTP mix (dATP, dTTP, dCTP, dGTP are each at 25 mM at 1X), 0.6 μM radiolabeled primer (3,521Ci/mmol), and 1X CE buffer. The reaction rate was measured by taking time points of the reaction from 3 to 21 minutes. The timer was set up to 25 minutes. At minute 24, 9.6U/μL MMLV-RT (USB Corp.) enzyme was added to the reaction vessel. At minute 21, the first aliquot was taken from the reaction vessel and quenched. In the same manner, every three minutes an aliquot is drawn from the reaction vessel and quenched into the respective time point quench solution. In total seven time points were taken.

The reverse transcription products from each time point were loaded in a different well, and electrophoresed using an 8% polyacrylamide/8 M urea gel. The gel was developed overnight in a phosphor screen. One gel was prepared for the reverse transcription of each ligase. The gels were scanned in a phosphorimager and product formation was quantified using ImageQuant software. Plots of the product formed vs. time were drawn using KaleidaGraph v. 4.1 (2009, Synergy Software). The data were fitted following the equation

$y=A(1-e^{-kt})$, where A is the asymptote and k is the slope of the curve. The slope of this fit was obtained for each ligase evaluated, and use as an estimation of the rate. The units are in "concentration of phosphorimager band" *per* minute, but are given as *per* minute in table 4.5B and in the Results and Discussion section, for simplification.

T7 RNAP kinetics

The rate of the forward transcription reaction was measured for each ligase ribozyme, by mixing clean PCR product with 1X rNTP mix (UTP, CTP, GTP, and ATP each at 2mM), 1X CE buffer, 0.1μM ^{32}P-γ-ATP (6,000Ci/mmol). The rates were measured by taking time points as done for the reverse transcription rate measurements. Time points were taken every three minutes for a period of 21 minutes, which is equivalent to the length of one burst of RNA production during the CE experiments. Once the timer reached minute 24, 2.4U/μL of enzyme T7 RNA polymerase (Ambion, Inc.) were added to the reaction vessel. Time points were taken by removing an aliquot from the reaction mix and quenching it in a 1:1 quench solution as described under measurement of the reverse transcription step.

The reaction products were loaded onto an 8% polyacrylamide/8 M urea gel and electrophoresed for about 1400V*h. The gel was developed in a phosphor

screen overnight and scanned in a phosphorimager. The bands were quantified with the software ImageQuant. Plots of the product formation vs. time were drawn in Excel v. 11.5612.5606 and the rate obtained from the slope of the line. The units are in “concentration of phosphorimager band” *per* minute, but are given as *per* minute in table 4.5C and in the Results and Discussion section, for simplification.

Multiple fitness components

The rates of the ligation, the forward transcription, and the reverse transcription reactions were used as fitness component of the ligases. To calculate the total absolute fitness value for each ligase, a multiplicative procedure following Arnold and Wade (1984) was used. The total relative fitness value was calculated as a rate of each ligase observation over the ligase with the highest observed value. Total absolute and relative rates were statistically processed using box plot method in Excel v. 11.5612.5606.

Results and Discussion

Ligases evolved during the CE

During the evolution of the ligase B16-19, new ligases emerged and persisted long enough in the population to be detected by direct nucleotide sequence analysis. During the evolution experiments in which a low mutation rate was

used, two ligases emerged with enough abundance to be chosen for our fitness assays. These two ligases have been previously evaluated for their k_{cat} values (Soll, *et al*., 2007). During the evolution experiments in which the mutation rate was increased by adding Mn^{2+}, three new ligases emerged. In total six ligases, including B16-19 (the "wildtype"), were evaluated for their fitness profiles. The name of the ligases is given by the line and burst in which they were originated from. The following is a short description of each ligase with its respective name:

1) **B16-19**: The ligase ribozyme (Figure 1.1A) used to seed the CE experiments. It was selected to initiate the lineages because it has a high catalytic rate, as previously measured (13 min^{-1}, Soll, *et al*., 2007) and hence a probable high fitness value. Mutations that arise in the system are more likely to have a deleterious effect on the reaction rate of the ligase and its fitness value would decrease (Lehman, 2004). As mutations accumulate, the mean fitness of the population can decrease, potentially driving the population to extinction. The sequence at the 3'-end of this ligase **5' (120) UCA CUG AAC CUU AAA G (135) 3'** was found to contain the main variations among the following mutant ligases.

2) **A27**: A ligase (Figure 2.6) evolved during the experiments that employed no added mutagen. This ligase is characterized by a long polyA tail and has been

observed in lineages that became extinct as a consequence of the onset of Muller's Ratchet and further mutational meltdown (Figure 2.3). This ligase actually represents a group of mutants with polyA tails on the 3'-end of the ligase that can have different lengths (Figure 2.5B), from a few nucleotides to several (or even over a hundred). For example, an A5 mutant has five additional *A*s, as follows: **5' (120) UCA CUG AAC CUU AAA AAA AAG (140) 3'**. The rate constant of a mutant (A23) from this group was previously measured (6 min^{-1}; Soll, *et al*., 2007) and found to have a deleterious effect on the population. The polyA ligase characterized in this study has 27 additionally As.

3) **B34**: This is a ligase (Figure 2.6) that also emerged during the no added mutagen studies. This ligase in contrast to the polyA, was never observed in lineages that went extinct. B34 was observed in lineages that evolved for 50 cycles or nearly so (Figure 2.5A and 2.5C). The rate constant of this ligase has been previously measured (18 min^{-1}; Soll, *et al*., 2007) and found to be an advantageous mutation. This ligase is characterized by the following sequence variation at the 3'-end **5' (120) UCA AAU CGA AG (130) 3'**, and a nucleotide substitution at the position 62 (U62$\rightarrow$A).

4) **6L22**: This ligase evolved during the high mutation rate experiments (Figure 3.8B). It is the least abundant of all the ligases evolved during these

experiments (Table 3.4) and is usually observed at the beginning of the lineages, usually being displaced out of the population by other ligases (Figure 3.9 and 3.10). It is characterized by a short sequence deletion (Figure 3.11) at the 3'-end of the ligase **5' (120) UCC (122) 3'**. This ligase also has the U62→A mutation.

5) **6H35**: This ligase evolved during the high mutation rate studies (Figure 3.8B), as well. The sequence at the 3'-end of the ligase is **5' (120) UCC CAA UCG AAC CUU GCG (137) 3'**. It has the mutation U62→A (Figure 3.11). This ligase has a relatively high abundance (Table 3.3) compared to other ligases evolved during the mutagenic experiments (Figure 3.9 and 3.10).

6) **6H50**: It is a highly dominant ligase in the populations evolved with high mutation rate. It usually emerges towards the end of the evolution path of the populations (Figure 3.9 and 3.10) and rapidly increases in abundance (Table 3.3). This ligase has a sequence similarity to 6H35, with only one nucleotide difference (Figure 3.11) **5' (120) UCC CAA UUC GAA CCU UGC G (138) 3'**. It has the U62→A mutation, as well.

These aforementioned ligases represent the most abundant ligases that have been observed in the evolution experiments, as a result of directional selection (e.g., B34), strong random drift (e.g., polyA mutants), or quasispecies (6L22,

6H35, 6H50). Although many other ligases have evolved during the CE, they were not selected for kinetic studies because none of them had a relative high abundance in the bursts studied. Therefore, the catalytic rates of these ligases serve as good references for the incidence of each catalytic step on the total absolute fitness of the ligases.

Estimating the rate of the enzymatic reactions of the CE

The three enzymatic-step components of ligase fitness were measured under environmental conditions very similar to the CE system. The ligation reaction was performed at a lower temperature to facilitate an accurate measurement of the rate by manual pipeting. Previous studies have shown that B16-19 catalyzes 13 turnovers *per* minute (Soll, *et al*., 2007), and pipeting 10 to 20 times per minute is not only very challenging but may also render inaccurate estimates of the rate of reaction. The lower temperature results in an underestimated value of the ligation rate compared to the rate during the CE experiments. Still, this measurement allows for a comparison among the rate constants of the various ligases evolved in the system.

Catalytic rate of the ligation

Given the fast nature of the ligase reaction, we carefully chose concentrations to measure the kinetic constants. Based on preliminary data (not shown), nine

concentrations of ligases were chosen below the upper bound limit of 1.0 μM. Seven time points of product formation were taken from the reactions and the k_{cat} values obtained from the slope of the curve of a modified Eaddie-Hofstee plot trace for each ligase (Figure 4.15).

B34 ligase has the highest k_{cat} value, with 18 min^{-1}.(Table 4.5A). This value coincides exactly with the one published previously (Soll, *et al*., 2007). This result is not surprising because this ligase has been detected in lineages that survived to burst 50 (Table 1.1, B34 = M; Figure 2.5A, and 2.5C) or nearly so. This mutant evolved in population sizes of 600 and 3000 molecules. These population sizes are large enough to allow for natural selection to operate. B34 has an advantageous catalytic rate that puts it ahead in the first step of the reaction cycles.

B16-19 ligase has a k_{cat} value of 11 min^{-1}; very close to the previously measured value (13 min^{-1}; Soll, *et al*., 2007). These two values are comparable, and given that in this study we used more time points to calculate the intercept, we choose to use this value in subsequent analyses. Although, the rate of the reaction of B16-19 is not as fast as B34, this is still a very fast reaction and it explains why B16-19 is a highly recurrent ligase during *in vitro* evolution experiments (Lehman, 2004; Díaz Arenas and Lehman, 2009).

Ligase A27 was found to have the lowest rate constant value, with 0.85 turnovers per minute (Table 4.5A). This value shows that this ligase has a big disadvantage compared to other ligases in the population, and supports the idea that this ligase in fact has a deleterious effect on the population. The deleterious effect of the polyA mutants becomes more acute as the number of *A*s in the 3'-end increases. For example, the rate of reaction of a previously measured mutant with 23 extra As is of 6.6 min^{-1} (Soll, *et al*., 2007). The more *A*s are added to the 3'-end of the ligase the more distorted its structure may become, affecting the folding and catalytic function of the ligase.

The polyA mutants were detected only in populations that went extinct and that were of small effective size (Table 1.1, Figure 2.3, and 2.5B). The strong effect of random drift and the slow rate of these polyA ligases can set up the population for extinction, because at such low rates of ligation it is very likely that the reverse transcription step starts with ligases that have not performed the ligation reaction, and therefore their lineage is doomed to an end at the transcription step, as the RNA polymerase will not find the promoter sequence (located in the substrate of the ligase) necessary for the transcription step. Additionally, the extra *A*s added to these mutant ligases may have a deleterious effect on the population as they can cause a depletion of ATP nucleotides from the pool which could otherwise be used by other ligases.

Ligases 6H35 and 6H50 have rate constant values of 13 min^{-1} and of 10 min^{-1} (Table 4.5A), respectively. These values are very close to the value of B16-19 and perhaps explain why these mutants could sustain themselves for long time and emerge in various lineages studied. A fast ligation reaction is crucial to ensure that most - if not only - ligase-substrate complexes are reverse transcribed. Definitely, these two reaction rates demonstrate that at the first stage of the evolution cycles, ligases 6H35 and 6H50 are as competitive as the wildtype B16-19.

Finally, ligase 6L22 has a rate constant of 1.6 min^{-1} (Table 4.5A). This low value may be the cause of the relatively low abundance of these ligase compared with ligases 6H35 and 6H50; the other ligases also evolved during the high mutation rate experiments. The contexts under which a mutant emerges affect its fate, and in this case the “slow” 6L22 ligase emerged in populations that developed a quasispecies structure. This ligase emerged early in the lineages and was quickly replaced by other ligases that rapidly became highly abundant.

In general the values for the first component of fitness have a large standard deviation –SD (mean=9.1, S.D.=6.7, ratio=0.74). This result shows that the rate of the ligation reaction is a contributing factor in the total absolute fitness of the ribozymes. Ligase A27 and 6L22 present very low values; located below

the first quartile (Figure 4.16A). B34 present the highest value; located above the third quartile. The other three ligases are located relatively close to the median and/or third quartile.

Catalytic rate of the reverse transcriptase

We measure the catalytic activity of the reverse transcription step for each ligase (Figure 4.17) and found that it has differential effects on each ligase's fitness (S.D.= 0.20). Ligase B16-19 has the highest value, with 0.43 min^{-1} (Table 4.5B). This ligase gets reverse transcribed at a faster pace than the other ligases which sets B16-19 at an advantage over all of the other ligases for use of the NTP resources in the next step; forward transcription. Ligase B34 follows with 0.40 min^{-1}, a high value as well. The other ligases have lower values given in decreasing order as follows: 6L22, 6H35, A27 and 6H50 with 0.26 min^{-1}, 0.11 min^{-1}, 0.086 min^{-1}, and 0.085 min^{-1}, respectively (Table 4.5B). These values seem variable; however ligases 6L22, 6H35 and 6H50 maintain competitiveness during the evolution cycles, as observed by the quasispecies clouds. The value of ligase 6H35 and 6H50 are similar, but the value of 6L22 is almost twice their value. 6L22 have a low value of ligation reaction compared to 6H35 and 6H50, so its higher value of reverse transcription acts as a "compensation" allowing 6L22 to stay in the competition with 6H35 and 6H50 for the dNTP and perhaps NTP resources. Ligase A27 has the same

value of reverse transcription rate as 6H50 but the difference in the ligation reaction (compare 0.85 min^{-1} to 10 min^{-1}, respectively) is big.

Importantly, ligase A27 has a reverse transcription value of 0.086 min^{-1} (Table 4.5B), slightly higher than its ligation rate. Although the ligation rate was measured at a lower temperature, the actual value in the CE system would be only slightly higher than the reverse transcription rate, or even the same as more As are added (compare to the ligation rate of A23 in Soll, *et al*., 2007). If that happens, the reverse transcription and the ligation values would be negatively correlated, because a faster reaction rate for the reverse transcription step than for the ligation step will likely cause most ligases with polyA tracks to be reverse transcribed before they ligate the substrate to itself, or the interaction with MMLV-RT with the ligase may impede completion of its ligation reaction and the substrate may not become covalently attached to the ligase. Substrate-ligase complexes that are only attached by hydrogen bonding (Figure 1.1B) are instable and can separate, causing the ligase to be copied into a cDNA without the substrate, and hence not being recognized by the T7 RNAP and not being transcribed.

The values of the ligases for the second component of the fitness have a large S.D. (0.16, mean=0.23; ratio=0.70). This large fluctuation indicates that this component of fitness has a differential incidence in the total absolute fitness of

the ribozymes. Ligases 6H50 and A27 have the lowest values, being the only ones located below the first quartile (Figure 4.16B). 6L22 and 6H35 have values between the first and third quartile, and ligases B34 and B16-19 have values above the third quartile.

Catalytic rate of the forward transcription

The forward transcription reaction was measured for each ligase as substrate of the T7 RNA polymerase (Figure 4.18) and found to be different for each one. The transcription step has a measurable incidence on the total lifetime fitness of the ligases. B34 has the highest rate of forward transcription, with 134 min^{-1} (Table 4.5C). This ligase is transcribed more quickly than any other ligase present in the population, which confers it with an advantage over the other ligases. Ligase B16-19 also has a high transcription rate value, 116 min^{-1} (Table 4.5C), which agrees with the fact that this ligase is highly recurrent (Lehman, 2004). Ligase 6L22, A27, 6H35 and 6H50 have values of 110 min^{-1}, 88.3 min^{-1}, 81.2 min^{-1}, and 78.4 min^{-1} (Table 4.5C), respectively. The values of A27, 6H35 and 6H50 are relatively close and may not determine a differential total fitness of these ligases. If the total fitness were measured only as estimated by this component, the observed abundance of these ligases would have been quite similar as well. Ligase 6L22 has a higher rate of transcription compared to 6H35 and 6H50, this higher rate gives 6L22 an advantage in the competition with 6H35 and 6H50 for resources such as NTPs.

The values for the third component of the fitness are also variable among the ligases (Figure 4.16C), indicating its incidence in the total absolute fitness of the ligases. The variation in the values (S.D.=22.2, mean=101, ratio=0.219) is less compared to the variation in the reverse transcription and ligation values (Figure 4.16). This result indicates that perhaps transcription may have a slightly lower incidence in the total fitness of the ligases.

Absolute lifetime fitness

We have discussed the effect of each component on the survivorship and and/or procreation ability of each ligase. In this section, we calculate the multiplicative effect of the fitness components in the total absolute and relative fitness of each ligase (Arnold and Wade, 1984). The relative fitness values facilitate the comparison of the different scales of the reaction rates (Figure 4.19, Table 4.5). The total absolute lifetime fitness of each ligase was calculated by multiplying the fitness components of the ligase (Figure 4.19). The total relative lifetime fitness of each ligase was calculated as the rate between its value and the value of the ligase with the highest total absolute fitness. The total relative lifetime fitness of the ligases obtained from the reaction rates measured is as follows:

From the relative fitness values (Table 4.5) calculated, we found that B34 has the highest value of relative fitness (1.0), this indicated that ligase B34 is an advantageous mutants and explains why the populations in which B34 was observed did not become extinct. The total fitness value of this ligase is relatively far from the distribution of the total fitness of the all other ligases measured (Figure 4.19). Ligase B16-19 follows with a high fitness value as well, 0.57, located also above the 75% percentile (Figure 4.19). Ligase 6H35, 6H50 and 6L22 emerged during the high mutation rate experiments and have values of total relative fitness of 0.12, 0.07, and 0.05, respectively. It is interesting that the abundance at which these ligases are observed do not correspond with the fitness values. Ligase 6H35 has the highest total fitness value of the three and although it was observed at high abundance in the populations its abundance was not as high as the one at which ligase 6H50 was observed. Ligase 6L22 was the least abundant of all the three and has the least total fitness value of the three. The populations in which these ligases were observed showed a dynamic equilibrium (Chapter 3) in which one ligase is replaced by other and then yet again by another one. The quasispecies phenomenon has being observed in association with the survival of the flattest phenomenon (Sardanyés, *et al*., 2008; Codoñer, *et al*., 2006; Sanjuán, *et al*., 2007; Comas, *et al*., 2005; Wilke, *et al*., 2001). During survival of the flattest the individual present at high abundance do not have high fitness values, and their fitness values do not differ significantly. These ligases, evolved during the

high mutation rate experiments, as in the case of the survival of the flattest experiments, have low total fitness values compared to the wildtype B16-19, also their fitness values are relatively close to each other (less than 10 units different among each other, but 60 units different when compared to B16-19).

Ligases A27 has a low value of total relative fitness, 0.01, as well. This ligase emerged in populations evolved at low mutation rate, which were observed to become extinct. In contrast, the populations evolved at high mutation rate in which a survival of the flattest phenomena may be in effect, never were observed to become extinct. In the populations were A27 was observed, the wildtype was replaced by a series of mutants with low fitness values, called polyadenylations (polyAs). A27 has 27 additional As but mutants with polyadenylation tracks of up to 150 has been observed in these populations. It is likely that the deleterious effect of the polyAs in the population mean fitness becomes stronger as the number of As increases (compare to the ligation rate of A23, Soll, *et al*., 2007).

The total fitness values of the ligases do not indicate a particular advantage that each ligase may have at the different stages of the evolution cycle. For example, ligase B34 has a lower reverse transcription rate than B16-19 but its ligation rate is faster (Figure 4.20A), this sets up B34 at an advantage over B16-19. Additionally B34 has a faster rate than B16-19 during the transcription

step, which gives B34 a definite advantage over the wildtype. Ligase 6H50 does not compete as well in the reverse transcription step compared to the 6H35 and 6L22 (Figure 4.20A), but it has a relatively close value to 6H35 for the ligation and transcription rates, which help 6H50 compete with 6H35 for resources. 6L22 ligase and A27 have relatively similar profiles, with 6L22 having larger values for all the rates. These two ligases have a negative correlation between the ligation and the reverse transcription. In general, each of the components of fitness studied has profound implications in the total absolute fitness of the ligases. The ligation and transcription components of fitness are the strength of B34, H35 and H50 (Figure 4.20A and 4.20B, left panel). The reverse transcription component is the strength of B16-19, while the transcription step is the strength of A27 and 6L22 (Figure 4.20B, right panel).

Evolutionary implications

It is important to realize that the total relative fitness values discussed above have a large deviation (S.D.=0.43; mean=0.33; ratio=1.3) because we are considering all the ligases together. These ligases were neither present in populations simultaneously, nor did they arise in the same evolutionary studies (Figure 4.21). Thus, these values should not be considered as an approximation of the mean fitness of the population because a number of other mutants were present in the same bursts where these ligases appeared.

The fitness values of these mutants were not studied because none of them were present abundantly enough, but together they represent a good fraction of the mean fitness.

This study demonstrates how different components of fitness can differentially affect the fate of the ligases, as it happens *in vivo* with organisms. Different fitness components can evolve to become more important under different environmental conditions.

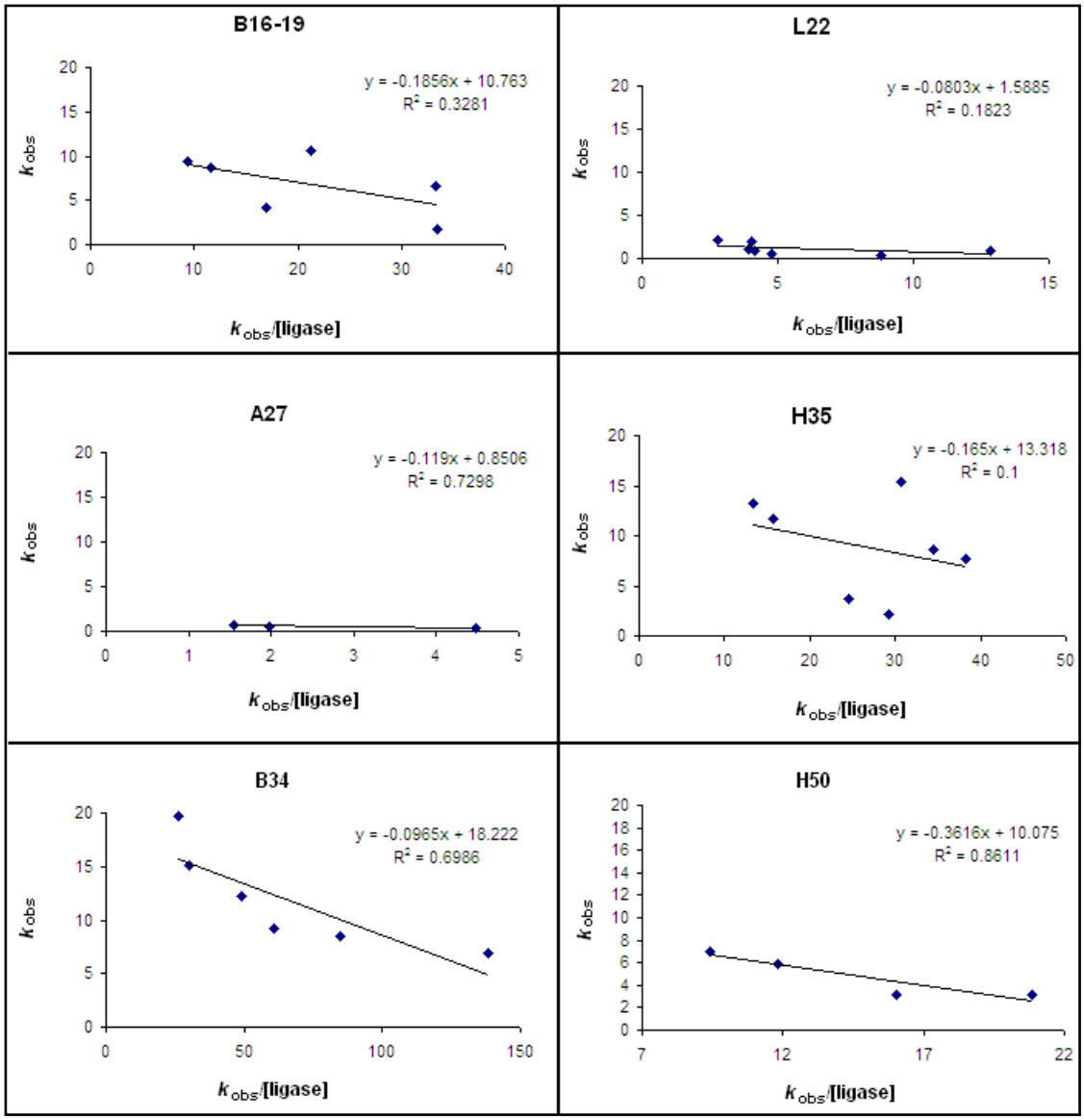


Figure 4.15. Plots for the rate of the ligation reaction. The ligation reaction was measured for each ligase, the product reacted per time plots were linearized using Eaddie-Hofstee method. The k_{cat} estimate of the rate is obtained from the slope of the curve of the k_{obs}/k_{obs}.

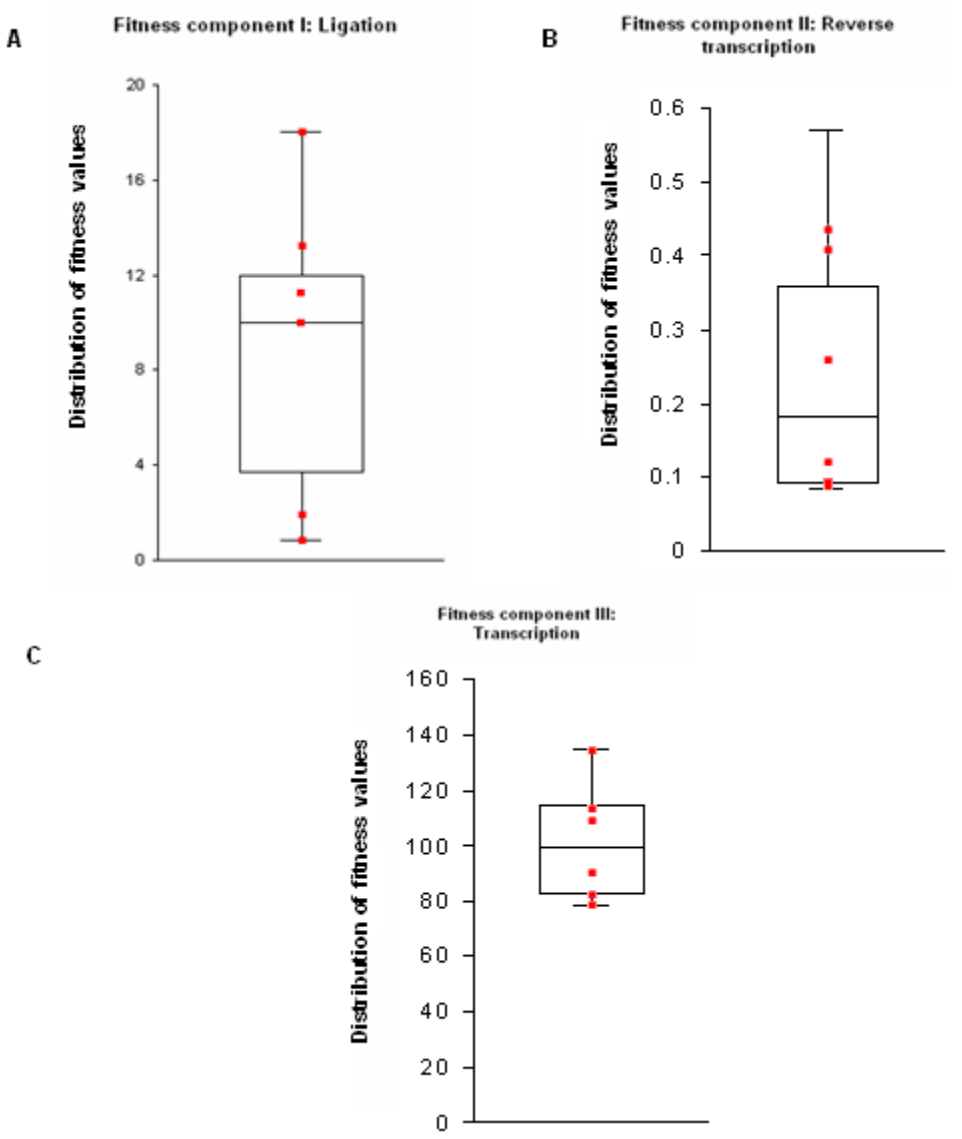


Figure 4.16. Box plot diagrams for each of the fitness components. (A) Depicts the distribution of the absolute fitness values obtained for the ligation reaction, (B) for the reverse transcription reaction, and (C) for the transcription reaction. The location of each ligase in the distribution is shown with a red dot. The measurement of the rates was different, for the ligation reaction the rate is in turnovers *per* minute, for the reverse transcription and transcription reactions the rate is in amount of product formed *per* minute; consequentially the distribution of the absolute fitness values (*y*-axis) appear in different scales.

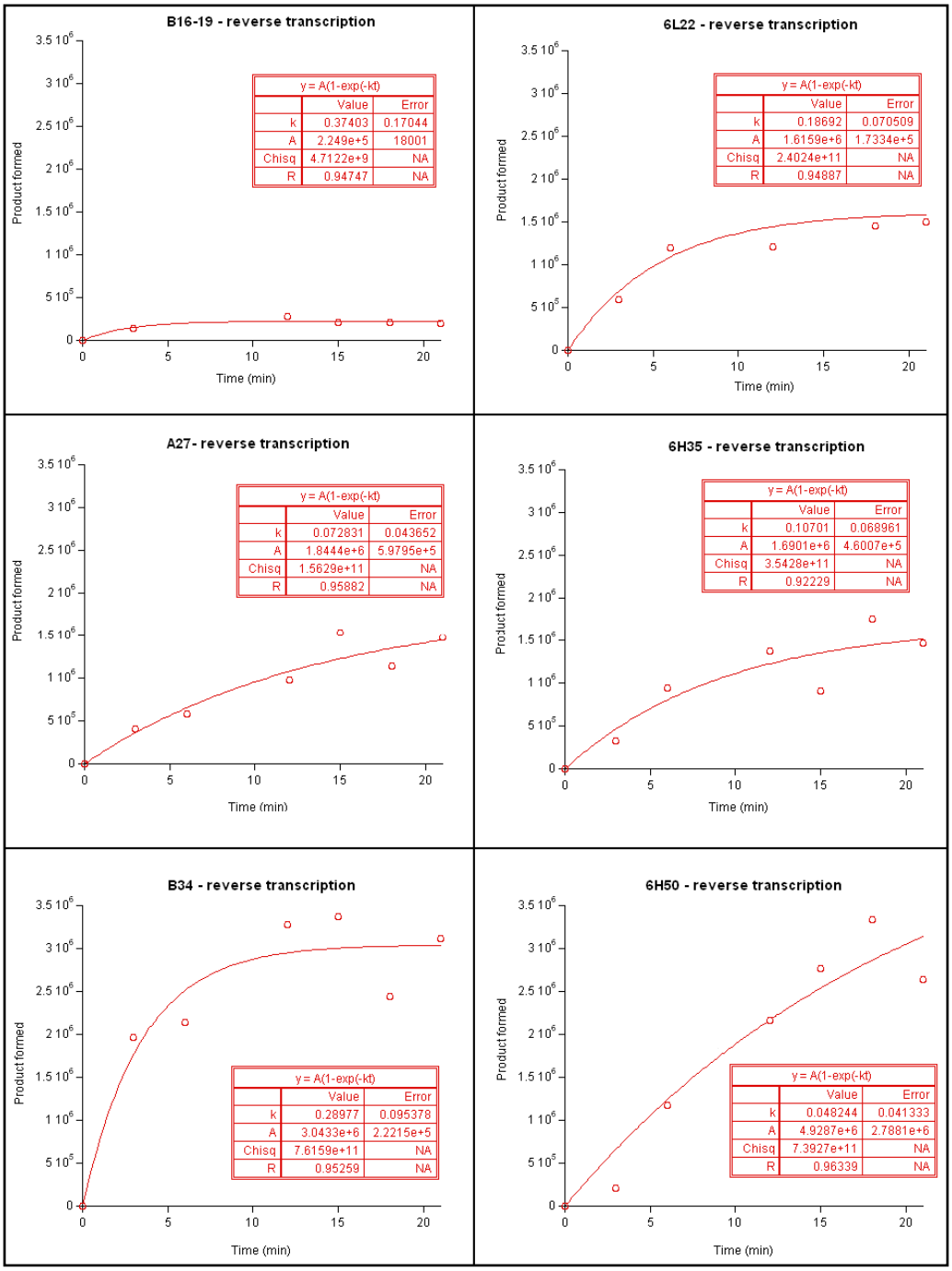


Figure 4.17. Plots for the rate of the reverse transcription reaction. This enzymatic step was measured, as the amount of product formed over time, for each ligase. The estimate of the rate is obtained from the slope (k) of the plot, using $y=A(1-\exp^{(-kt)})$, where *k* is the slope and *A* is the asymptote. The *k* values were calibrated for the slight variation in the ligase concentration used for the input experiment.

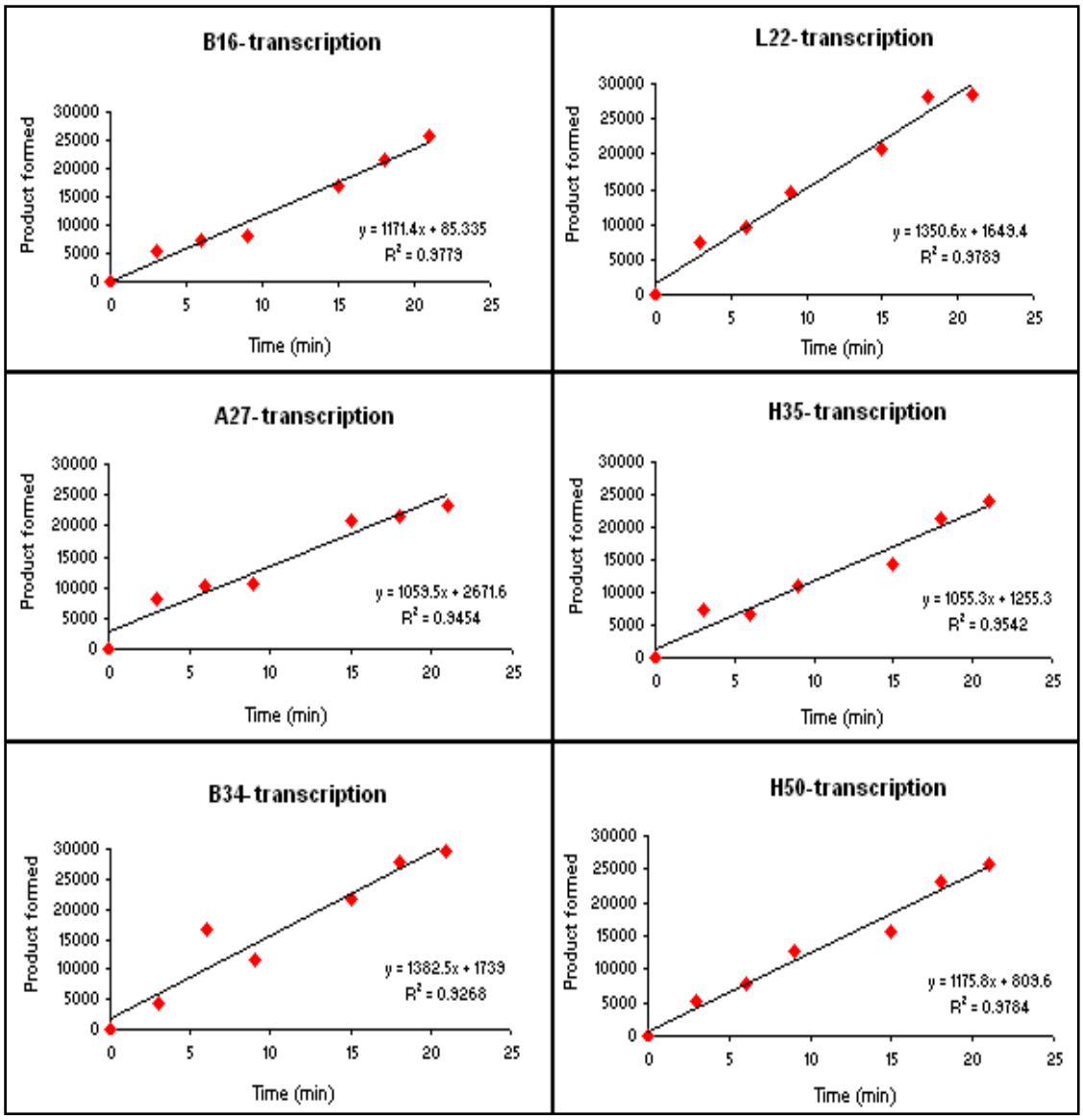


Figure 4.18. Plots for the rate of the transcription reaction. The enzymatic step was measured, as the amount of product formed over time, for each ligase. The estimate of the rate is obtained from the slope of the line and calibrated for the slight variation in the ligase concentration used for the input experiment.

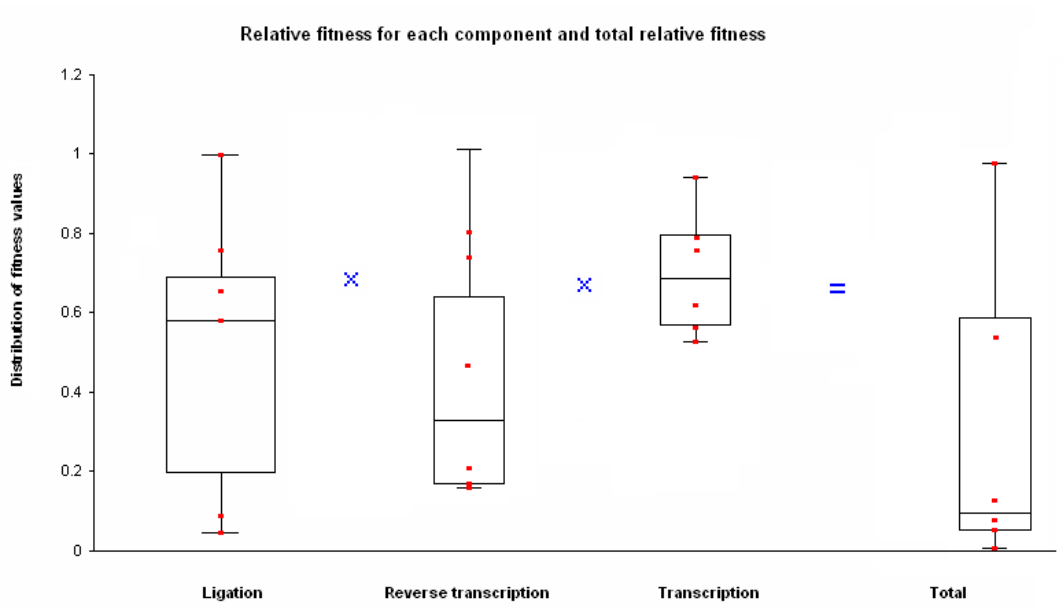


Figure 4.19. Box plot of relative fitnesses, for each component and the total. Box-plots are presented here in relative values to facilitate comparison. The total relative fitness value is calculated by the multiplicative effects of the relative fitness of each fitness component (*x*-axis). The variations in the fitness distribution (*y*-axis) demonstrate that the ligases have a variable usage of each component of the fitness (ligation, reverse transcription, and transcription) and hence different values of total relative fitness.

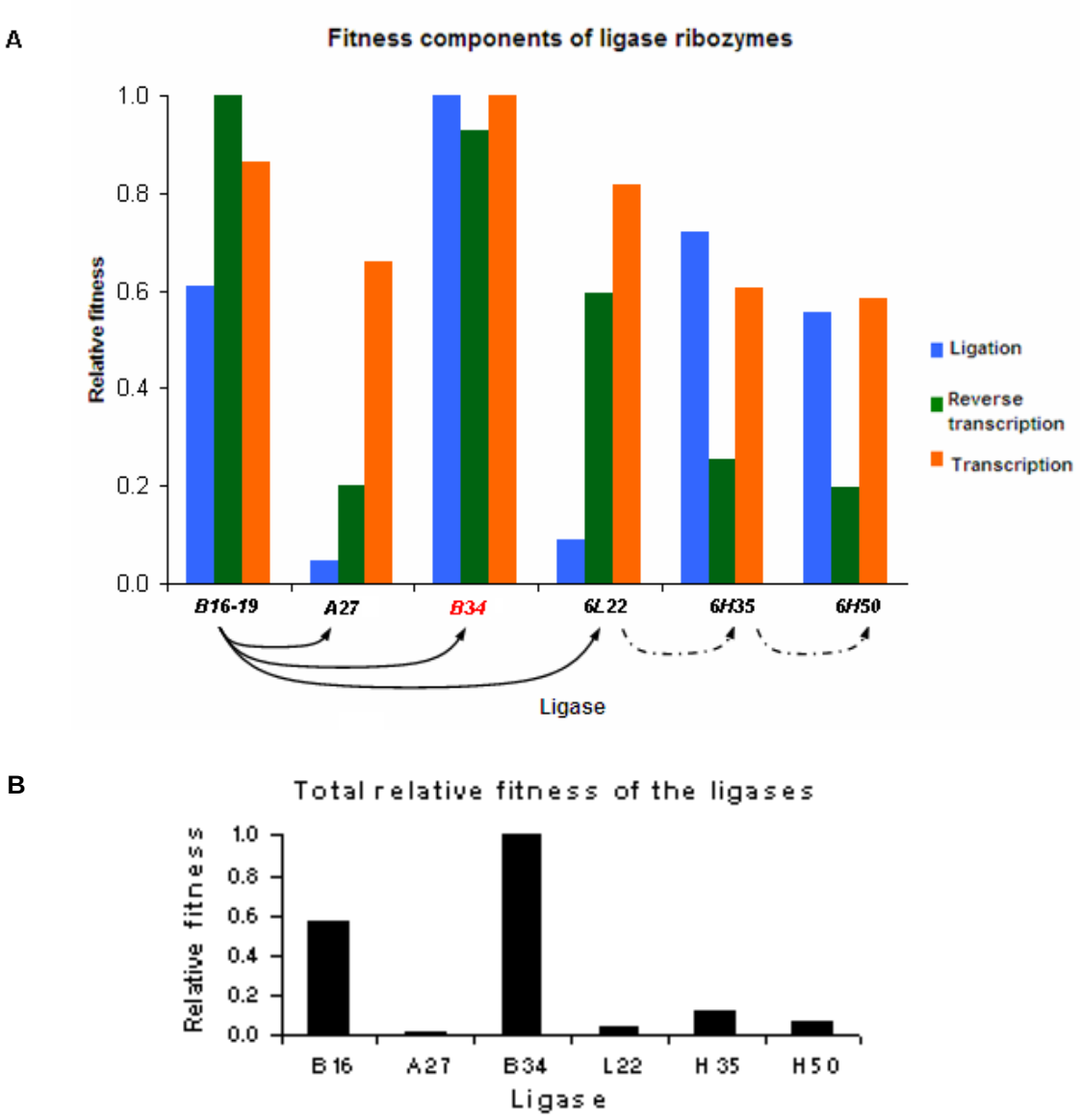


Figure 4.20. Component of fitness of the ligase ribozymes. (A) The three enzymatic steps of ligation, reverse transcription, and transcription were measured for each ligase (*x*-axis). The strength of each fitness component over the relative fitness (*y*-axis) is depicted here by the high of the columns. Notice that different ligases have particular strength in one of the fitness components. (B) The total relative fitness value obtained by the multiplicative effect of the components of fitness in A. For easy comparison, the values of total fitness for all the ligases are relative to the ligase B34 that has the maximum value measured.

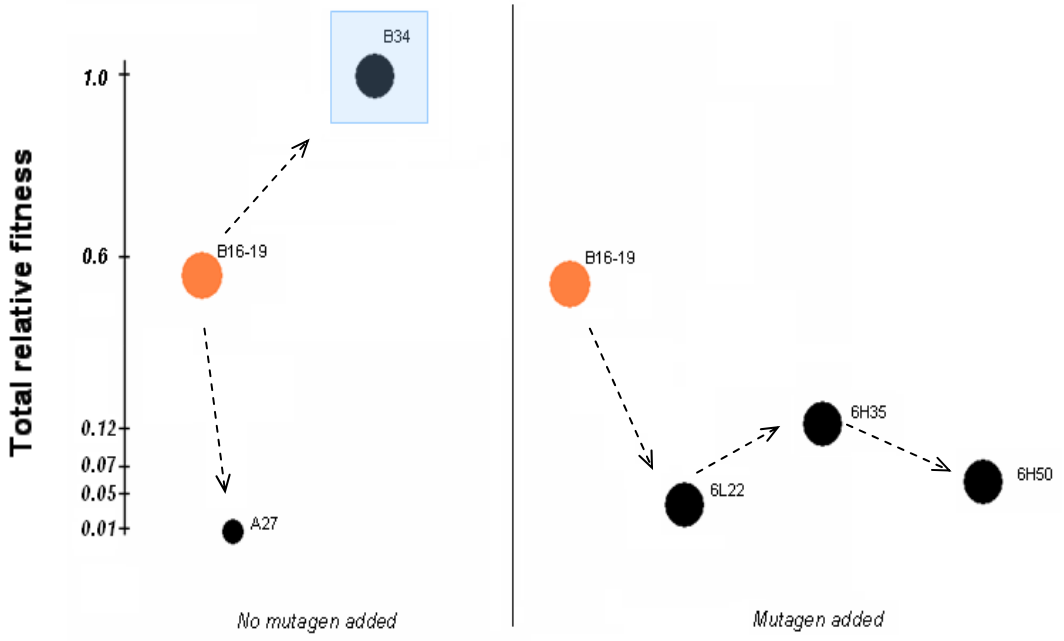


Figure 4.21. Fitness of the ligases and their relationship in evolutionary time. The evolution experiments start with a clonal population of B16-19 ribozymes, depicted in orange. As evolution cycles go (*x*-axis), mutant ligases emerge with different total fitness values (*y*-axis). Polymorphic populations (black) can persist through time or extinct depending on the mean fitness value and the mutation rate. Ligases A27 and B34 evolved from B16-19 in experiments with no Mn (II) added to the reaction vessel, whereas the evolution of 6L22, 6H35 and 6H50 occurred with Mn (II) added to the reaction vessel. The exact evolutionary path followed by an emergent mutant from the wild type is not know, but the relationship of these mutants in time is known, and depicted here as break lines arrows. Two evolutionary histories are shown, one in which evolution experiments were carried out under no mutagenic pressure, and other in which the evolution experiments were carried out under added mutagenic pressure ($MnCl_2$). The blue area highlights that this population is of 3000 molecules, and not of 100 molecules as all others. Notice population were polyA mutants (e.g., A27) are abundant, is decreasing in size, and are at risk of extinction, due to the low mean fitness value of the population, which keeps decreasing as adenylation tracks are extended. The mutants 6L22, 6H35 and 6H50 have relative low values of total fitness, but they emerged in population evolved at high mutation rate, in which a quasispecies were observed. This is perhaps evidence of the survival of the flattest, a phenomenon that has been show associated with quasispecies.

A

Enzyme	Absolute rate of ligation (min^{-1})	Relative rate
B16-19	11	0.61
A27	0.85	0.05
B34	18	1.00
6L22	1.6	0.09
6H35	13	0.72
6H50	10	0.56

B

Enzyme	Absolute rate of RT (min^{-1})	Relative rate
B16-19	0.43	1.00
A27	0.086	0.20
B34	0.040	0.93
6L22	0.26	0.60
6H35	0.11	0.26
6H50	0.085	0.20

C

Enzyme	Absolute rate of transcription (min^{-1})	Relative rate
B16-19	116	0.86
A27	88.3	0.66
B34	134	1.00
6L22	110	0.82
6H35	81.2	0.60
6H50	78.4	0.58

Table 4.5. Rate of the enzymatic reactions in the CE. (A) Ligation reaction, (B) reverse transcription (RT) reaction, and (C) transcription reaction. Each enzymatic reaction was measured for all the ligase enzymes (first column). The rates are presented in absolute values (central column), with the ligation values in turnovers *per* minute and the transcription and reverse transcription in amount of product formed *per* minute. To facilitate comparison, the rates are presented in relative values (third column), in relation to the ligase with highest value (1.00).

μ \ N_e	100 molecules (167ym/8.2μL)			3000 molecules (5zm/8.2μL)		
	Name	t_E *(b)*	E[t_E]	*Name*	t_E *(b)*	E[t_E]
MMLV-RT (1/30,000)	3Z	18		2I	36	
	4R	18		2K	37	
	4U	34		2A	45	
	● 4V	18	24.3	2J	49	44.5
	4W	30		3A	50+	
	4Y	28		● 3C	50+	
MMLV-RT + Mn^{2+} 6-30 fold higher than (1/30,000)	6E	50		6B	50	
	● 6H	50	50	6N	50	50
	6K	50		● 6O	50	
	● 6L	50		6P	50	

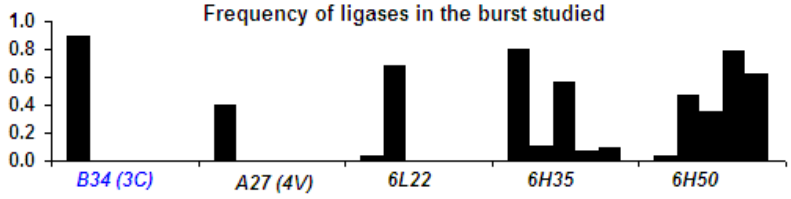


Table 4.6. Lineages evolved with the CE. The name of the lineages evolved with the CE are given in the first column of the body of the table, the observed time to extinction (t_E) is given in the second column and the mean of the extinction time is given in the third column as an estimate of the expected time to extinction (E[t_E]). The effective size (N*e*) of the populations is given in the top row of the table and the approximate mutational rate (μ) is given in the first column. These mutational rates are based on literature review (Ji and Loeb, 1992; El-Deiry, *et al*., 1984; Lazcano, *et al*., 1992; Vartanian, *et al*., 1999). The lineages that were studied for population diversity have a red dot. From those lineages mutants arose that have a high relative abundance in the populations and that have evolutionary implication for the fate of the lineages they were present in. At the bottom of the table there is a frequency bar diagram that shows the frequency at which these mutants appeared in the different lineages that they were observed. Each bar represents a burst evaluated. Overall it can be seen that the abundance of these mutants in the different burst studied is over 40%. The name of the mutants is given according to the lineage and/or burst in which they emerged, with the exception of A27. Note that B34 is the from a 3000-molecule population as opposed to the other mutants that are from a 100-molecule population.

CHAPTER FIVE

Conclusions

Mutations can act as a major evolutionary force leading to a vast diversity of forms and functions in organisms, which can then explore many available niches. In contrast, mutation can also be a major cause of population extinction if the genetic load generated by the accumulation of mutations is high and the random sampling of genetic variants is strong. Given this double-edged sword, I wanted to understand how the accumulation of mutations may have acted on the survival of populations at the origins of life, where high diversity may have been required to explore and create novel functions that could have lead to cellular life, but the small size of the populations posed a risk of extinction via Muller's Ratchet.

Many current organismal populations that have experienced bottlenecks are at risk of extinction due to the high genetic load imposed by the accumulation of mutations in synergism with the effect of random drift (reducing genetic variability) on their small effective population sizes. The genetic deterioration caused each generation in a ratchet-like mode (Muller, 1950) can eventually cause the extinction of the populations by a mutational meltdown (Lynch, 1993). The problem in studying genetic deterioration in organisms and the potential for extinction is that the interplay of random genetic drift, effective population size, and Darwinian natural selection can become fairly

complicated. There are at least two issues here: (1) Epistatic effects, linkage disequilibrium, and hitchhiking, can all interfere with our elucidation of the speed of genetic deterioration via Muller's ratchet; and (2) The length of the generational time can constitute a major problem because the fixation of new mutants takes certain amount of time depending on their selective advantage (Danforth, 1923; Haldane, 1935; Kimura, 1983). Advantageous mutants, for example, take 4Ne generations to become fixed, so if the generation time of the organisms is long, the requirement to study many generations and individuals to elucidate potential mechanisms that allow populations to overcome Muller's ratchet (*e.g*. via fixation of advantageous mutants) can be prohibitive.

So, in order to better our understanding of the effects of mutation accumulation and genetic load, different models have being used: *E. coli* (Kibota and Lynch, 1996), *Caenorhabditis elegans* (Estes *et al*., 2004), *Drosophila melanogaster* (Mukai, 1964), *Arabidopsis thaliana* (Schultz *et al*., 1999), *Saccharomyces cerevisiae* (Zeyl, *et al*., 2001), and RNA bacteriophage □6 populations (Chao, 1990), just to name a few. These model organisms have shortened generational times compared to larger organisms, but they still have complex genomes. Consequently, the use of populations of catalytic RNA molecules, not only allows us a better understanding of such phenomena, because of the absence of genomic interactions and short

generation times, but it also provides a test of how abiotic populations of RNA molecules at the origins of life could have faced or perhaps avoided altogether the onset of Muller's ratchet and eventual mutational meltdown.

In this research we used a catalytically proficient ligase ribozyme, called B16-19, and the continuous evolution in vitro (CE) methodology (Wright and Joyce, 1997). We use clonal populations of B16-19 to start the experiments because this ligase is already well adapted to the CE environment, and therefore most mutations in this sequence would be deleterious to either survival or reproduction (Joyce, 2004). We started the experiments with different population sizes and found that the small populations went extinct via Muller's Ratchet and mutational meltdown (Chapter 2), as predicted based on what happens in organismal populations. Larger populations were able to persist as a consequence of the presence of advantageous mutants (Figure 2.6, Table 2.1, Chapter 2). Although in the CE system, the individual ligases can develop advantages at different steps of their "life" cycle (analogous to animals that can take advantage of variable mating abilities, evading predators, *etc*.), the ligases that emerged in the smaller populations have detrimental effects on the mean fitness because they have very low fitness values (Table 4.5, Figure 4.20, Chapter 4). In contrast, the ligases evolved in the large populations have developed fitness advantages in all the stages of the cycle compared to B16-19 (Table 4.5, Figure 4.20, Chapter 4).

This set of experiments provides a clear correlation between the effective population size and the average time to extinction due to genetic deterioration caused by the accumulation of mutations. Our small populations of 100 molecules went extinct at an average of 24.3 cycles, the populations of 300 molecules at 29 cycles, the 600-molecule populations at 36.4 cycles and the 3000-molecule populations at 44.5 cycles (Figure 2.4, Chapter 2). Because there are no complex genomic interactions (linkage disequilibrium, hitchhiking, etc) in our population of RNA molecules, we could directly measure the effect of random drift and population size on populations of naked genes. We can conclude that populations of 600 molecules or less struggle to survive under the pressure of random drift and accumulation of mutations; and this is a large population if we locate ourselves at the origins of life (ca. 4 billion years ago).

The stability of an RNA sequence largely depends on its secondary structure through the number and types of non-covalent bonding interactions. It has been proposed (Lehman *et al*., 2000; Kun, *et al*., 2005; Holmes, 2005) that the secondary structure of catalytic RNAs served as a mutational buffer at the origins of life. Through epistatic interactions manifest through RNA motifs such as bulges, stems, loops, and pseudoknots, these intramolecular forces could strongly stabilize the RNA structure such that a major number of mutations have a neutral effect on fitness, and therefore are tolerated. Kun *et*

al. (2005) have calculated the actual error rate of the hairpin and *Neurospora* VS ribozymes based on phylogenetic comparison of the known secondary structures. The error rate that they calculate suggests that ribozymes could have potentially increased enough in sequence length, through "epistatic" cooperation of secondary structural motifs, as to give rise to the first ribo-organism with a genome size long enough to code for as many as 70 genes of a tRNA size (about 7 Kb in total).

If this is the case, the RNA populations should be able to tolerate a relative large number of mutations, compared to populations of polymers that are less flexible and cannot acquired the secondary structure arrangements of RNA molecules. We tested an increased mutation rate during the replication of our B16-19 ligase populations. There were two potential routes our populations could have taken: (1) a purifying selection scenario, in which as in the previous experiments (Chapter 2) the populations persist if they are able to "purge" deleterious mutations (*e.g*., populations of 3000 and some of 600 molecules), if not, they become extinct via Muller's ratchet and mutational meltdown (*e.g*., populations of 100, 300 and some of 600 molecules); or (2) a survival of the flattest scenario, in which mutant molecules with lower fitness values are able to proliferate and be sustained in the population as a consequence of a change in the fitness landscape from one with a single high peak mandating the operation of purifying selection (Diaz Arenas and Lehman, 2009), to one

with a low mesa with many equally-most-fit genotypes. The survival of the flattest phenomenon has been observed in association with quasispecies structure and high mutation rates in viral, viroids and cellular automata (computer simulations of evolutionary population dynamics), but never in such simple populations such as our catalytic RNA populations.

Our results show that the RNA populations of 100 molecules evolved within the same environmental conditions as described earlier (Chapter 2) except with higher mutation rate (Chapter 3) did not actually go extinct at all, and that these ligases have develop a quasispecies population structure (Figure 3.9, Figure 3.10, Chapter 3). Furthermore, the most abundant ligases in the quasispecies observed (called the “master sequences”) have low fitness values compared to B16-19. Although the different master sequences have different advantages at one or other steps of the ribozyme life cycle, they in general have similarly low fitness values (Figure 4.20, Chapter 4) as the survival of the flattest model proposes. Our results demonstrate that quasispecies structure in association with the survival of the flattest phenomenon serves as a mechanism for small catalytic RNA populations to avoid Muller’s ratchet and mutational meltdown. This could have been a strategy that small (*e.g.* 100 molecules) abiotic populations at the origins of life could have exploited to persist and eventually approach a minimum genome size for more sophisticated functions.

References

Albert R, Jeong H, and Barabasi A. 2000. Error and attack tolerance of complex networks. *Nature,* 406:378-382.

Arnold S, and Wade M. 1984. On the measurements of natural and sexual selection: theory. *Evolution*, 38: 709-719.

Baltimore D 1970. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 226: 1209-11.

Bandelt H, Forster P, and Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.*, 16:37-48.

Bergman N, Johnston W, and Bartel D. 2000. Kinetic framework for ligation by an efficient RNA ligase ribozyme. *Biochemistry* 39: 3115–3123.

Bergman N, Lau N, Lehnert V, Westhof E, and Bartel D. 2004. The three-dimensional architecture of the class I ligase ribozyme. *RNA*, 10: 176–184.

Bisswanger H. 2008. Enzyme kinetics. Second edition. Wiley-VCH Verlag GmbH & Co. KGaA.

Borenstein E, and Ruppin E. 2006. Direct evolution of genetic robustness in microRNA. *Proc. Nat. Acad. Sci. USA.,* 103:6593-6598.

Bowie J, Reidhaar-Olson J, Lim W, Sauer R. 1990. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science,* 247:1306-1310.

Bull J, Ancel M, and Lachmann M. 2005. Quasispecies made simple. *PLoS Comput. Biol.*, 1:e61.

Burton A, Madix R, Vaidya N, Riley C, Hayden E, Chepetan A, et al. 2009. Gel purification of radiolabeled nucleic acids via phosphorimaging: Dip-N-Dot. *Anal. Biochem.*, 388: 351-352.

Carothers J, and Szostak J. 2006. *In vitro* selection of functional oligonucleotides and the origins of biochemical activity. In the aptamer handbook: functional oligonucleotides and their applications. Edited by Klussmann S. Weinheim: Wiley-VHC publisher, 3–28.

Cech T. 1987. The chemistry of self-splicing RNA and RNA enzymes. *Science*, 236: 1532-1539.

ChemDraw v. 10. 2005. www.cambridgesoft.com/software/ChemDraw/

Codoñer F, Daròs J, Solé R, and Elena S. 2006. The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. *PLoS Pathogens*, 2:e136.

Comas J, Moya A, Gonzalez-Candelas F. 2005. Validating viral quasispecies with digital organisms: A re-examination of the critical mutational rate. *BMC Evol Biol*, 5:5.

Conner J, and Hartl D. 2004. A primer of ecological genetics. Sinauer Associates, Inc.

Danforth C. 1923. The frequency of mutation and the incidence of hereditary traits in man. In: Eugenics, Genetics and the Family, Vol. 1, pp. 120-28.

Davies E, Peters A, and Keightley P. 1999. High frequency of cryptic deleterious mutations in *Caenorhabditis elegans*. *Science* 285: 1748–1751.

De Angioletti M, Rovira A, Sadelain M, Luzzatto L, and Notaro R. 2002. Frequency of mis-sense mutation in the coding region of a eukaryotic gene transferred by retroviral vectors. *J. Virol*. 76: 1991–1994.

Denver D, Morris K, Lynch M, and Thomas W. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* 430: 679–682.

Díaz Arenas C, and Lehman N. 2009. Darwin's concept in a test tube: Parallels between organismal and *in vitro* evolution. *Int J Biochem Cell Biol*, 41:266-273.

Díaz Arenas C, and Lehman N. 2010a. The continuous evolution *in vitro* technique. *Curr Protoc Mol Biol*, Unit 9.7.

Díaz Arenas C, and Lehman N. 2010b. Quasispecies behavior observed in catalytic RNA populations evolving in a test tube. *BMC Evol Biol*, in press.

Eigen M. 1971. Self-organization of matter and evolution of biological macromolecules. *Naturwissenschaften*, 58: 465.

Eigen M, Schuster P. 1979. The hypercycle. A principle of natural self-organization. Berlin: Springer-Verlag.

Eigen M. 2000. Natural selection: a phase transition? *Biophys. Chem*., 85:101-123.

Eigen M. 1993. The origin of genetic information: viruses as models. *Gene*, 135:37-47.

Eigen M, and Schuster P. 1977. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Die Naturwissenschaften*, 64:541-565.

El-Deiry W, Downey K, and So A. 1984. Molecular mechanisms of manganese mutagenesis. *Proc. Natl. Acad. Sci. USA.*, 81:7378-7382.

Ellington A, and Szostak J. 1990. *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, 346:818-822.

Estes S, Phillips P, Denver D, Thomas W, and Lynch M. 2004. Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in *Caenorhabditis elegans*. *Genetics* 166: 1269–1279.

Falconer D. 1981. Introduction to quantitiative genetics. Second edition. Longman, London.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics*, 78:737-756.

Fernandez G, Bonaventura C, and Martinez M. 2007. Fitness landscape of human immunodeficiency virus type 1 protease quasispecies. *J. Virol*., 81:2485-2496.

Fisher R. 1930. The genetical theory of natural selection. Oxford: Oxford University Press.

Fluxus technology ltd. [http://www.fluxus-engineering.com]

Gabriel W, Lynch M, and Bürger R. 1993. Muller's ratchet and mutational meltdown. *Evolution* 47: 1744–1757.

Gesteland RF, Cech TR, and Atkins JF. 2006. The RNA World. Third edition. Cold springs Harbor. Cold Springs Harbor Laboratory Press: 49-77.

Gilbert, W 1986. Origins of life: the RNA World. *Nature* 319: 618.

Gould S. 1989. Wonderful life: The Burgess shale and the nature of history. W. W. Norton and Co. Publishers. Pp 347.

Haldane J. 1929. The origin of life. *Ration*. *Ann*. 3: 3–10.

Haldane J. 1935. The rate of spontaneous mutation of a human gene. *J. Genet.*, 31: 317–326.

Haldane J. 1937. The effect of variation on fitness. *Am*. *Nat*. 71: 337–349.

Hanczyc M, and Dorit R. 2000. Replicability and recurrence in the experimental evolution of a group I ribozyme. *Mol. Biol. Evol*., 17:1050-1060.

Hayden E, and Lehman N. 2006. Self-assembly of a group I intro from inactive oligonucleotide fragments. *Chem. Biol*., 13:909-918.

Heineman R, Molineux I, and Bull J. 2005. Evolutionary robustness of an optimal phenotype: re-evolution of lysis in a bacteriophage deleted for its lysin gene. *J*. *Mol*. *Evol*., 61: 181–191.

Hill W, and Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet. Res*., 8:269–294.

Holmes E. 2005. On being the right size. *Nature Genet.*, 37:923.

Hu W, and Temin H. 1990. Retroviral recombination and reverse transcription. *Science,* 250: 1227–1233.

Jhaveri S, and Ellington A. 2002. *In vitro* selection of RNA aptamers to a small molecule target. *Curr. Protoc. Nucleic Acid. Chem.*, 9.5.1- 9.5.14.

Johns G, and Joyce G. 2005. The promise and peril of continuous *in vitro* evolution. *J*. *Mol*. *Evol*., 61: 253–263.

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

Joyce G. 1989. RNA evolution and the origins of life. *Nature*, 33: 217-224.

Joyce G. 2004. Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem*., 73:791-836.

Joyce, G. 2007. Forty years of *in vitro* evolution. *Angew*. *Chemie*, 46: 6420-6436.

Kauffman S. 1991. Antichaos and adaptation. *Scientific American*, 265:78-84.

Kauffman S. 1993. Origins of order: self-organization and selection in evolution. USA: Oxford University Press, Inc.

Keightley P, and Eyre-walker A. 2000. Deleterious mutations and the evolution of sex. *Science,* 290: 331–333.

Kibota T, and Lynch M. 1996. Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. *Nature,* 381: 694–696.

Kimura M. 1983. The neutral theory of molecular evolution. Great Britain. Cambridge University Press.

Kun A., Santos M, and Szathmáry E. 2005. Real ribozymes suggest a relaxed error threshold. *Nat*. *Genet*., 37: 1008–1011.

Lande R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain: body size allometry. *Evolution*, 33:402-416.

Lande R, and Arnold S. 1983. The measurement of selection on correlated characters. *Evolution*, 37:1210-1226.

Langhammer U. 2003. Artificial evolution: the RNA model. Institute for theoretical chemistry and structural biology. University of Vienna, Austria. Seminar of January 13th.

Lazcano A, Valverde V, Hernandez G, Gariglio P, Fox G, and Oro J. 1992. On the early emergence of reverse transcription: theoretical basis and experimental evidence. *J. Mol. Evol.*, 35:524-536.

Lehman N, Delle Donne M, West M, Dewey G. 2000. The genotypic landscape during *in vitro* evolution of a catalytic RNA: implication for genotypic buffering. *J. Mol. Evol.*, 50:481-490.

Lehman N, and Joyce G. 1993. Evolution *in vitro*: analysis of a lineage of ribozymes. *Curr. Biol.*, 3:723-734.

Lehman N. 2004. Assessing the likelihood of recurrence during RNA evolution *in vitro*. *Artificial Life*, 10:1-22.

Lehman N, and Unrau P. 2005. Recombination during *in vitro* evolution. *J*. *Mol*. *Evol*., 61: 245–254.

Lehman N, and Wayne RK. 1991. Analysis of coyote mitochondrial DNA genotype frequencies: Estimation of the effective number of alleles. *Genetics*, 128: 405-416.

Lewis E. 1974. Investigation of rats and mechanism of reaction. Part I, v. VI. Third edition. Chapter I and VIII. Willey Interscience Publication.

Lozupone C, Changayil S, Majerfeld J, and Yarus M. 2003. Selection of the simplest RNA that binds isoleucine. *RNA*, 9:1315-1322.

Lynch M, and Conery J. 2003. The origins of genome complexity. *Science,* 302: 1401–1404.

Lynch M, and Gabriel W. 1990. Mutation load and the survival of small populations. *Evolution,* 44: 1725–1737.

Lynch M, and Walsh B. 1998. Genetics and analysis of quantitative traits. Sinauer, Sunderland, MA.

Lynch M, Burger R, Butcher D, and Gabriel W. 1993. The mutational meltdown in asexual populations. *J*. *Hered*., 84: 339–344.

Lynch M, Blanchard J, Houle D, Kibota T, Schultz S, Vassilieva L, and Willis J. 1999. Perspective: spontaneous deleterious mutation. *Evolution,* 53: 645–663.

McGinness K, Wright M, and Joyce G. 2002. Continuous *in vitro* evolution of a ribozyme that catalyzes three successive nucleotidyl addition reactions. *Chem. Biol*., 9: 585-596.

Maynard Smith J. 1964. Group selection and kin selection. *Nature*, 201:1145–1147.

Maynard Smith J, and Szathmáry E. 1999. The origins of life. From the birth of life to the origins of language. UK: Oxford University Press.

Montville R, Froissart R, Remold S, Tenaillon O, and Turner P. 2005. Evolution of mutational robustness in an RNA virus. *PLoS Biol.,* 3:1939-1945.

Muller H. 1950. Our load of mutations. *Am*. *J*. *Hum*. *Genet*., 2:111-176.

Muller H. 1932. Some genetic aspects of sex. *Amer. Nat*., 66:118–138.

Mukai T. 1964. The genetic structure of populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. *Genetics* 72: 335–355.

Muller H. 1950. Our load of mutations. *Am*. *J*. *Hum*. *Genet.,* 2: 111–176.

Negroni M, and Buc H. 2001. Mechanisms of retroviral recombination. *Annu*. *Rev*. *Genet*., 35: 275–302.

Nimwegen E, Crutchfild J, and Huynen M. 1999. Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA*., 96:9716-9720.

Ordoukhanian P, and Joyce J. 1999. A molecular description of the evolution of resistance. *Chem*. *Biol*., 6:881–889.

Orgel l. 1979. Selection *in vitro*. *Proc. Roy. Soc. Lond*., 205:435-442.

Paegel B, and Joyce G. 2008. Darwinian evolution on a chip. *PLoS Biol*., 6: 900-906.

Pearson K. 1903. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Phil*. *Trans*. *Roy*. *Soc*. *Lond*., 200: 1-66.

Pfrender M, and Lynch M. 2000. Quantitative genetic variation in Daphnia: temporal changes in genetic architecture. *Evolution,* 54: 1502–1509.

Random numbers [http://www.random.org].

Salehi-Ashtiani K, and Szostak J. 2001. *In vitro* evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, 414:82-84.

Sanjuán R, Moya A, and Elena S. 2004. The contribution of epistasis to the architecture of fitness in an RNA virus. *Proc*. *Natl*. *Acad*. *Sci*. *USA.,* 101: 15376–15379.

Sanjuán R, Cuevas J, Furió V, Holmes E, and Moya A. 2007. Selection for robustness in mutagenized RNA viruses. *PLoS Genet.*, 3(6): e93.

Santos M, Zintzaras E, and Szathmáry E. 2004. Recombination in primeval genomes: a step forward but still a long leap from maintaining a sizable genome. *J. Mol. Evol.*, 59:507-519.

Sardanyés J, Elena S, and Solé R. 2008. Simple quasispecies models for the survival-of-the-flattest effect: the role of space. *J. Theor. Biol.*, 250:560-568.

Schlosser K, and Li Y. 2005. Diverse evolutionary trajectories characterize a community of RNA-cleaving deoxyribozymes: a case study into the population dynamics of *in vitro* selection. *J. Mol. Evol.*, 61:192-206.

Schlosser K, Lam J, and Li Y. 2009. A genotype-to-phenotype map of *in vitro* selected RNA-cleaving DNAzymes: implications for accessing the target phenotype. *Nucleic Acids Res*., 37:3545-3557.

Schmitt T, and Lehman N. 1999. Non-unity molecular heritability demonstrated by continuous evolution *in vitro*. *Chem*. *Biol*., 6: 857–869.

Schultz S, Lynch M, and Willis J. 1999. Spontaneous deleterious mutation in *Arabidopsis thaliana*. *Proc*. *Natl*. *Acad*. *Sci*. *USA.,* 96: 11393–11398.

Shabarova Z, and Bogdanov A. 1994. Advance organic chemistry of nucleic acids. VCH editorial. Federal Republic of Germany.

Smith R, and Pereira-Smith O. 1977. Colony size distribution as a measure of age in cultured human cells. A brief note. *Mech*. *Ageing Dev*., 6:283–286.

Soll S, Arenas CD, and Lehman N. 2007. Accumulation of deleterious mutations in small abiotic populations of RNA. *Genetics*, 175:267-275.

Szathmáry E, and Maynard-Smith J. 1997. From replicators to reproducers: the first major transitions leading to life. *J. Theor. Biol*., 187:555-571.

Tagaki Y, and Yoshida M. 1980. Clonal death associated with the number of fissions in *Paramecium caudatum*. *J*. *Cell Sci*., 41: 177–191.

Tarasow T, Tarasow S, and Eaton B. 1997. RNA-catalyzed carbon-carbon bond formation. *Nature*, 389: 54-57.

Temin HM and Mizutani S. 1970. RNA-dependent DNA polymerase in virions of Rous Sarcoma virus. *Nature,* 226: 1211-13.

Tuerk C, and Gold I. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249:505–510.

Vartanian J, Sala M, Henry M, Wain-Hobson S, and Meyerhans A. 1999. Manganese cations increase the mutational rate of human immunodeficiency virus type 1 *ex vivo*. *J. Gen. Virol*., 80:1983-1986.

Voytek S, and Joyce G. 2007. Emergence of a fast-reacting ribozymes that is capable of undergoing continuous evolution. *Proc. Natl. Acad. Sci. USA*., 104:15288-93.

Wagner A. 2008. Robustness and evolvability: a paradox resolved. *Proc. R. Soc. B.,* 275:91-100.

Wallace B. 1987. Fifty years of genetic load. *J*. *Hered*., 78: 134–142.

Watson J, Hopkins N, Roberts J, Argetsinger J, and Weiner A. 1987. The molecular biology of the gene. Volume II. Fourth edition.

Wilke C. 2005. Quasispecies theory in the context of population genetics. *BMC Evol. Biol.*, 5:44.

Wilke C. 2003. Probability of fixation of an advantageous mutant in a viral quasispecies. *Genetics*, 163:467-474.

Wilke C, Wang J, Ofria C, Lenski R, and Adami C. 2001. Evolution of digital organism at high mutational rate leads to survival of the flattest. *Nature*, 412:331-333.

Wright M, and Joyce G. 1997. Continuous *in vitro* evolution of catalytic function. *Science*, 276:614-617.

Zeyl C, Mizesko M, and Devisser J. 2001. Mutational meltdown in laboratory yeast populations. *Evolution,* 55: 909–917.

Zhang B, and Cech T. 1997. Peptide bond formation by *in vitro* selected ribozymes. *Nature*, 390: 96-100.

Zuker M. 2003. mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., 31: 3406–3415.

Appendix: The Continuous Evolution *In Vitro* Technique

(Díaz Arenas and Lehman. 2010. *Curr. Prot. Mol. Biol.*, Unit 9.7.)

Background

The complex evolutionary dynamics that shape organismal populations can be mimicked in the laboratory using molecules as genotypic models and the continuous evolution *in vitro* (CE) technique (Wright and Joyce, 1997; Schmitt and Lehman, 1999; McGineness *et al*., 2002; Jonhs and Joyce, 2005; Soll *et al*., 2007; Voytek and Joyce, 2007; Joyce, 2007; Paegel and Joyce, 2008). *In vitro* experimentation has been used since the late 1960's to study how selective forces operate in individuals or group of individuals and how the populations evolve in response to those forces. Many approaches has been utilized, but most of these *in vitro* techniques aim for the search of specific, rare, and/or, improved functionalities, and they require constant intervention from the researcher. For a more extensive review, refer to the following papers: (1) Development of *in vitro* experimentation techniques: Saffhill *et al*., 1970; Orgel, 1979; Joyce, 1989, Guatelli *et al*., 1990. (2) Key experiments: Ellington and Szostak, 1990; Beaudry and Joyce, 1992; Lehman and Joyce, 1993; Bartel and Szostak, 1993; Tarasow *et al*., 1997; Ellinger *et al*., 1998; Seeling and Jäschke, 1999; Johnston *et al*., 2001; Lehman, 2004. (3) Review papers: Breaker and Joyce, 1994; Wilson and Szostak, 1999; Joyce, 2004; Joyce 2007, Díaz Arenas and Lehman, 2009; Ellington *et al.*, 2009.

The molecules that are suitable for in vitro selection and in vitro evolution experiments have the following two characteristics (1) they must have both an evolvable genotype and a distinct phenotype, such that they can really behave as organisms in term of selective and evolutionary forces (Joyce, 1989; Lehman et al., 2000; Ancel and Fontana, 2000; Kun et al., 2005), and (2) they must be able to store and transfer the genetic information (e.g., be readable). Such molecules with those two characteristics are the catalytic RNAs or ribozymes, which have a genotype constituted by a sequence of nucleotides and a phenotype constituted by the catalytic function codified in the sequence (Cech, 1987). The CE technique takes advantage of these ribozymes features, and with the incorporation of protein enzymes physically present in the evolution system, successive rounds of amplification and selection can be performed in a single reaction vessel prior to dilution into fresh reagents. Thus, The molecules that are suitable for *in vitro* selection and *in vitro* evolution experiments have the following two characteristics (1) they must have both an evolvable genotype and a distinct phenotype, such that they can really behave as organisms in term of selective and evolutionary forces (Joyce, 1989; Lehman *et al*., 2000; Ancel and Fontana, 2000; Kun *et al*., 2005), and (2) they must be able to store and transfer the genetic information (*e.g*., be readable). Such molecules with those two characteristics are the catalytic RNAs or ribozymes, which have a genotype constituted by a sequence of nucleotides and a phenotype constituted by the catalytic function codified in the sequence

(Cech, 1987). The CE technique (Fig. 1) takes advantage of these ribozymes the CE system is akin to the original extracellular Darwinian experiments of Spiegelman and colleagues (Mills *et al*., 1967; Levisohn and Spiegelman, 1969), in which serial transfers were performed to facilitate the ability of the protein enzyme Qβ replicase to effect the rapid evolution of Qβ RNA.

The more commonly used laboratory technique of stepwise evolution employs a protocol that requires frequent researcher intervention in order to carry on successive rounds of selection and amplification. The selection and amplification cycles occur in different reaction vessels even for a single generation. Specific selection pressures are commonly applied to the system in order to obtain molecules with a desired property, and often at the start of a new cycle or after a certain number of cycles, the selection pressures are intensified, favoring genotypes that can evolve with the desired properties. Generalized protocols for stepwise evolution are described elsewhere (Jhaveri and Ellington, 2002; Breaker, 2000; Davidson *et al*., 2009; Piasecki *et al*., 2009).

In contrast, CE offers the advantages of more rapid evolutionary progress with less physical intervention by the investigator. In CE, the selection pressures are determined by the evolving system itself, because the population evolves more “naturally” following the dynamics dictated by the mere interaction

among the individual molecules and a relatively constant environment. During a long lineage of CE, different molecular dynamics can emerge that affect the selection forces operating on the molecules, and consequently modeling the evolutionary pathway that the population follows. Therefore, if all the "ingredients" for the evolution are included in the reaction vessel and side products (*e.g.*, pyrophosphate) are frequently removed, the catalysis and selective amplification cycles can be performed for as long as the population is still evolving.

These aforementioned characteristics make the CE technique a better tool to parallel the evolutionary dynamics that occur in organismal populations. The current paper is dedicated to a step-by-step description of the CE experimental procedure that should guide the researcher who aims to better their understanding of the process and to exploit its use in molecular evolution laboratory experiments.

Characteristics of the CE system

The CE system (Fig. 1) can use a class I ligase ribozyme, as devised by Wright and Joyce (1997), or a DSL ligase ribozyme as devised by Voytek and Joyce (2007). These two RNA ligases (Fig. 2) have completely different evolutionary origins but they share an important characteristic required for the CE system: an extremely fast rate of self-ligation of an exogenous substrate.

Both ligases have been evolved and adapted to work effectively under the experimental conditions required for the CE. The class I ligase was the first to be adapted to CE, and a particularly fit genotype B16-19, with a catalytic rate constant of ~13 min^{-1}, has been used in several CE experiments (Soll *et al*., 2007). The DSL ligase (Ikawa *et al*., 2004) has more recently been adapted to work on the CE environment. An active derivative, T100-1 with a k_{obs} value of ~ 3 min^{-1}, has more recently been used in some CE experiments (Voytek and Joyce, 2007). These ligases are specifically able to catalyze the attack of the 3′-end hydroxyl group of an exogenous oligomer substrate on the α-phosphate of the 5′-end of the ligase ribozyme itself. This reaction is required for the CE system to work because the substrate contains the promoter sequence that the T7 RNA polymerase recognizes to initiate the transcription process (Fig. 1). This scheme determines an explicit selection process at the time that the ligase genomes are to be transcribed. cDNA of reacted ligases possess the promoter sequence and will pass to the next generation, while the cDNA of non-reacted ligases will not have the promoter sequence and therefore will not be transcribed, and will not proceed to the next generation.

After the ligation reaction has occurred, given that the ligation reaction occurs at a very high speed, a reverse transcription step is likely to occur. The reverse transcriptase (RT) that is already present in the reaction vessel will initiate from the primer binding site of the ligase located on its 3′-end. For this

reason, the RT enzyme can produce cDNA copies of both reacted and non-reacted ligases. If a ligase fails to catalyze the ligation reaction or is too slow at it, the RT will likely make a cDNA copy of the ligase without inclusion of the substrate containing the T7 promoter sequence.

Once the cDNA has been produced, transcription can occur. The T7 RNA polymerase recognizes the promoter sequence, which is located on the 5′-end of the substrate-ligase cDNA complex, and transcribes it. A selection process occurs in which only catalytically active ribozymes can pass to the next generation and undergo sequential rounds of amplification. This selection event has two important consequences: (1) as non-reacted ribozymes undergo a negative selection, the population size decreases proportional to the increase in the number of non-reacted ribozymes. Hence, the potential of becoming extinct if the number of non-reacted ribozymes exceeds the number of reacted ribozymes (*e.g.*, slow ribozymes that get reverse transcribed before they have catalyzed the ligation reaction of the exogenous substrate). (2) Contamination of the work environment with exoribonuclease I and oligoribonuclease can cause the degradation or partial degradation of the 5′-end of the ligase and the substrate respectively.

After the transcription step, newly transcribed ligases will undergo another round of selection and amplification. In this fashion, generations can be carried

on for as long as fresh supply of the ingredients is provided. Consequently, a series of stock solutions needs to be prepared, as explained in the next section.

Critical Parameters

The CE cycles need to be performed under strict control of parameters and variables, such that the evolutionary outcomes of the population dynamics adhere solely to the catalytic events of the molecular populations rather than to artifacts caused by unwanted variations in the parameters along the lineage evolution. Parameters such as pH, salt concentration, metal ions, stock concentrations, and temperature all affect the enzymes kinetics and therefore the amount of amplification achieved during the specified time that is being used. Variation in these parameters can cause the size of the population to vary and affect the evolutionary outcomes, giving confounding results. By the same token, the lack of precision in the length of burst used each cycle can cause population size to fluctuate and also cause altered results. It is important to try to avoid more than the slight variations intrinsic to open systems.

Anticipated Results

If clonal populations of a ligase were initially used the genetic diversity of the population will increase over time as a consequence of the lack of error correction (*e.g.*, 3′ to 5′ exonuclease activity) of the reverse transcriptase enzyme. To inspect if polymorphisms have arisen, an RFLP test can be done using restriction enzymes that are known to cut (or fail to cut) the ligase genome at a particular sequence. If genetic variants are detected, a more in-depth genotypic characterization of the populations can be achieved through an extensive cloning and nucleotide sequence analysis of specific burst of the evolved lineages.

Computer software can be used to aid in the analysis of the population composition and structure. The sequences can be aligned to detect the location of mutation in the primary structure of the ligases, and to determine whether those mutations are randomly distributed throughout the sequence or are more commonly located in a specific part of the sequence. Analysis of the type and location of the mutations is important to predict the effect of the accumulation of mutations on the fitness of the ligases. Network diagrams can be built to display the structure of the population and detect relationships between mutant sequences, which may not be that easily detected otherwise. The secondary structure of the ligases can be predicted using the appropriate software for RNA secondary structure prediction, taking into account that the

ligases used to date have a pseudoknot motif, and that not all programs are efficient at calculating such complex secondary structure arrangements.

Finally, the effect of the genotype on the phenotype can be inferred by performing kinetic assays to test the catalytic rate of enhancement of the mutant ligases. This measurement should be done under simulated CE conditions, and with a substrate radiolabeled with γ-ATP at its 5′ end. The reaction occurs very rapidly, so time points are mostly accurate if taken at a minimum of 15 sec by hand, or on the order of milliseconds if a quench flow system is used (Voytek and Joyce, 2007).

Time Considerations

Performing evolutionary studies with the CE system requires a dedicated amount of time. A wide variety of questions can be explored with this system and the length of time necessary to gather enough data depends on the question itself. Careful planning needs to be done to chose the initial size and composition of the population. In the case of the experiments reported on Soll *et al*., 2007, for example, an amplification/dilution set-up needed to be done first, to determine the different population sizes. Once they were established, the actual CE cycles could be performed for each population. Replicate lineages are necessary to account for slight random variations intrinsic of open systems. Each lineage is carried out for a established number of bursts

depending on the question (*e.g.*, 50). Once the replicate lineages have been finished for all the populations, data analysis requires additional processing such as cloning and sequencing. Finally, software processing and analysis of the genotypes sequenced also needs to be considered in the experimental timeline.

EXPERIMENTAL PLANNING

In order to initiate the CE experiments it is necessary to prepare all the stock solutions that are needed to be combined in the reaction vessel. The following is a description of the preparation processes prior to the actual CE experimentation.

Ligase ribozymes

The CE system can use one or two types of ligase ribozymes, depending on the purpose of the experiment. Stock DNA mini-preps can be kept in storage at -20°C and aliquots can be taken from them when a set of evolution experiments is to be initiated. The DNA aliquot can be amplified via PCR and transcribed following standard laboratory procedures. Once fresh RNA has been prepared, the CE experiments can be all performed with the same RNA preparation.

> *If clonal populations are to be used, a class I ligase ribozyme (Wright and Joyce in 1997; Soll et al., 2007) or a DSL ligase ribozyme (Voytek and Joyce, 2007) miniprep stocks can be kept in storage at -20°C, and used to prepare fresh RNA following standard in vitro transcription*

protocols. Alternatively, the evolution process can be initiated with a randomized population of ribozymes. In this case, error-prone PCR (Caldwell and Joyce, 1992; Vartanian et al., 1996) can be used at the desired mutagenic level to produce an initial library of RNA variants to initiate the evolution experiments. Also, ecological studies (e.g., of competition or of cooperation) can be pursued by starting the CE experiments with a mixed population of class I and DSL ligases (Voytek and Joyce, 2009).

Substrate

The stock substrate is acquired from the company of choice, and should be purified to ensure size uniformity. A 10% polyacrylamide / 8M urea gel electrophoresis can be used, followed by a standard gel extraction procedure of the RNA. A 100 μM stock is prepared and stored at -20°C, preferably in a dedicated freezer.

It is important to keep in mind that the substrate is a DNA/RNA chimera and, although the DNA constitutes most of the sequence, the handling of it should be done with the care that a labile RNA demands. Also, it is recommended to prepare the substrate at high concentration (100 μM) because it is required in excess in the CE reaction vessel. For example a concentration of approximately 6 μM of trans *substrate is typically utilized in the CE reaction (Soll et al., 2007).*

Primer for reverse transcription

The cDNA primer can be acquired from various companies at the desired concentration. A CE stock at 100 μM should be prepared and stored in the freezer at -20°C.

CE buffer

The CE system uses a buffer that can support the catalytic function of both the ligase ribozymes and the protein enzymes. A 3.3X buffer stock is prepared that will dilute to 25 mM total Mg^{2+} in the 1X reaction (16.2 mM free Mg^{2+}). The Mg^{2+} is required for the ribozymes' proper folding and consequent catalytic activity. The recipe is as follows:

82.5 µL 2M KCl

100 µL of 1M EPPS (pH 8.3)

83.2 µL of 1M $MgCl_2$

16.5 µL of 1M DTT

6.6 µL of 1M spermidine

711.2 µL of DEPC-treated and/or RNase-free water (*e.g.*, from Ambion, Inc.)

Σ = 1 mL of total buffer volume

The NTP mix

A mixture of the nucleotides (NTPs), deoxynucleotides (dNTPs), primer for reverse transcription, and substrate oligonucleotide is prepared as follows:

10.8 µL of 100 µM substrate oligomer

3.26 µL of 100 µM primer for reverse transcription

11.03 µL of 100 mM each rNTP

4.39 µL of 25 mM dNTP mix (25 mM each of dATP, dGTP,dUTP, and dCTP)

104.48 µL of DEPC-treated and/or RNase-free water

Σ = 167.05 µL of total NTP mix volume

The PE mix

The CE system uses two protein enzymes: a reverse transcriptase (RT) and an RNA polymerase which amplifies the ligase during the CE cycles. The protein enzymes (PE) are mixed in one reaction vessel right before the initiation of the continuous evolution process and are kept in a small bench-top cooler at -20°C near the CE hood. A ratio of 4 to 1 between the RT and the RNAP enzymes is recommended, because the reverse transcriptase is required at high concentrations in the mixture.

> *In the Lehman laboratory the enzyme MMLV reverse transcriptase from USB Corp. has been used (Soll et al, 2007). This enzyme comes in a concentration of 200U/µL, and has RNase H activity. In the Joyce laboratory the SuperScript and Superscript II reverse transcriptase from Invitrogen Corp. has been used with good results (Voytek and Joyce, 2009). T7 RNA polymerase can be purchased from Ambion, Inc (200U/µL) or prepared in situ in the laboratory if preferred.*

CONTINUOUS EVOLUTION *IN VITRO* (CE) EXPERIMENTS

Once all stock solutions have been prepared, the CE experiments can be started. The actual evolution experiments can be performed in a warm room at 37°C, in a hood dedicated exclusively for the CE experiments (Fig. 3), or in a microfluidics system. If a warm room is used, strict regulation of room usage must be kept to avoid RNase contamination issues. Also, careful regulation of the temperature must be maintained to avoid introducing variables that may

have unwanted effects on the evolutionary dynamics of the populations. If a hood is chosen for the CE experiments, a heat block must be placed in it, and the temperature should be kept at 37°C during the selection/amplification cycles. The advantages of using a hood are that the PE mix can be kept cold more easily and that contamination can be significantly reduced.

Strict adherence to the following procedure is recommended to prevent contamination of the work environment by RNases. A small contamination of RNases that may not be easily detected can degrade the ligase ribozymes and hence cause an unintended reduction of the population size. A misinterpretation of this process is that the population size is decreasing as a consequence of the reduction in the fitness of the ligases. By the same token, strict timing of the serial transfers of the RNAs is recommended to avoid over-amplifying the population and affecting the evolutionary dynamics happening in the test tube (see Selection and Amplification section).

Materials

Stock ribozymes [50 μM]

Substrate oligonucleotide [100 μM]

Reverse transcription primer [100 μM]

Each of the four rNTPs [100 mM]

Freshly prepared dNTP mix [25 mM]

Reverse Transcriptase 200U/μL

T7 RNA polymerase 200U/μL

Potassium chloride 2M

4-(2-hydroxethyl)-1-piperazinepropanesulfonic acid (EPPS) 1M at pH 8.3

Magnesium chloride 1M

Dithiothreitol (DTT) 1M

Spermidine 1M

DEPC-treated and/or RNase-free water

CE reaction cycles

The following step-by-step experimental procedure description can be used for CE experiment carried out on a warm room or in a dedicated hood. Only two serial transfers of the protocol followed for the CE are described here. Each serial transfer involves a proliferation of RNA molecules that are termed a "burst" (Schmitt and Lehman, 1999). The proliferated RNA molecules of each serial transfer are used to seed the next one in a continuous fashion, as follows:

1. Place the RNA dilution, buffer, and NTP mix tubes in an ice bucket next to the CE hood to thaw.

2. Label 600 μL tubes with the experimental code and the burst number. These are the CE reaction vessels.

3. Set up the timers for the amount of time required to complete the cycles (see Dilution and Amplification section).

4. Fill various 1.5mL tubes with the necessary amount of DEPC-treated and/or RNase-free water required for the dilution process (see Dilution and Amplification section). Close the lids and label them with the appropriate experiment code and burst number.

5. Move the thawed stock tubes from the ice bucket into the hood, place them dry in the “reaction tubes” rack. Keep their lids closed.

6. Get the bench-top cooler with the PE mix out of the freezer and place it outside, but next to, the CE hood (Fig. 3).

6. Add 7.5 µL of the CE buffer to the reaction tube corresponding to the first burst.

7. Add 7.8 µL of the NTP mix into the same reaction vessel for the first burst.

8. Add 8.2 µL of the RNA ligase and mix by quickly pipetting the solution up and down.

9. Place the reaction tube with the lid open in the heat block at 37°C.

10. Take a 1.5 µL aliquot of the PE mixture from the bench-top cooler and add to the reaction vessel. Close the lid, and mix by gentle finger flicking of the bottom of the tube a few times and spin a few seconds in a bench top centrifuge.

11. Rapidly place the tube back into the heat block and start the timer.

12. Return the bench-top cooler to the freezer.

13. Return all the stock tubes into the ice bucket and keep them with the lids closed.

14. Remove the dilution tube for the first burst from the “dilution tubes” rack and place it in the “reaction tubes” rack. Keep its lid closed at all times.

15. Be aware of the timer, and five minutes before the reaction has to be stopped, take the stock solution tubes from the ice bucket, quickly dry them and place them back into the CE hood in the “reactions tube” rack.

16. Three minutes before the reaction has to be stopped, get the second burst reaction tube ready by adding 7.5 µL of the buffer and 7.8 µL of the NTP mix. Close the lids of the tubes at all times, except when pipetting solutions in or out of them.

17. One minute before the reaction needs to be stopped, check that the pipet is set to 3 µL and have it ready to take an aliquot from the reaction vessel.

18. About 20 seconds to time zero, take the reaction vessel from the heat block uncap it and take a 3 µL aliquot.

19. Place the tube in the “reaction tubes” rack and take the dilution tube and uncap it.

20. Hold the pipet tip close to the dilution tube mouth, and when the timer goes off, add the 3 µL of reaction into the dilution tube.

The time the reaction is maintained at room temperature prior to dilution is relatively insignificant and can be decreased with experience.

21. Close the lid of the dilution tube and immediately and mix by inversion two to three times. Spin the tube few seconds and place it back into the “reaction tubes” rack.

22. Take a 8.2 µL aliquot of the dilution tube from this first burst to seed a second burst; add the 8.2 µL RNA to the reaction vessel labeled for the second burst, and which already contains the CE buffer and the NTP mix.

23. Place the tube in the heat block, and repeat steps (9 – 22) to seed further bursts.

> *Iterate this procedure as many times as possible in one working day, and keep the dilution tubes on ice until the end of the experimental session. At the end of the session, place the reminder of the CE buffer and NTP mix stock solutions in the -20°C freezer located next to the CE hood, and which is designated only for CE experiments. Next, take all the dilution tubes from the ice bucket where they have been kept, dry them all, re-mix by inversion a couple of times, and spin in the centrifuge*

The procedure described above is for one lineage, although more than one lineage can be carried on in parallel, even using manual pipetting methods. However, it is recommended to do a maximum of three lineages at the time, including a water “lineage” (a negative control with no RNA added). One timer for each lineage should be used. Also, because it requires more time to propagate multiple lineages, to get the tubes warmed up, to add the buffer and the NTP mix to all tubes, to have them capped, and to have the pipet ready to take the 3 µL from the reaction vessel, it is recommended that one be ready at

the hood at an earlier time than when doing one lineage (*e.g.*, five minutes before the timer goes off).

Once the CE session is finished, do a PCR amplification of each burst done in the session to test the status of the population. The PCR amplification products can be loaded on an agarose gel, each burst in one well, and visualized to evidence the presence of the bands for each burst. If bands are observed for all the bursts (Fig. 4), another set of CE experiments can be initiated, seeding the next burst from the last saved dilution tube. Simply take the saved dilution tubes from the freezer and thaw them as indicated above. In this fashion, lineages can be evolved for as long as the populations stay extant. If absence of the proper size band is observed the population can be declared extinct and a search for the causes of this extinction can be initiated.

Dilution and amplification scheme

The effective population size (N_e) first described by Sewall Wright (1931) has strong impact on the evolutionary dynamics of the population: it determines the genetic variability of the population and has an implication for the survival of the population. The amount of fluctuation in gene frequencies is expected to be become larger as the N_e is reduced, because of the effect of random genetic drift (Kimura, 1983). Populations with small N_e will more readily built

up a genetic load by the accumulation of deleterious mutations than populations with larger effective sizes. The variable N_e in this context can be defined as the harmonic mean of the population size in each burst. For the aforementioned reasons, it is desirable that the effective size of the population does not change as a consequence of experimental effects, because it will introduce an external influence in the evolutionary trajectory of the populations under study.

To ensure that the size of the populations is not being effected by the amplification time used in the CE experiment, preliminary experiments are necessary to calculate the maximum amount of amplification that can be achieved before enzyme degradation and inhibition by side product build-up occurs. At this point, the population will cease to grow, as evidenced by the absence of the correctly sized band of the PCR product upon agarose gel analysis. The maximum amplification determines when fresh “nutrients” have to be provided to the reaction. The length of the burst time calculated is relative to the N_e used, and should be kept constant to avoid fluctuations of the N_e. Preliminary CE experiments are required to determine the right amount of time for the amplification cycle and the correspondent degree of dilution required between bursts.

The CE cycles

1. Follows the standard procedure above described, with the difference that a 1.5 µL of radiolabeled α-^{32}P-ATP (10 µCi/µL) is added to the NTP mix vessel.

2. Remove 3 µL aliquot from the reaction vessel about one minute before the reaction has to be stopped.

3. Place these 3 µL into a 600 µL tube that contains 3 µL of 2X XC^{-} acrylamide gel-loading dye (bromophenol blue with no xylene cylenol added).

4. Do approximately 10 bursts.

5. Keep all tubes on ice until the experimental section has been completed.

6. When the bursts are done, prepare a PCR sample from the dilution tubes. Run the amplified products on an agarose gel and obtain an image of the band distribution.

7. Prepare a 5% polyacrylamide / 8M urea gel to run the radiolabeled samples. Load the 6 µL of the RNA/dye of each burst into a single well of the gel. Let run for about 1000 V*hr and expose to a phosphorimager screen.

8. Compare the images of the PAGE and the PCR agarose gel (Fig. 5) to determine if over- or under-amplification is occurring.

> *If the population is not being over-amplified, the amount of the radiolabeled sample is close to the detection limit of the phosphorimager scanner, and hence an image of the bands for each burst can be barely detected, if at all, compared to a population size positive control used to verify the effectiveness of the radiolabeling process. The PCR gel, in contrast, should show a band for each burst as normally observed when the population is still "alive" (Fig. 4). If bands are not observed in the agarose gel or if they disappear, an excessive dilution of the population may be occurring. In the opposite*

case, if bands are observed in both PAGE and PCR gels, an under-dilution (over-amplification) must be occurring (Fig. 6).

9. According to the results of step 8, estimate a new length for the amplification time.

10. Repeat steps 1 – 8 and check the results.

If the proper burst time has not being found, iterate this procedure until the proper amplification/dilution scheme has been found. The burst time can be varied typically between 15-30 minutes.

Once the time necessary for the amplification process and the corresponding dilution scheme have been found for the population size of interest (Fig. 7), the evolution experiments can be started (Basic Protocol I). An ample variety of questions can be tested with the CE experiments. For example, in the Lehman laboratory, the effect that mutation accumulation (MA) has on the survival of populations has been tested. Clonal populations of ligases of 100, 300, 600 and 3000 molecules were used, and a direct relationship was found between MA and time to extinction (Soll *et al*., 2007). More recently, the effects of an increased mutational rate in the survival of bottleneck populations of 100 molecules has been evaluated, and unexpected population dynamics were found to evolve in the ligase populations (Diaz Arenas and Lehman, in review). In the Joyce laboratory, experiments of substrate competition has been performed, revealing ecological dynamics of nutrient selection and competition (Voytek and Joyce, 2007; 2009).

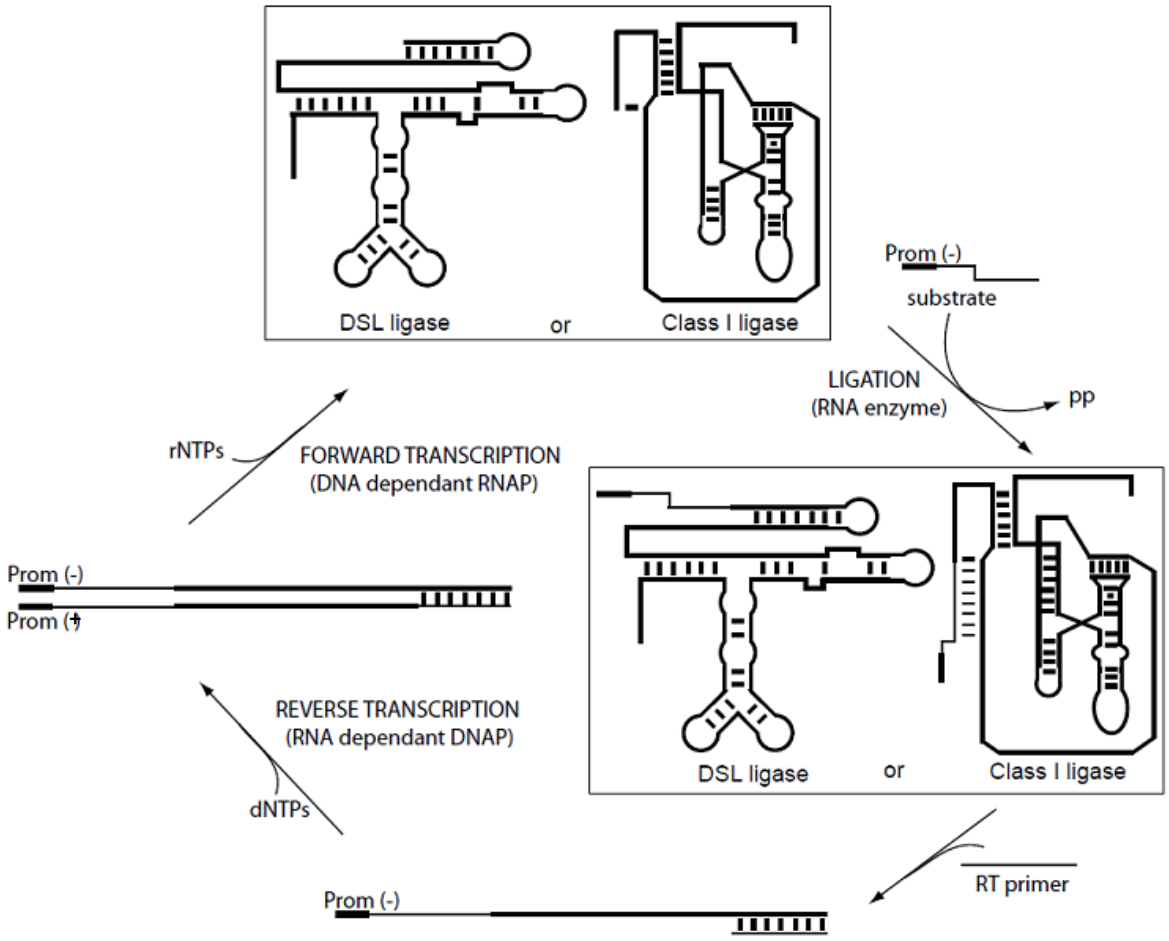
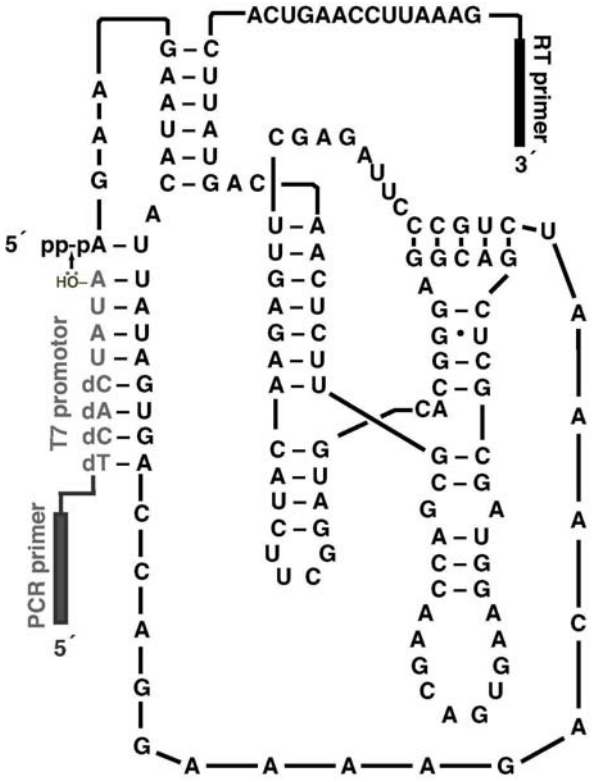


Figure 1. The CE system. The continuous *in vitro* evolution (CE) scheme showing the reaction cycles where the ribozymes are amplified. All reaction steps take place in one homogeneous environment. The cycles start when the ribozymes (secondary structures shown at the top) catalyze the ligation of the *trans* oligomer substrate to themselves. Next a reverse transcription occurs. The cDNA copies are then transcribed if they posses the T7 RNA polymerase promoter sequence, which is contained in the substrate, to initiate another cycle. At this point, cDNA products of non reacted ligases are not transcribed, RNA strands are represented by thick solid lines, while the cDNA strands are represented by thin solid lines.



B16-19 Class I ligase ribozyme

Figure 2. Secondary structure of B16-19 ligase ribozyme with the exogenous substrate. The ribozyme effects the ligation of this substrate oligomer (in gray) to itself by catalyzing the attack of the 3´-hydroxyl group of the substrate onto the 5´-α-phosphate of the ribozyme (arrow). The substrate contains the T7 RNA transcriptase promoter sequence that is required for the CE cycles to occur. The primer binding sites for reverse transcription (3´ end) and PCR amplification at the 5´ end of the promoter are denoted by solid rectangles.

Figure 3. CE hood. Photograph of the CE hood that shows it organization inside and its surroundings. Next to the hood, at the right side, notice the ice bucket, the small bench cooler where the PE mix is kept, and the small freezer dedicated to the CE experiments. Inside the hood there are two distinctive areas with a spatial separation between them: the reaction area and the dilution area. The reaction area is localized towards the right side of the hood. It contains the timers, the 600 µL rack that has attached to it five holders for 1.5 µL tubes (for CE buffer and three dilution tubes), the heat block, pipet tips and a dedicated pipet for the reactions. Notice that this area is close to the side of the hood where the bench-top cooler is kept, to minimize motion. Stock tubes with NTP mix and CE buffer are placed on this rack towards the right side, and kept closed at all times except when being used. The 600 µL reaction tubes are kept in the rack by the far left side, and moved to the front when in use (*e.g.*, five minutes before stopping the reaction). The dilution area is towards the left side of the hood and it contains a water bottle, a couple of racks for 1.5 µL tubes, a pipet tip box and a pipet dedicated to fill the dilution tubes, which are kept in these rack with their lids close at all times. The dilution tubes necessary to stop a reaction are moved from these racks to the holders attached to the reaction rack.

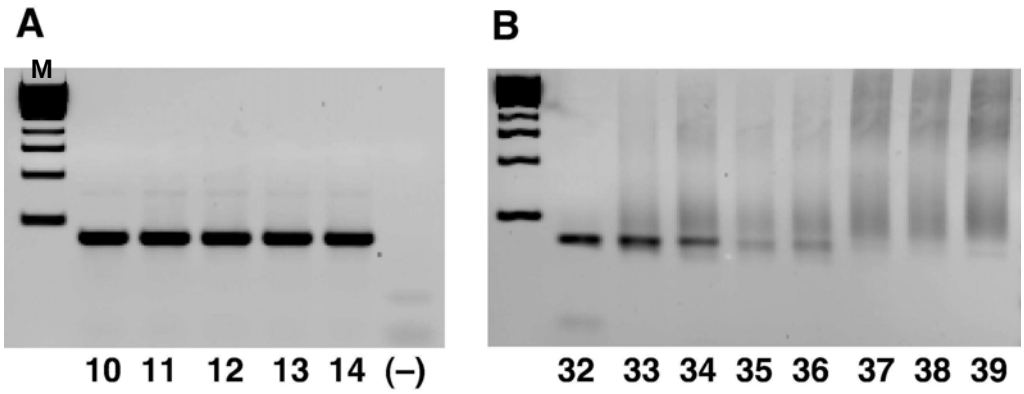


Figure 4. Agarose gel analysis of population persistence or extinction. Agarose gel images of two populations evolved with the CE system. Numbers beneath the gels track the specific bursts within a CE lineage. The symbol (–) denotes a negative control in the PCR. Panel A depicts a healthy population, with the presence of a robust PCR band being sustained through time. Panel B depict a population that is going extinct, with the disappearance of the proper sized band over time, a consequence of a lack of sufficient RNA amplification of the population during CE.

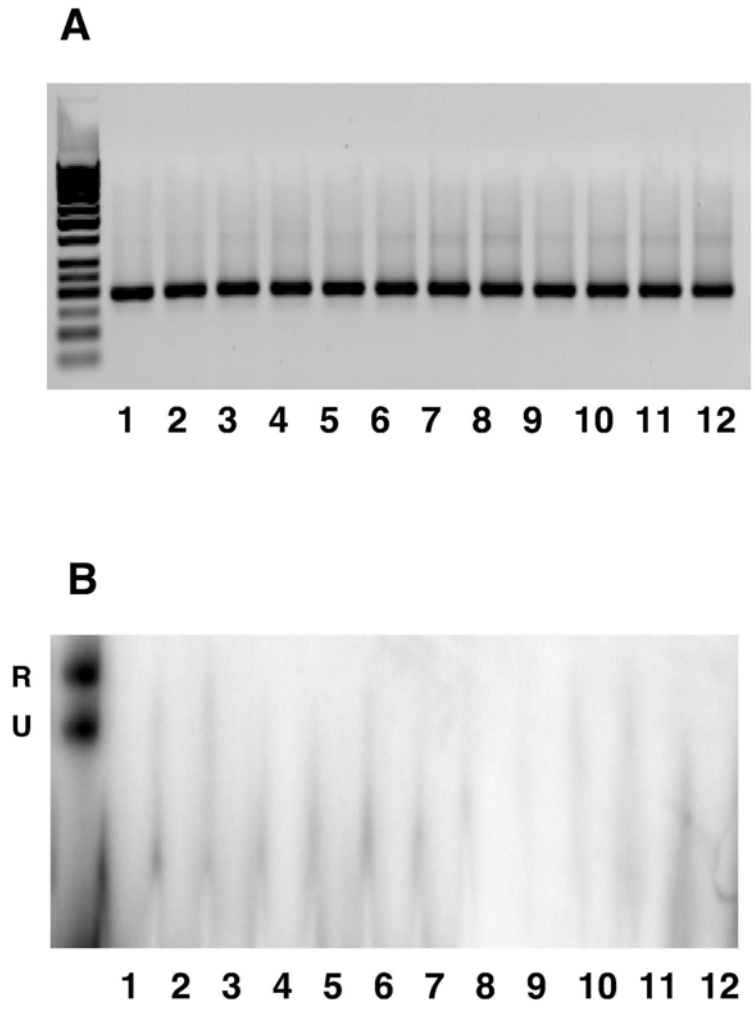


Figure 5. Adjusting the amplification factor to maintain N_e. The number of minutes in each CE burst can be adjusted to achieve an RNA amplification that matches the dilution factor between bursts. These gels depict a population in which this match is correct. Panel A shows the PCR products resulting from the cDNA amplification of a healthy lineage as in Fig. 4. Panel B shows a PAGE image of the same lineage in which α-^{32}P-ATP has been included in the CE mixture. The left-most lane is a positive control where reacted (R) and unreacted (U) class I ligase ribozymes are located. Note that no RNA can be seen in any of the bursts, signifying that the RNA population is not being overamplified with time.

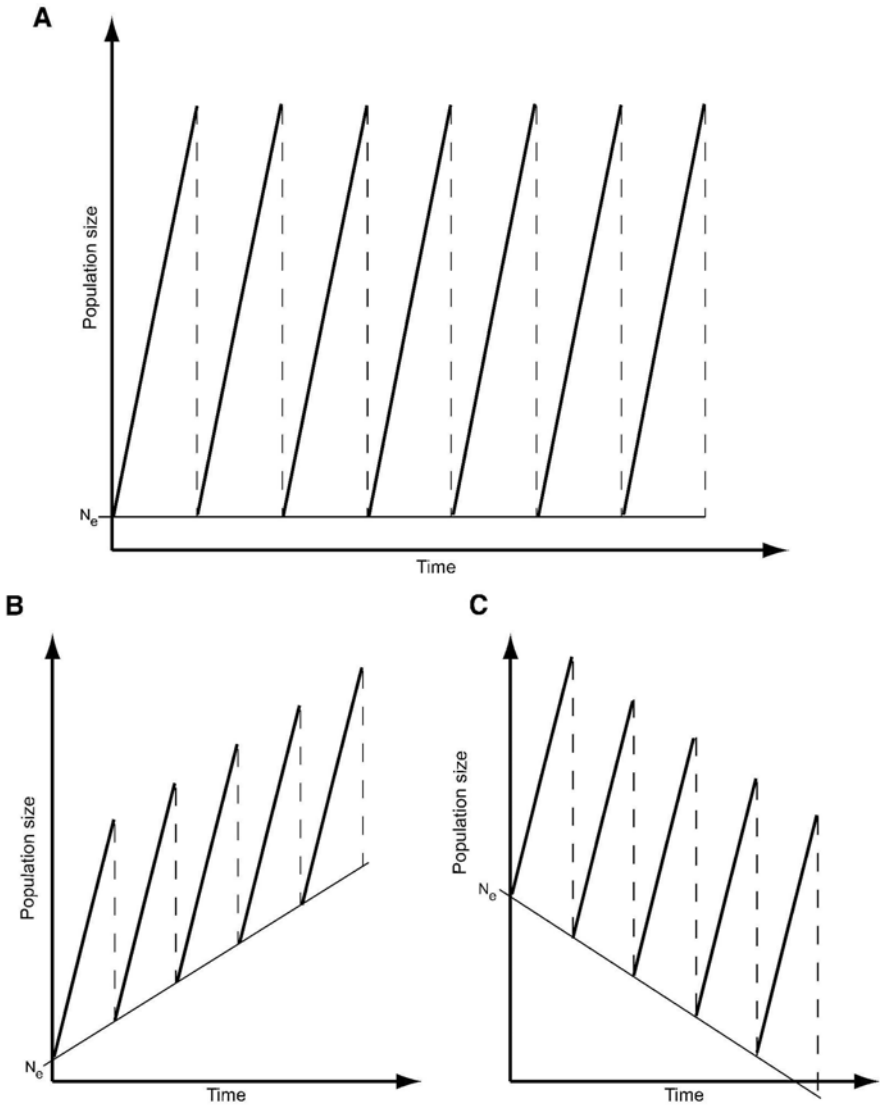


Figure 6. Amplification dilution scheme. During the CE the ribozymes are subject through cycles of amplification (bold lines). At the end of each cycle (burst) the amplified population size is subject to a dilution (dashed lines) that serves two purposes: (1) stop the reaction, and (2) keep the effective population size (N_e) nearly constant on average. The *x*-axis indicates time and the *y*-axis indicates population size. Panel A shows a case in which the dilution matches the amplification and the N_e stays nearly constant. Panel B shows a case of underdilution and the N_e increases with time. Panel C shows a case of overdilution and N_e decreases with time; notice that by the end of the fifth burst the population size has dropped below zero (extinction). The cases shown in panels B and C are to be avoided because they mask increases or decreases of the N_e due to evolutionary dynamics.

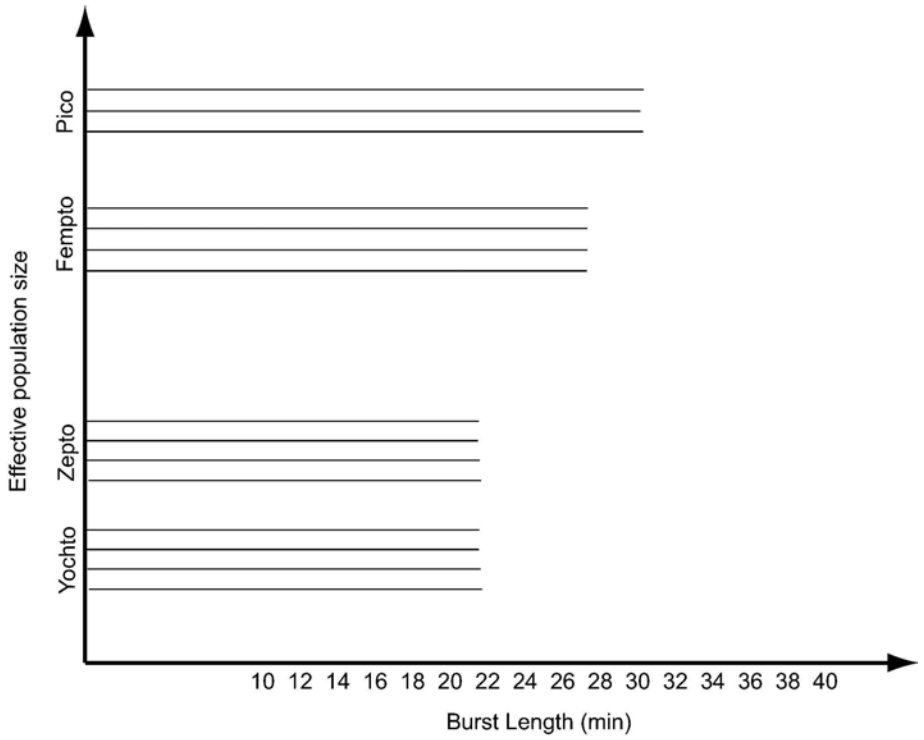


Figure 7. Burst length vs. N_e scheme. The length of the burst is given by the amount of amplification that can be achieved before nutrients become exhausted and side product builds up. The amount of amplification achieved given the nutrients provided is dependent on the population size. Therefore, the burst length is directly related to population size, as shown. The *x*-axis indicates the burst length time (in minutes) and the *y*-axis indicated the population size (in moles). Approximate burst lengths and population sizes are given as an example.

References

Ancel, L.W., and Fontana, W. 2000. Plascticity, evolvability, and modularity in RNA. *J. Exp. Zool.* 288:242–283.

Bartel, D.P., and Szostak, J.W. 1993. Isolation of new ribozymes from a large pool of random sequences. *Science* 261:1411–1418.

Beaudry, A.A., and Joyce, G.F. 1992. Directed evolution of an RNA enzyme. *Science* 257:635–641.

Breaker, R.R., and Joyce, G.F. 1994. Emergence of a replicating species from an *in vitro* RNA evolution reaction. *Proc. Natl. Acad. Sci. USA.* 91:6093–6097.

Breaker, R. 2000. Selection for catalytic function with nucleic acids. *Curr. Protoc. Nucleic Acid Chem*. 9.4.1–9.4.17

Caldwell, R.C., and Joyce, G.F. 1992. Randomization of genes by PCR mutagenesis. *PCR Methods Appl*. 2:28–33.

Cech, T.R. 1987. The chemistry of self-splicing RNA and RNA enzymes. *Science* 236:1532–1539.

Davidson, E.A., Dlugosz, P.J., Levy, M., and Ellington, A.D. 2009. Directed evolution of proteins *in vitro* using compartmentalization in emulsions. *Curr. Protoc. Mol. Biol*. Chapter 24:Unit 24.6.

Díaz Arenas, C., and Lehman, N. 2009. Darwin's concept in a test tube: Parallels between organismal and *in vitro* evolution. *Int. J. Biochem. Cell Biol.* 41:266–273.

Díaz Arenas, C., and Lehman, N. 2010. Quasispecies behavior observed in RNA populations evolving in a test tube. In review in *BMC Evol Biol.*

Ellinger, T., Ehricht, R., and McCaskill, J.S. 1998. *In vitro* evolution of molecular cooperation in CATCH, a cooperatively coupled amplification system. *Chem. Biol.* 5:729–741.

Ellington, A.D., and Szostak, J.W. 1990. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* 346: 818–822.

Ellington, A.D., Chen, X., Robertson, M., and Syrett, A. 2009. Evolutionary origins and directed evolution of RNA. *Int. J. Biochem. Cell Biol*. 41:254–265.

Guatelli, J.C., Whitfield, K.M., Kwoh, D.Y., Barringer, K.J., Richman, D.D., and Gingeras, T.R. 1990. Isothermal, *in vitro* amplification of nucleic acids by a multienzyme reaction modeled after retroviral replication. *Proc. Natl. Acad. Sci. U.S.A.* 87:1874–1878.

Ikawa, Y., Tsuda, K., Matsumura, S., and Inoue, T. 2004. *De novo* synthesis and development of an RNA enzyme. *Proc. Natl. Acad. Sci. U.S.A*. 101:13750–13755.

Jhaveri, S., and Ellington, A. 2002. *In vitro* selection of RNA aptamers to a small molecule target. *Current Protoc. Nucleic Acid Chemistry* 9.5.1–9.5.14.

Johns, G.C., and Joyce, G.F. 2005. The promise and peril of continuous *in vitro* evolution. *J. Mol. Evol.* 61:253–263.

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

Joyce, G.F. 1989. Amplification, mutation and selection of catalytic RNA. *Gene* 82:83–87.

Joyce, G.F. 2004. Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem.* 73:791–836.

Joyce, G.F. 2007. Forty years of *in vitro* evolution. *Angew. Chem. Int. Ed.* 46:2–19.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Great Britain: Cambridge University Press.

Kun, Á., Santos, M., and Szathmáry, E. 2005. Real ribozymes suggest a relaxed error threshold. *Nature Genet.* 37:1008–1011.

Lehman, N., Delle-Donne, M., West, M., and Dewey, G. 2000. The genotypic landscape during *in vitro* evolution of a catalytic RNA: implication for genotypic buffering. *J. Mol. Evol.* 50:481–490.

Lehman, N. 2004. Assessing the likelihood of recurrence during RNA evolution *in vitro*. *Artif. Life* 10:1–22.

Lehman, N., and Joyce, G.F., 1993. Evolution *in vitro* of an RNA enzyme with altered metal dependence. *Nature* 361:182–185.

Levisohn, R., and Spiegelman, S. 1969. Further extracellular Darwinian experiments with replicating RNA molecules; diverse variants isolated under different selective conditions. *Proc. Natl. Acad. Sci. U.S.A.* 63:805–811.

McGinness, K.E., Wright, M.C., Joyce, G.F. 2002. Continuous *in vitro* evolution of a ribozyme that catalyzes three successive nucleotidyl addition reactions. *Chem. Biol.* 9:585–596.

Mills, D.R., Peterson, R.L., and Spiegelman, S. 1967. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad*. *Sci. U.S.A.* 58:217–224.

Orgel, L.E. 1979. Selection *in vitro*. *Proc. Roy. Soc. Lond. B Biol. Sci*. 205:435–442.

Paegel, B.M., and Joyce, G.F. 2008. Darwinian evolution on a chip. *PLoS Biol*. 6:900–906.

Paegel, B.M., Grover, W.H., Skelley, A.M., Mathies, R.A., and Joyce, G.F. 2006. Microfluidic serial dilution circuit. *Anal. Chem.* 78:7522–7527.

Piasecki, S. K., Hall, B., and Ellington, A.D. 2009. Nucleic acid pool preparation and characterization. *Methods Mol. Biol.* 535:3–18.

Saffhill, R., Schneider-Bernloehr, H., and Orgel L.E. 1970. *In vitro* selection of bacteriophage Qβ ribonucleic acid variants resistant to ethidium bromide. *J. Mol. Biol.* 51:531–539.

Schmitt, T., and Lehman, N. 1999. Non-unity heritability demonstrated by continuous evolution *in vitro*. *Chem. Biol.* 6:857–869.

Seelig, B., and Jäschke, A. 1999. A small catalytic RNA with Diels-Alderase activity. *Chem. Biol*. 6:167–176.

Soll, S., Díaz Arenas C., and Lehman, N. 2007. Accumulation of deleterious mutations in small abiotic populations of RNA. *Genetics* 175:267–275.

Tarasow, T.M., Tarasow, S.L., and Eaton, B.E. 1997. RNA-catalysed carbon–carbon bond formation. *Nature* 389:54–57.

Vartanian, J. -P., Henry, M., and Wain-Hobson, S. 1996. Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversion. *Nucleic Acids Res.* 14:2627–2631.

Voytek, S.B., and Joyce, G.F. 2007. Emergence of a fast reacting ribozyme that is capable of undergoing continuous evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104:15288–15293.

Voytek, S.B., and Joyce, G.F. 2009. Niche partitioning in the coevolution of 2 distictive RNA enzymes. *Proc. Natl. Acad. Aci U.S.A.* 106:7780–7785.

Whitesides, G.M. 2006. The origins and the future of microfluidics. *Nature* 442: 368–373.

Wilson, C., and Szostak, J.W. 1999. *In vitro* selection of functional nucleic acids. *Annu. Rev. Biochem.* 68:611–647.

Wright, M.C., and Joyce, G.F. 1997. Continuous *in vitro* evolution of catalytic function. *Science* 276: 614–617.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16:97-159.