Nucleic acid phase separation and replication drive selection in prebiotic pools

Adriana Calaça Serrão

München 2024

Nucleic acid phase separation and replication drive selection in prebiotic pools

Adriana Calaça Serrão

Dissertation
an der Fakultät für Physik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Adriana Calaça Serrão
aus Almada, Portugal

München, den 25.03.2024

Erstgutachter: Prof. Dr. Dieter Braun
Zweitgutachter: Prof. Dr. Hannes Mutschler
Tag der mündlichen Prüfung: 08.05.2024

Zusammenfassung

Nukleinsäuren gelten aufgrund ihrer Fähigkeit, Informationen zu speichern und gleichzeitig Reaktionen zu katalysieren, als evolutionär ursprüngliche Moleküle. Die Entstehung komplexer funktioneller Nukleinsäuren aus einem Pool von Bausteinen oder kurzen Fragmenten impliziert die Selektion bestimmter Sequenzen. Der Sequenzraum von Nukleinsäurepools mit Dutzenden oder Hunderten von Basen ist zu groß, um vollständig erforscht zu werden. Aus diesem Grund geht man davon aus, dass eine Selektion bereits dann stattfindet, wenn die Pools komplexer werden, und nicht erst, wenn sie die zur Erzeugung eines Ribozyms ausreichende Länge erreicht haben. In allen Phasen der Nukleinsäureverlängerung und -replikation sind die Stränge verschiedenen Selektionsmechanismen ausgesetzt. Diese entstehen entweder aus dem Replikations- oder dem Verlängerungsmechanismus oder aus äußeren Einflüssen der Umgebung.

In Kapitel 1 wird gezeigt, dass die Fähigkeit von Nukleinsäuren, Phasentrennung zu vollführen, möglicherweise als extrinsischer Selektionsdruck für Sequenzen mit spezifischen Sekundärstrukturen wirkt, indem sie Schutz vor Verdünnung und Hydrolyse bietet. Unsere Studie ergab, dass Sequenzen, die zur Phasentrennung fähig sind und sich dadurch auf dem Boden einer Steinpore absetzen können, bei wiederholter Entfernung des Überstandes gegenüber anderen Sequenzen angereichert werden. Diese Bedingungen ahmen geologische Gegebenheiten mit sich verändernden Flüssigkeitsströmen nach, wie z.B. Gewässer, die einem Tag-Nacht-Zyklus ausgesetzt sind. Genauer gesagt sind die angereicherten Sequenzen jene, die dazu neigen, netzwerkartige Sekundärstrukturen zu bilden. Systeme mit geringer Vielfalt in der Komposition – etwa binäre Systeme aus ausschließlich **AT** oder **GC** – wiesen keine Phasentrennung auf. Die Einführung einer einzigen Mutation des jeweils anderen Alphabets stellte diese Fähigkeit jedoch wieder her, was auf eine hohe Sequenzspezifität der Phasentrennung als treibende Kraft für die Selektion hinweist.

In Kapitel 2 wurde der Einfluss von Replikation auf Pools kurzer Oligomere mit verzerrter Anfangsverteilung als ein intrinsischer Faktor untersucht. Zufällige Pools kurzer DNA-Stränge, die im Durchschnitt in Richtung eines bestimmten Nukleotids verzerrt sind, wurden mit der Polymerase *Bst* templatbezogen polymerisiert. Während die Gesamtvielfalt der Poolzusammensetzung auf Nukleotidebene aufgrund der templatbezogenen Natur des Mechanismus zunahm, blieben Reste der anfänglichen Zusammensetzungsverzerrung an bestimmten Positionen der replizierten Stränge bestehen. Ursprünglich unstrukturierte Pools entwickelten Muster wie z.B. Periodizität, die die Replizierbarkeit durch vermehrte Bindungsstellen innerhalb eines und zwischen verschiedenen Strängen verbesserten. Auch hier begünstigte der durch den Replikationsmechanismus ausgeübte Selektionsdrucks eine bestimmte Gesamtzusammensetzung des Pools sowie bestimmte Sekundärstrukturen und Sequenzen.

In Kapitel 3 wird ein nicht-enzymatisches RNA-Replikationssystem beschrieben, das 2′, 3′-zyklische Phosphate verwendet. Diese Aktivierungsgruppe, die während der RNA-Hydrolyse, der Nukleobasenpolymerisation und der präbiotischen Phosphorylierung gebildet wird, ist präbiotisch leicht verfügbar. Zum ersten Mal wurde eine templatbezogene RNA-Replikation

durch Ligation ohne Zugabe von organischen Katalysatoren erreicht. Dieses System wies eine hohe Sequenzspezifität auf, mit einer Ligationsgenauigkeit von über 82%. Lange Sequenzen bis zu 100-mer wurden durch konsekutive Ligationen synthetisiert und ebneten den Weg zur Erzeugung von Pools, die molekulare Evolution ermöglichen.

Abstract

Nucleic acids are considered to be early molecules owing to their dual ability to store information and catalyze reactions. The evolution of complex functional nucleic acids from a diverse pool of building blocks or short fragments necessitates the selective preference of certain sequences over others. The sequence space of nucleic acid pools with dozens or hundreds of bases is too vast to be fully explored. For this reason, it is believed that selection occurs as the pools become more complex, not only after attaining a sufficient length to support a ribozyme. Throughout the stages of nucleic acid elongation and copying, strands are subjected to various selection pressures, intrinsic to the replication or elongation mechanisms, or extrinsic, where certain sequences are better suited to survive in specific environments.

In Chapter 1, selection for nucleic acids capable of phase separation was studied as the result of extrinsic selective pressures such as flux changes and dilution. Sequences capable of phase separation were enriched over cycles of supernatant removal by sedimenting at the bottom of the pore while other sequences were washed-out. These enriched sequences were the ones prone to forming network-like secondary structures. The wash-out and refeeding conditions mimic geological settings with varying fluxes such as bodies of water subjected to day-night cycles or tidal waves. Sequence pools with low compositional diversity (i.e. those that contained a binary alphabet - **A**/**T** or **G**/**C**) did not exhibit phase separation. However, the introduction of a single mutation of the opposite alphabet restored this capacity, indicating a high sequence specificity of phase separation as a driving force for selection.

In Chapter 2, the influence of replication on pools of short, biased oligomers was examined as an intrinsic factor. Random pools of short DNA strands, on average biased toward a specific nucleotide, underwent templated polymerization with *Bst* [1]. While the overall pool compositional diversity increased at the nucleotide level due to the templated nature of the mechanism, remnants of the initial bias persisted at specific positions of the replicated strands. Initially unstructured pools developed patterns such as periodicity, enhancing replicability through increased intra- and inter-strand binding sites. In this case the overall pool composition also shifted to specific secondary structures, and therefore sequence motifs, due to the selective pressure exerted by the replication mechanism.

In Chapter 3, a non-enzymatic RNA replication system employing 2′, 3′-cyclic phosphates is described. This activation group, formed during RNA hydrolysis, nucleotide polymerization, and prebiotic phosphorylation, is readily available prebiotically. For the first time, templated RNA replication through ligation was achieved without the addition of organic catalysts. This system exhibited high sequence specificity, with a ligation fidelity exceeding 82%. Long sequences up to 100-mer were synthesized through consecutive ligation, paving the way for the creation of pools capable of hosting molecular evolution. Describing such a system represents the initial step in investigating the intrinsic influence of replication on a prebiotically plausible pool.

[1] short for *Bacillus stereathermophilus* polymerase I

List of Acronyms and Abbreviations

DNA Deoxyribonucleic acid

RNA Ribonucleic acid

TNA Threose nucleic acid

XNA Xeno nucleic acid

ssDNA Single-stranded DNA

dsDNA Double-stranded DNA

dNTP Deoxynucleoside triphosphate

dATP Deoxyadenosine triphosphate

dTTP Deoxythymidine triphosphate

dGTP Deoxyguanosine triphosphate

dCTP Deoxycytidine triphosphate

$T_\mathbf{m}$ Melting temperature

HPLC High-performance liquid chromatography

UV Ultraviolet radiation

SP Sequence Pair

LLPS Liquid-liquid phase separation

PAGE Polyacrylamide Gel Electrophoresis

NGS Next-Generation Sequencing

NAR Nucleic Acids Research

PNAS Proceedings of the National Academy of Sciences

JACS Journal of the American Chemical Society

Bst *Bacillus stereathermophilus* polymerase I

PCR Polymerase Chain Reaction

RegEx *Regular Expression*

HFIP 1,1,1,3,3,3-Hexafluor-2-propanol

EDTA Ethylenediaminetetraacetic acid

TBE Tris-Borate-EDTA

TAE Tris-Acetate-EDTA

TEA Triethylamine

CHES N-cyclohexyl-2-aminoethanesulfonic acid

MES (2-(N-morpholino)ethanesulfonic acid)

EDC 1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide

APS Ammonium persulphate

TEMED Tetramethylethylenediamine

UDP Uridine diphosphate

ESI TOF Electrospray ionisation time-of-flight

>P 2′,3′-cyclic phosphate

DAP Diamidophosphate

TMP Trimetaphosphate

PCR Polymerase chain reaction

Contents

Introduction

Chemical origins of life

The emergence of life needed a source of local non-equilibrium, facilitating self-organization, up-concentration, and a continuous flux of nutrients and energy to lower entropy [49, 119]. These non-equilibria, such as concentration and temperature gradients, salt and pH cycling, could have been initiated by external forces like geothermal heating, solar irradiation, and day-night cycles, propelling the system towards self-organization [7, 57]. Many geological scenarios can provide with these non-equilibria forces [56], such hydrothermal vents [23, 113], heated rock pores [56, 87], water ponds [26, 44], and more.

Complex environments as the ones mentioned above likely resulted in the formation of rich and diverse pools of organic molecules. While contemporary life relies on essential building blocks such as amino acids, nucleotides, and fatty acids, early life forms were likely simpler, potentially comprising only a subset of these components alongside other molecules no longer prominent in modern organisms. The molecular evolution of organic precursors of biomolecules over time complicates our understanding of the potential primordial soup, as from present-day biology. For example, though DNA is the prevalent information-carrier nucleic acid today, it is considered a later molecule compared to RNA due to the ribose's reducibility to deoxyribose [108]. Research on non-canonical nucleic acids[1], suggests other simpler nucleic acids might have preceded RNA. This early set of organic molecules likely comprised a diverse array of molecules, possibly also including non-canonical nucleobases and amino acids, with selection pressures leading to the current canonical nucleotides and amino acids to be selected [10, 79].

Emergence of nucleic acids

Nucleic acids play a crucial role in modern biology, primarily serving as carriers of genetic information. While proteins (enzymes) perform most of the catalytic functions, nucleic acids are implicated in storing and transmitting genetic data. Given the complex nature of modern proteins, indicative of their evolutionary development, it is theorized they were preceded by a simpler chemical mechanism for replication. The RNA world hypothesis posits that RNA could have played a dual role in early life, possessing both informational and catalytic properties [48]. The structure of the contemporary ribosome, where the active site is a central core of RNA and the peptides are mostly in the outskirts, supports this theory [6, 146, 153].

The interpretations of the RNA world hypothesis vary considerably. For the purpose of this study, a core definition as proposed by Robertson et. al [109] is adopted, which focuses on three fundamental assumptions:

- At a certain stage in the evolution of life, genetic continuity was maintained through RNA replication;

[1] such as α-L-threose nucleic acid (TNA) [152]. Such non-canonical nucleic acids are also known as XNAs

- Watson-Crick base-pairing was key to replication;
- Genetically encoded proteins were not involved as catalysts.

These assumptions do not exclude the potential existence of other replicating and evolving molecules preceding RNA, akin to how RNA is believed to precede DNA and complex proteins. They also do not disregard the possibility of RNA's co-evolution with other smaller molecules, such as amino acids or small peptides. These could have co-evolved as the peptides protected RNA from degradation or stabilized specific conformations [95, 96, 137].

In the evolutionary time line, this work explores the time frame where this core RNA world theory applies. This comprises the time period between the prebiotic synthesis of very small RNA oligomers and the development of replication networks based on RNA. The evolution of nucleic acids likely began with the assembly of simpler, shorter sequences that underwent elongation, replication, and molecular evolution to yield more advanced structures. Such functional sequence are often long, highly structured and rare [129]. For this reason, the process of elongation and selection likely occurred simultaneously. Early ribozymes probably emerged from initial pools synthesized through non-enzymatic oligomerization, composed predominantly of short RNA segments [27, 138, 139] and exhibiting low diversity [30]. The shift of nucleotide, sequence and structural diversity over time is then tied with the selection for functional motifs.

Sequence selection mechanisms

Biases in nucleic acid sequences can arise through both intrinsic and extrinsic selection factors. Mechanisms of replication and elongation may inherently favor certain nucleotides or sequences due to stability, kinetics, or recognition by the replication apparatus [32, 51, 70, 84]. Additionally, environmental selection pressures can introduce biases in sequence evolution. For instance, in aqueous environments, sequences capable of encapsulation and local accumulation may have a survival advantage [35, 36, 55], while in environments more exposed to the irradition of the early sun, the presence of sequence-specific UV-induced mutations [24, 68, 132] may become deleterious for propagation, favouring sequences that do not have those UV-damage prone motifs.

In Chapter 1 cyclic flux changes are explored as an extrinsic factor for selection of phase separating sequences. In this project the liquid-liquid phase separation (LLPS) of mixtures of DNA is triggered, in a system where part of the dilute phase is cyclically re-fed. Over cycles, specific sequences enrich in the dense phase, effectively selecting them out of the original pool.

In Chapter 2 the influence of replication on pools of short biased oligomers is studied as an intrinsic factor. Biased random pools of DNA strands were subjected to templated polymerization with *Bst*. Periodicity emerged through replication of the initially unstructured pools. This selection was accompanied by a shift in the overall nucleotide composition.

In Chapter 3 an non-enzymatic RNA replication chemistry is characterized. The previous systems were proof-of-principle approaches with DNA and, in the case of Chapter 2 with an enzyme. This allows for higher stability and the exploration of evolutionary time scales, due to the faster enzymatic kinetics. In this chapter however, a plausible replication mechanism relying on the templated ligation of short RNA fragments containing 2′, 3′-cyclic phosphates is described.

1 Nucleic acid selection by cyclic phase separation

Summary

Cyclic processes such as daily temperature oscillations and tides are ubiquitous and could have triggered oligonucleotide phase separation on early Earth. Here, nucleic acid phase separation is proposed as a mechanism to select subsets of short oligomer pools. This was experimentally realized through the discontinuous feeding of an initial pool to a chamber were the dense phase sedimented and was therefore sequestered. Sequence-specific enrichment of DNA in the sedimented dense phase was shown, in particular of short 22-mer DNA sequences. The underlying mechanism selects for complementarity, as it enriches sequences that form fully base-paired networks. A 4-letter alphabet (i.e. with **A**, **T**, **C** and **G** incorporated in the sequences) was found to be necessary for phase separation to occur within this system, potentially due to higher prevalence of non-specific secondary structures in binary (**A**/**T** or **G**/**C** only) systems. One single base mutation to a binary system was found to trigger phase separation. These findings provide an example of a selection mechanism towards sequence networks with high cross-complementarity. This enrichment in inter-sequence interactions could lead to the emergence of auto-catalytic oligonucleotides.

This chapter was published by Barlotucci, Serrão and Schwintek et. al [8] in PNAS and is here adapted and reprinted in part with permission from PNAS. It was based on a collaborative project with Giacomo Bartolucci and Christoph A. Weber. Full article attached in the List of Publications.

1.1 Motivation

Liquid-liquid phase separation (LLPS) is the process by which a homogenous solution of molecules separates into two distinct phases: a dense and a diluted phase. Nucleic acids are known to be capable of phase separation, forming condensates such as coacervates [5, 60], liquid crystals [93, 155] or hydrogels [94, 149]. The intermolecular interactions between strands of nucleic acids lead to a local enrichment of certain oligonucleotides. When these interactions are sequence-specific, such as Watson-Crick base-pairing, the strands recruited and enriched in the dense phase are highly correlated, as they are to some degree complementary. This subset of strands has a evolutionary advantage in comparison to the remainder of the pool which is two-fold. On the one hand, the dense phase provides with a microenvironment which shelters from dilution and potentially hydrolysis. On the other hand, higher-order complementary networks are the starting point for the development of functional nucleic acids capable of catalysis [99]. Depending on the environmental conditions, LLPS could have played a role in the selection and evolution of prebiotic oligonucleotide pools.

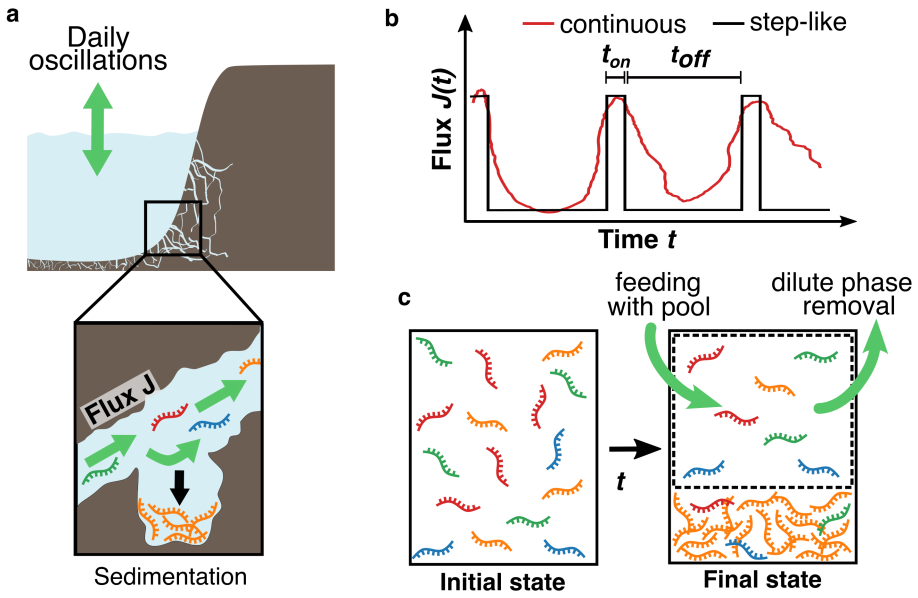


Figure 1.1: Prebiotic rocky pore subjected to continuous oscillating flux of material selects for phase-separating sequences. a, Scheme of a continuous oligonucleotide flux through porous rocks on early Earth. By sedimentation, phase separating sequences (depicted in orange) are enriched in the bottom forming a dense phase. Meanwhile the remaining dilute phase is washed-out. **b**, The oscillatory continuous flux can be approximated by a step-like cyclic flow profile experimentally. **c**, Experimental abstraction of **a** where a closed system is fed in a step-like manner. Cyclically, the dilute phase is removed and the same volume of initial pool is provided. The feeding and removal steps are separated in time from phase separation kinetics.

Environmental factors were involved in the selection of some sequence subsets over others on early Earth, exerting a selection pressure, with varying influence over time and geological setting. A range of geological environments, including shallow water ponds [26], freeze-thaw

cycles within ice [62] and hydrothermal vents [23], have been suggested as potential settings where the origins of life might have emerged during the late Hadean and early Archaean [57].

An underwater rocky pore is a plausible prebiotic setting that could act as a physical reservoir for phase separated oligonucleotides, see Figure 1.1 **a**. An alternating flux of a pool of oligonucleotides (Figure 1.1 **b**) could lead to part of the pool to be retained in the interim time between cycles. A subset of sequences that interact tightly could then form a dense phase and sediment, being therefore more sheltered from the next cycle's wash-out than the dilute phase. Over time, the dense phase would grow, by recruiting more sequences from the pool, potentially enriching in certain motifs, and altering nucleotide composition. Without phase separation, the oligomer composition approaches or remains at the composition of the environment, independent of oligonucleotide sequence. This work hypothesizes that such a setting can provide a physical mechanism of selection of specific sequences.

1.2 Scientific approach

The hypothesis behind these experiments was: "Do sequences that phase separate *i)* selectively enrich in the sedimented dense phase and *ii)* are therefore protected from wash-out of the remaining pool?". To answer these questions an initial pool of sequences was artificially designed, with some of the sequences (or sequence combinations) being able to phase separate while the remaining are not. Mimicking the step-wise flux profile (Figure 1.1 **b**), a series of cycles of feeding and dilute phase removal were manually done, Figure (Figure 1.1 **c**). An analysis of the removed dilute phase samples and the remaining dense phase, at the final state, allowed to differentiate the composition of both phases. This approach implies a series of decisions and assumptions that facilitate researching the hypothesis in a laboratory and isolating the contribution of each of the parameters to strand selection.

- **DNA vs. RNA**

  RNA is considered to be an earlier molecule than DNA, both due to its capacity to information storage and catalysis [48], see Introduction for more details on the RNA World theory. Since RNA is more labile than DNA to hydrolysis [1] and RNAses are common contaminants, DNA is often used as a replacement in proof-of-principle studies such as this work. Both nucleic acids share similar properties, particularly comparable base-pairing dependence on temperature and secondary structure formation [118, 126]. Phase separation through Watson-Crick base-pairing as the source intermolecular interaction was the focus of this work as it is sequence specific, and for this reason DNA was used as the polymer for the initial pools.

- **Artificial strand design vs. pool generated in prebiotic conditions**

  Prebiotic nucleic acid synthesis is one the cornerstones for understanding the molecular origin of life. Several different condensation chemistries have been explored in this context yielding pools of mostly short length [27, 31, 72, 142] and/or low diversity (e.g. only consisting of `A` or `G`) oligomers [43, 88, 139]. Most of the oligonucleotides that have been demonstrated to phase separate in the basis of forming long base pairing network have been at least about 20 nt long and consisting of the full 4-letter alphabet [60, 94, 117, 149] (with the exception of [87] which demonstrated hydrogel formation of `A`/`T` only or `G`/`C` only strands). As the generation of diverse oligonucleotide pools through prebiotic

synthesis in a length regime of tens of base pairs is not yet understood, an artificially designed and synthesized pool was used instead.

- **Low diversity pool vs. random pool**

  Prebiotic pools are not thought to be fully random (equal incorporation of the 4 bases) since the nucleotide abundance in the environment is imbalanced due to different rates of nucleotide formation and degradation, depending on the conditions [22, 30, 70, 71, 145] and to the different rate of condensation [27, 32, 84]. Assuming a length of approximately 20 nt, the possible sequences with a 4-letter alphabet would be $4^{20} \approx 1.1 * 10^{12}$ and with a 2-letter alphabet $2^{20}) \approx 1.1 * 10^{6}$. Even taking into account the asymmetry in nucleotide incorporation, the sequence space resulting from condensation would still be within these two boundaries and therefore too large to screen in laboratory time scales. Such a large sequence space entails that individual sequence would be most likely too dilute to interact. While phase separation is hypothesized here to select for particular sequences due to base pairing networks, likely other selection mechanisms were at play, both before and concomitantly, that would up-concentrate nucleic acids to the critical phase separation concentration. For these reason, the initial pool used here was artificially "pre-selected" as a proof-of-concept approach. Sequences that are known to phase separate were mixed with others that do not, while still having high cross complementarity, at concentrations above the critical phase separation concentration.

- **Reaction tube vs. rocky pore**

  While it is plausible that many selection mechanisms were at play in the origin of life, as previously discussed, using a real rocky pore would introduce additional interactions with the system that would not allow to isolate the influence of phase separation. Mineral surfaces have been shown to be selectively adsorb longer oligonucleotides, effectively enriching them [29, 85]. Inhomogeneities in temperature across the pore were also extensively studied for selectively accumulating sequences [81, 86, 116]. In turn, using a reaction tube assures homogeneous temperature across the sample, with limited convection and condensation, and a chemically inert inner surface.

1.3 Results and Discussion

1.3.1 Sequence design

In order to experimentally test selection through phase separation, short nucleic acid sequences were designed and tested for their capacity to undergo LLPS and form condensates. This design aimed to create a sequence pair (SP) that interacts robustly and which would be able to phase separate only when both sequences are simultaneously present. The choice of an SP, in opposition to a single phase separating DNA sequence, aims to prevent the formation of stable self-folded structures and increases the chances of incorporation into a network.

Previous work has been done to design and characterize DNA sequences that phase separate. These have the common denominator that the secondary structures formed can grow through the continuous incorporation of more sequences in a network formation. Eventually, through the formation of networks, phase separation occurs. These have been reported both for systems of single sequences [86] or for systems of two or more cooperative sequences [3, 12, 60, 94, 117, 148, 149]. The formation of such branched DNA aggregates with short sequences leads

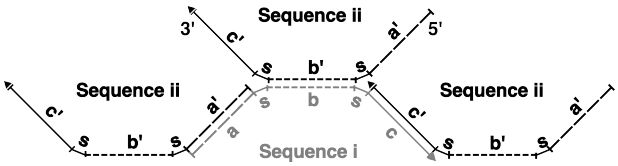


Figure 1.2: Schematics of expected secondary structure for designed SPs. Sequence i is composed of three segments a, b, and c with spacer s. Its pair ii consists of reverse complements a', b' and c'. The inverted arrangement of a' and b' creates a network from three binding sites and prevents the formation of a linear double-stranded duplex. The spacers were included to add more flexibility to the individual segments and avoid angular constraints.

to a dense phase. However, these studies investigated rather long strands and solely [3] studied phase separation of short DNA strands in the length regime of 20 to 25 bp. Understanding the process of phase separation within the range of tens of base pairs is particularly significant within the prebiotic chemistry framework as early, non-enzymatic nucleic acid synthesis was initially slow and likely yielded short fragments.

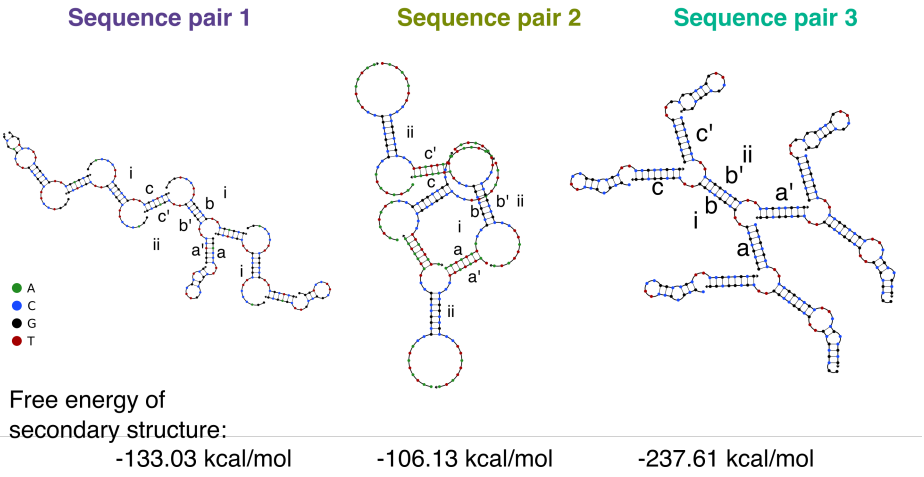


Figure 1.3: NUPACK simulation for the three designed SPs. Sequence i is composed of three segments a, b, and c with spacer s. Its pair ii consists of reverse complements a', b' and c'. The inverted arrangement of a' and c' creates a network from three binding sites and prevents the formation of a linear double-stranded duplex.

SPs were designed to be composed of sequences with three binding regions each (a, b, and c or a' , b' and c') which are separated by dimeric spacer sequences 1.2. These segments are individually reverse-complementary, i.e., a' is the reverse complement of a. However, the sequences i and ii are not the reverse complement of each other, because the order of the individual segments is not reversed (since i = 5′ a,b,c 3 ′ and ii =5′ a',b',c' 3′). The goal was to design a SP composed of two sequences containing at least 3 unique binding regions such that every strand bound to the first one results in an additional binding site available. This yields $b = 3 + n$ as the number of vacant binding sites (b) with n being the number of strands bound in the network. Therefore, the strand network will grow faster the more strands are already bound. This choice avoids a fully complementary double-stranded structure and allows each sequence to bind to three other sequences, forming a branched structure.

DNA oligonucleotide systems were designed using the NUPACK software package 3.2.2 [154]. This search was performed in an automated manner, with the appended script in Appendix 1.A and three systems that formed the planned secondary structure were chosen. The shortest possible SPs that still formed a branching network upon hybridization were chosen. While keeping the center part of each sequence restricted to **G** or **C** only, the outer sections ("arms") were varied, to either consist of all four bases (SP 1), **A**/**T** only (SP 2) or **G**/**C** only (SP 3). The resulting output sequences were iteratively mutated afterwards. Upon each mutation of the strands, the sequences were analysed using the Analysis segment of the online NUPACK tool in order to check their ability to form a network. Figure 1.3 shows the final iteration for SPs 1, 2 and 3, where the sequences form the intended secondary structure. Specifically, the presence of secondary structures that included strands bound to three other strands were considered indicative of network formation. Spacers of two bases were inserted between the three segments of each sequence. These were chosen to be either **TT** (SP 1 and 3) or **CC** (SP 2) in order to not have complementarity with the arms and reduce their participation in the overall secondary structure. Inspired by Ref. [94], these spacers allow for more flexibility of the segments by minimizing angular constraints. The corresponding sequences for each of the systems are shown in Table 2.1.

Table 1.1: Sequences of the designed SPs, divided into the three segments. Sequence *i* is shaded in light grey and sequence *ii* is shaded in dark grey. For SP 1, the nucleotide 'mutation' is highlighted in red.

Seq. pair	Sequence		a/a′	s	b/b′	s	c/ c′	
1	i	5′	GGACCC	TT	CGGCCG	TT	CGCTCG	3′
	ii	5′	GGGTCC	TT	CGGCCG	TT	CGAGCG	3′
2	i	5′	AATATATA	CC	GCGGCCGG	CC	TATAATAA	3′
	ii	5′	TATATATT	CC	CCGGCCGC	CC	TTATTATA	3′
3	i	5′	GGCGCGCG	TT	GCGGCCGG	TT	CGCGGCGG	3′
	ii	5′	CGCGCGCC	TT	CCGGCCGC	TT	CCGCCGCG	3′

1.3.2 Thermal melting curves

The stability of a nucleic acid complex can be measured by a thermal denaturation experiment. Varying the temperature changes the fraction of DNA strands that is in the random coil or single-stranded state versus the fraction that is in the double-stranded form. The melting temperature (T_m) is defined as the temperature at which half of the DNA is in single-stranded (ssDNA) state, and is generally higher for longer or GC-rich sequences. The fraction of bound sequences is commonly measured by absorbance at 260 nm or fluorescence with an intercalating dye, but many other methodologies have been developed [82].

Melting curves for all SPs were measured in triplicates at neutral pH and with close to physiological salt conditions, Figure 1.4. The methodology is detailed in Section 1.5.3. The mixtures contained SYBR Green I, which is an intercalating dye that preferentially binds to double-stranded (dsDNA) [161]. The SYBR Green I fluorescence signal across temperature was converted into a double-stranded fraction (fraction bound) through the baseline normalization described in [82]. This analysis yielded a T_m of 57°C for SP 1, 71°C for SP 2 and 65°C for SP 3. The shortest SP, 1, expectedly displayed the lowest melting temperature. However, between

SP 2 and 3, which both are 29 nt in length, the lowest melting temperature was found for 3, which has a lower GC-content. While the SPs were designed to fold into the structure shown in Figure 1.2, other alternative secondary structures are possible and may be contributing to the difference in melting temperature.

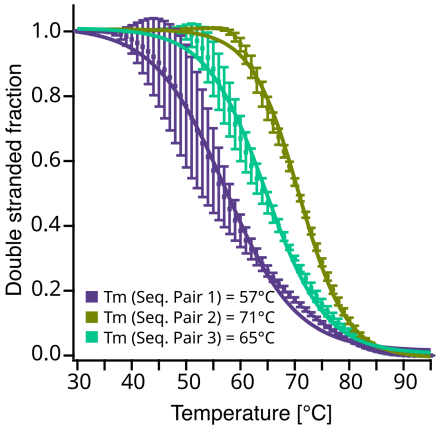


Figure 1.4: Melting curves of SP 1, 2 and 3. The mixtures contained both strands of each system at 2 µM each DNA strand, 10 mM Tris buffer pH 7, 125mM NaCl and 10 mM $MgCl_2$. SYBR Green I concentration was 5X. Potentially due to the shorter duplex forming segments (a, b and c), the T_m of SP 1 is the lowest. Data depicted as mean ± one standard deviation for three independent replicates.

1.3.3 Condensation and sedimentation

The phase separation propensity of the designed systems was measured with time-lapse fluorescence microscopy. In particular, each system was imaged over time in thin temperature-controlled microfluidic chambers, Section 1.5.4. For each SP, both strands were at 25 µM (remaining buffer concentrations described in Section 1.5.2). The samples were then slowly cooled at a rate of 6 K/min to 15°C and incubated at that temperature for at least 3 h. This temperature is lower than the melting temperature for all of the SPs, corresponding to a scenario where most of the sequences are bound (Section 1.3.2).

Microscope images for all three systems are shown in Figure 1.5 **a**, where "0h" corresponds to the moment when the cooling step has reduced the temperature to 15°C. This was taken as the initial micrograph since from that time point onwards the temperature did not change, and there was no temperature influence on fluorescence. Within about 10 min, the first dense phase DNA nucleated for SP 1, grew within 1.5 h to a size of a few micrometers and sedimented. As a result, a phase of sedimented DNA accumulated at the bottom of the chamber. The DNA concentration in the dense phase increased up to 13-fold for SP 1 (Fig. 1.2 **b**).

The total amount of molecules that sedimented saturates at about 8% of the initial material at about 3 h, decreasing then only slightly over time (Figure 1.5 **c**, black data points). The height of sedimented DNA reached a maximum of about 100 µm at 3 h but then compacted to about half the height Figure 1.5 **c**, red data points). This volume reduction over time has been observed in literature for systems composed of longer DNA strands [94]. This phenomenon, denominated contractile aging of DNA hydrogels, is attributed to the rearrangement of base-pairing within the network. Sequences are in dynamic equilibrium and through the shuffling

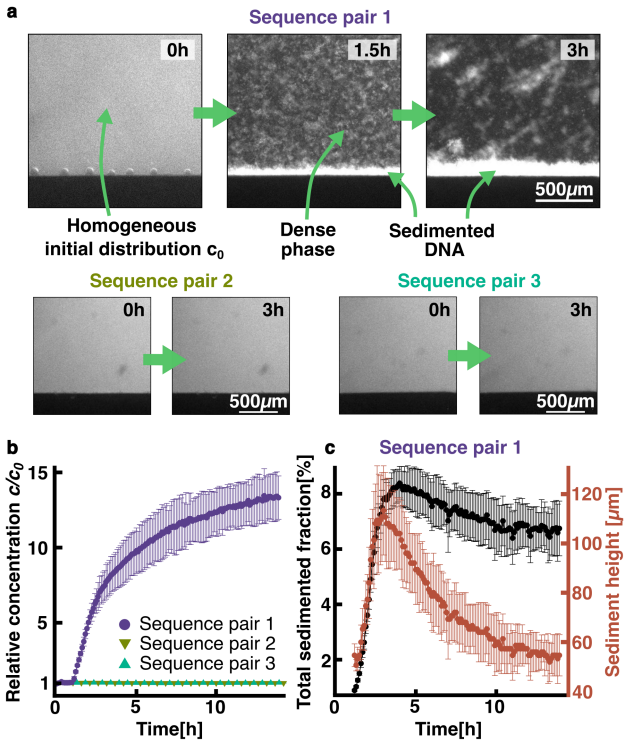


Figure 1.5: Phase separation and sedimentation behavior of three SPs. a, Fluorescence time lapse images in a vertical, 500μm thin microfluidic chamber to prevent convection flow, Section 1.5.4. DNA concentration was 25 μM (for each sequence in the SPs) in a 10 mM Tris-HCl pH 7 buffer with 10 mM $MgCl_2$ and 125 mM NaCl. Fluorescence labelling was provided by 5X SYBR Green I. After cooling from 65°C to 15°C, SP 1 phase separated and sedimented to the bottom of the chamber. SPs 2 and 3 did not form a dense phase, and thus showed a homogeneous fluorescence signal. **b**, SP 1 showed an up to 13-fold enhanced relative concentration while SPs 2 and 3 showed no phase separation. **c**, The sedimentation behaviour of SP 1 is studied by measuring its fluorescence. The total amount of sedimented DNA plateaued at 6 to 8% after 5 h while the sedimented DNA contracted about two-fold. The sticking of dense phase DNA to the chamber walls could not be fully prevented. Data depicted as mean ± one standard deviation for three independent replicates.

of hybridization partners, a more favourable state is reached, where a larger fraction of strands have their arms subsegments bound, minimizing the strain in the overall structure.

No dense phase was observed for SPs 2 or 3, despite having longer arms (a,a') and (c,c') than SP 1, which should lead to stronger binding interactions through base-pairing, and therefore more stable networks. A possible explanation for this would be that SPs 2 and 3 have arms with a binary alphabet composed of only **G**/**C** or **A**/**T**, respectively. This could cause non-specific hybridization (especially within G- and C-rich) sequences, which can lead to alternative secondary structures, such as hairpin-rich configurations, internal loops or G-quadruplexes [128]. The reduced alphabet also increases the likelihood of finding a sequence that binds partially, effectively preventing network formation.

1.3.4 Cyclic phase separation

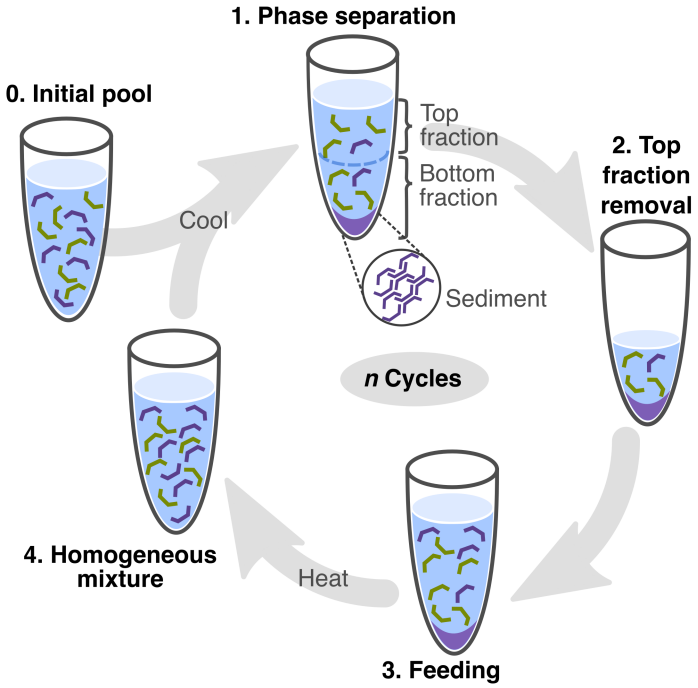


Figure 1.6: Cyclic experimental protocol to mimic a rocky-pore subjected to an oscillating flux of material. An initial pool containing both oligonucleotides that phase separate (purple) and that do not (green) is incubated at room-temperature for hours. This leads to liquid-liquid phase separation followed by sedimentation, where the dense phase mostly consists of the purple oligonucleotides. Part of dilute phase is then "washed-out", or manually removed, and re-fed with the same volume of initial pool. The mixture is then subjected to a high temperature, that homogenizes the mixture and simulates a worst-case scenario of memory loss, ending one full cycle.

Based on the observation that SP 1 can form a condensed DNA-rich phase, an experimental implementation of the selection mechanism shown in Figure 1.1 **c** that relied on a cyclic influx of material was performed. The phase-separating DNA was subjected to cyclic feeding steps by replacing the top fraction of the supernatant phase in the vial by the pool. The experimental approach of such a cycle is schematized in Figure 1.6. The cycle starts with the phase separation of a homogeneous initial pool (step 0.) into dense and a dilute phases (step 1.).

The theory developed for multi-component phase separation subject to cyclic material exchanges suggests that exchanging the complete supernatant phase with the pool reaches the final stationary state with minimal amount of cycles (Appendix 1.B for the theoretical description). However, a complete removal of the supernatant phase by pipetting turns out to be experimentally difficult since this also risks to remove condensed DNA. For this reason, in step 2. of Figure 1.6, only a top fraction is removed. This corresponds in this approach to half of the total volume and is entirely composed of dilute phase. The bottom fraction therefore has both the totality of the dense phase and a fraction of the dilute phase.

Afterwards, an equal volume of initial pool is fed to the system in step 3., in this case specifically corresponding to half of the total volume. To avoid kinetically trapped states of

condensing oligonucleotides, annealing and melting steps are included in the cycle (Fig. 1.6, steps 3 to 4 and back to step 1). This procedure enabled a fast relaxation to thermodynamic equilibrium after each feeding step.

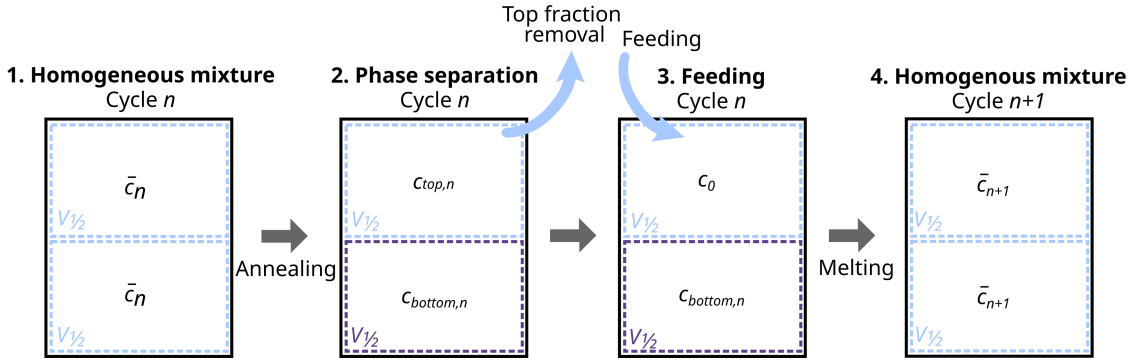


Figure 1.7: Schematic of a feeding cycle and the corresponding concentrations for top and bottom volume fractions.

The system composed of equal fractions of the SPs 1 and 2, where solely the SP 1 showed phase separation before. As control a system composed of SPs 2 and 3 was considered, where the formation of a condensed DNA phase was not observed (Fig. 1.5 **a**). The strand concentrations of each system in the top and bottom fractions of the vial was determined using HPLC[1] UV absorbance (Section 1.5.5 for methodology). The concentration in the top fraction could be measured directly, as it is removed in each cycle. However, the concentration of the intermediate bottom fractions cannot be determined, since they proceed into the next cycle, and only the last bottom fraction can be retrieved and directly analysed.

To determine the concentration of the bottom fraction for each cycle n a mass balance is used (Figure 1.7). Using the concentration of the removed top fraction obtained through HPLC analysis, and the known concentration of the initial pool c_0, all the concentrations of the intermediate bottom fractions are calculated. At step 1., the system phase separates which conserves the total mass:

$$\bar{c}_n \, 2V_{\frac{1}{2}} = c_{\text{top},n} \, V_{\frac{1}{2}} + c_{\text{bottom},n} \, V_{\frac{1}{2}} \,, \qquad \text{and thus} \quad c_{\text{bottom},n} = 2\bar{c}_n - c_{\text{top},n} \,, \tag{1.1}$$

where $V_{\frac{1}{2}} = V/2$ denotes half the total volume (V is the total volume of the system), $c_{\text{top},n}$ and $c_{\text{bottom},n}$ the concentrations of material in the top and bottom fractions in the n-th cycle, respectively. Note that the bottom fraction includes the sedimented dense phase. Moreover, $\bar{c}_n$ describes the concentration in the whole volume during the n-th cycle, which is conserved during phase separation. At step 2. the top half of the system is removed and fed at step 3 by the pool of concentration c_{pool}. The corresponding mass balance reads:

$$c_{\text{pool}} \, V_{\frac{1}{2}} + c_{\text{bottom},n} \, V_{\frac{1}{2}} = \bar{c}_{n+1} \, 2V \,. \tag{1.2}$$

[1]short for High-Performance Liquid Chromatography

Using Equation (1.1) to substitute Equation (1.2), $\bar{c}_{n+1}$ can be written as a function of $c_{\text{top},n}$, c_n and c_{pool}. This relation can be used to obtain an equation for $c_{\text{bottom},n+1}$, through substitution in Equation (1.1):

$$\bar{c}_{n+1} = \bar{c}_n + \frac{c_{\text{pool}} - c_{\text{top},n}}{2}\,, \tag{1.3}$$

$$c_{\text{bottom},n+1} = 2\bar{c}_n + c_{\text{pool}} - c_{\text{top},n} - c_{\text{top},n+1}\,. \tag{1.4}$$

Since for $n = 0$, $\bar{c}_n = c_{\text{pool}}$ and by measuring the top fraction concentrations $c_{\text{top},n+1}$, all the bottom fraction concentrations using Equation (1.4) can be calculated.

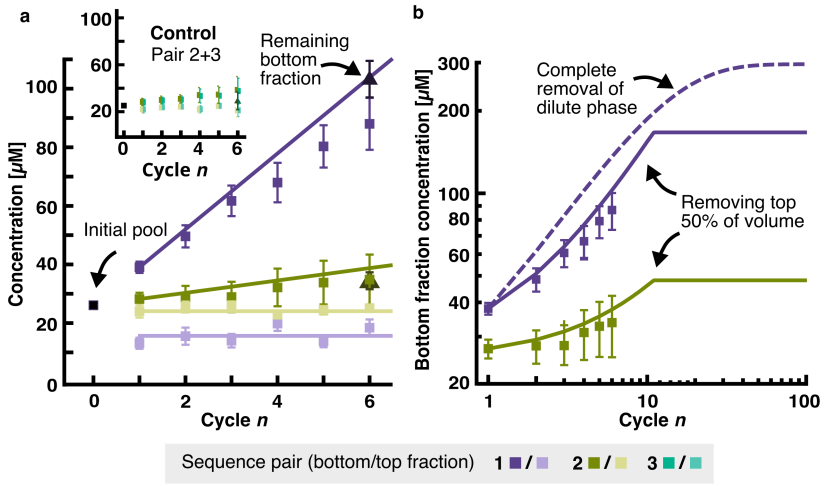


Figure 1.8: Cycles of phase separation and feeding steps select specific oligonucleotide sequences from the initial pool. **a**, The initial pool contained a 25 μM concentration of SPs 1 and 2. After sedimentation, the top half of the volume (top fraction) was removed (α= 0.5, see Appendix 1.B) and fed after each cycle with the same volume of the initial pool. Using quantification with HPLC, SP 1 (purple) was found to be enriched while the concentration of SP 2 (green) remained approximately constant. The same flat dynamics was found for a control system using non-phase-separating SPs 2 and 3 (inset, due to significant data overlap markers are halved for clarity). In addition, the concentration of all supernatants and the final sediment were measured by absorbance at 260 nm (triangular markers, Section 1.5.5). **b**, Solid lines show theoretical predictions using the mass-balances detailed in Figure 1.7. The bottom fraction concentration saturates once the sedimented DNA has filled the bottom fraction of the chamber. If the whole supernatant could have been removed at each step of the cycle, only slightly amplified selection would be predicted (dashed line).

The kinetics over six feeding cycles for the system composed of SPs 1 and 2 were monitored (Figure 1.8 **a**) and compared it to the non phase-separating control (inset). Both systems were intialized with equimolar concentrations of the two respective SPs. The concentrations for the control hardly increased per cycles with slopes about or less than 2 μM/cycle. This non-zero increase is probably due to the adhesion of strands to the vials surface. For the phase-separating system with SP 1 the concentration strongly increased, approximately linearly with a slope of about (10.2 ± 0.4) μM/cycle (purple), while the SP 2 in the mixture got only weakly enriched by the cycling. This observation confirmed that specific sequences could get selected by phase separation from the dilute phase.

In the experiments, the selection occurred concomitant to the growth of the condensed phase. As cycles proceeded, the dense phase grew and lead to an increase of the concentration of the bottom fraction (Figure 1.8 **a**). In contrast, the concentration of the top fraction remained constant at about 14.7 μM (Fig. 1.8 **a**, light purple). A constant supernatant concentration during cycles implies that system remained on the same tie line while the volume of the sedimented DNA was growing. This corresponds to the simple theoretical scenario where the system is initialized at the pool tie line. The experimental results for the bottom fraction concentration were compared with the theoretical model presented in Appendix 1.B and shown in Figure 1.8 **b**. Since the experimental selection kinetics occurred on a single tie line, the dense phase and dilute phase concentrations remained constant over time. For the dilute phase concentration the experimental concentration value of the top fraction was used. The sediment concentration could be estimated for the theory using the experimental value for the initial average sequence concentration and the initial sedimented DNA size, taken from Figure 1.5 **c**.

Using these values the theoretical results could be computed, Figure 1.8 **b**, solid lines. Based on the agreement between experiment and theory, the theory was used to extrapolate the selection kinetics for a larger amount of cycles. For the experimental partial removal of the dilute phase it was found that the selection approximately doubled after 20 cycles. The selection kinetics saturates because the condensed phase has grown to the volume corresponding to the bottom fraction. Finally, the theory could also consider the ideal case of a complete removal of the dilute phase. For this ideal case the SP can enrich by two-fold better than the for partial dilute phase removal and more than 10 fold compared to the initial pool (Figure 1.8 **b**, dashed lines). This experimentally demonstrates that discrete cycles of feeding, i.e., replacing the dilute phase with a reference pool, leads to the enrichment of specific SPs based on the formation of a dense phase which confirms the theoretically proposed selection mechanism in Appendix 1.B.

1.3.5 Influence of SYBR Green I

The initial pool mixtures in Section 1.3.4 also contained 5X SYBR Green I. This was done in order to be able to apply the knowledge about the dynamics of phase separation of individual systems, obtained through fluorescence microscopy in the presence of SYBR Green I and shown in Figure 1.5, to the development of theoretical data for the cyclic phase separation. The presence of SYBR Green I however is not necessary for HPLC detection and could interfere with secondary structure and therefore dense phase formation.

In order to understand the potential influence of SYBR Green I on the phase separation and sedimentation of the SP 1, the feeding cycle experiment presented in Fig 1.8 **a** was repeated without including SYBR Green I 5X. The concentration of each of the SPs over cycles of refeeding, for a mixture of SP 1 and 2 is plotted in Fig. 1.9 and for SPs 2 and 3 in the corresponding inset. The quantification was performed with HPLC at 260 nm UV detection. Similarly to the original experiment containing SYBR Green I, the concentration of SP 1 increases linearly in the bottom fraction and is depleted in the top fraction. Additionally, SP 2 also does not partition between the top and bottom fractions, maintaining an approximately constant concentration around 25 μM, the initial concentration. For the control experiment, with SPs 2 and 3, the concentration of both systems stays approximately constant over cycles.

The general behaviour of the pools is similar in the presence and absence of SYBR Green I at 5X concentration. However, the linear increase in concentration of SP 1 in the bottom

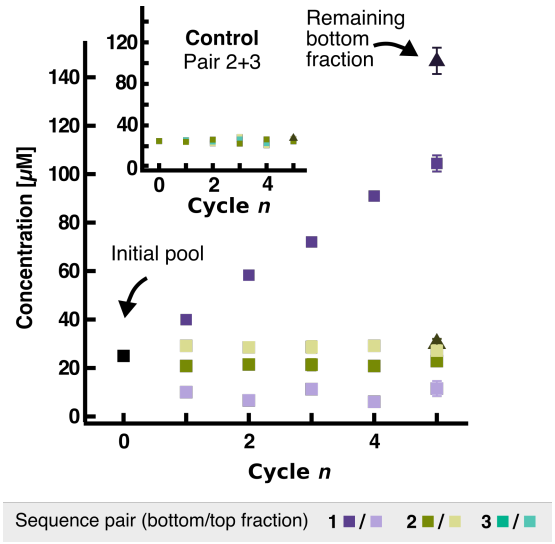


Figure 1.9: Concentration of the SPs over five cycles of refeeding in the absence of SYBR green I. The initial pool contained 25 µM of SPs 1 and 2. The sequences were slowly cooled to 15°C at 6 K/min similarly to the remaining sedimentation experiments (see Figure 1.8). Similarly, the volume removed for the top fraction was half of the total volume. The quantification was performed with HPLC through the measurement of absorbance at 260 nm. For the mixture of SPs 1 and 2, the concentration of SP 1 increased linearly with a slope of approximately (16.1 ± 0.3) µM/cycle . Data are the average of three independent repeats and error bars correspond to one standard deviation of the mean.

fraction is about 1.6 times higher without it, with an approximate slope of (16.1 ± 0.3) µM/cycle (compared to 10.2 ± 0.4 with SYBR Green I). This difference could be explained by the fact that SYBR Green I is an intercalating agent, which has been described to bind to the minor groove of a DNA helix [161]. This could interfere with the structure of the dsDNA, as well as changing its T_m [52]. This has also been reported for other intercalating dyes [66], and could influence the phase separation.

1.3.6 Salt and pH dependence of SP 3

In order to understand the dependence of the SP 1 phase separation on the salt and pH conditions a screening was performed. Firstly, the buffer conditions were fixed to 10 mM TRIS pH 7 and both NaCl and $MgCl_2$ were varied between 0-250 mM and 0-10 mM respectively. The fluorescence micrographs are shown in Figure 1.10 **a**. In the absence of $MgCl_2$, phase separation is only observed at the highest concentration of NaCl tested (250 mM). In the presence of 10 mM $MgCl_2$, phase separation occurs for any concentration of NaCl in the range.

Additionally, the influence of the buffer conditions were tested by fixing the salt concentration to the standard conditions used throughout the manuscript (10 mM $MgCl_2$ and 125 mM NaCl) and varying the pH between 5.5 and 9 (see Fig. 1.10 **b**). Coexisting phases are observed for both pH 7 buffers screened, Tris and HEPES[2], suggesting that the phase separation is

[2]short for (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid)

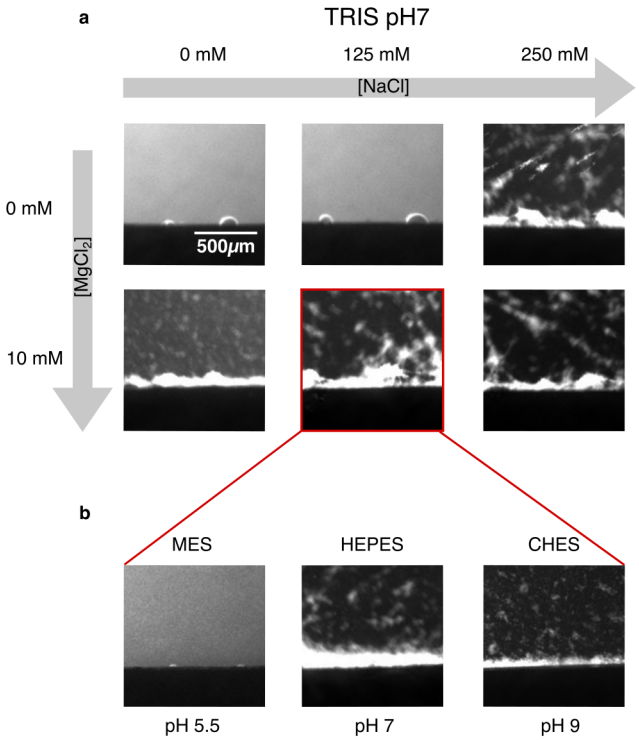


Figure 1.10: pH and salt concentration screening for the phase separation of SP 1. **a**, Mixtures with 25 µM of SP 1 in 10 mM TRIS buffer pH 7 and varying concentrations of NaCl and $MgCl_2$ were analysed through fluorescence microscopy (5X Sybr Green I was added). The temperature protocol is as described for Figure 1.5, denaturing the samples at 95°C, then fast cooling to 65°C and lastly cooling to 15°C at a rate of 6 K/min. The micrographs show a cut-out of the bottom of the microfluidic chamber after at least 3 h have passed since the temperature reached 15°C. **b**, Keeping the salt concentration fixed at 125 mM NaCl and 10 mM $MgCl_2$, the buffer (and consequently the pH) was now varied between 5.5 and 9. Concentration of SP 1 was 25 µM, SYBR Green I was 5X and the temperature protocol was kept the same. The micrographs show either homogeneous fluorescence for the conditions no phase separation is observed, or a layer of sedimented DNA at the bottom of the chamber.

not buffer dependent. Decreasing the pH, by using MES[3] pH 5.5, led to no observable DNA aggregates. Meanwhile, increasing the pH to 9, now buffering with CHES[4], did not have detrimental effect to the phase separation. Different buffers were used in order to assure buffering capacity for the desired pH range.

1.3.7 4-letter alphabet vs. 2-letter alphabet

Additionally, five more 22 nt SPs similar to SP 1 were analysed, systematically varying the base composition of the **G**/**C**-only binding segments. Notably, the inclusion of a single **A**/**T** nucleotide in the outer segments a/a' or c/c', thus utilizing the full 4-letter alphabet, proved crucial for phase separation.

[3] short for 2-(N-morpholino)ethanesulfonic acid
[4] short for (N-cyclohexyl-2-aminoethanesulfonic acid)

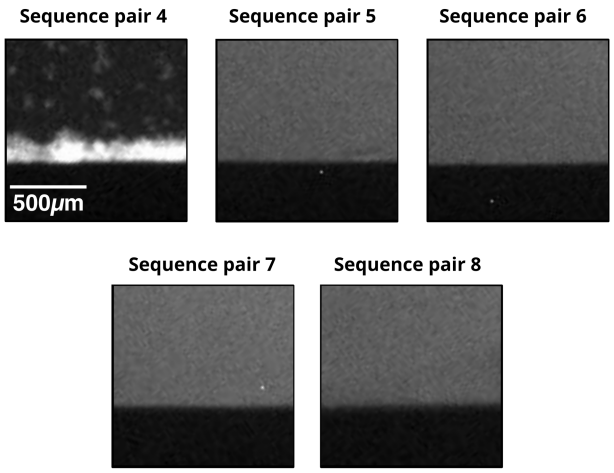


Figure 1.11: Phase separation and sedimentation behaviour of the additional SPs. Samples were initially heated to 60°C to ensure dehybridization and then slowly (6 K/min) cooled to 10°C. Buffer conditions were 10 mM TRIS pH7, 10 mM $MgCl_2$, 500 mM NaCl, 25 µM of each DNA sequence and 5X SYBR Green I). Images were taken continuously and the samples were given up to 7 h to sediment. All micrographs depict the bottom of the well after 7 h. The only system to phase separate and sediment is SP 4, which is also the only one to contain a full-alphabet in the a and b segments of the sequences.

The SPs 4-8 are described in Table 1.2. All the Sequences have 6 nucleotide long segments (like SP 1), which contain only **G** / **C**. The exception is SP 4, which is identical to SP 5, but one nucleotide in segments a and c is changed to **A** / **T** (highlighted in red in Table 1.2). Confirming the previous observations for SP 1, this change triggers SP 4 to phase separate and sediment in contrast to its counterpart SP 5. This suggests that the selection mechanism is able to separate single base changes and heavily favors sequence including all four nucleotides. For all the possible iterations of SPs with **G** / **C** -only segments, sedimentation was not observed (See Figure 1.11). Removing the **TT** spacers from SP 5, did also not lead to observable phase separation (see SP 7).

1.4 Conclusion

LLPS of oligonucleotides was experimentally demonstrated to lead to an evolutionary selection mechanism under feeding cycles. Specifically, by replacing the dilute phase with a consistent pool containing different oligonucleotide sequences, a dense phase grew that was enriched in specific sequences and depleted in others. The experiments with cyclic removal and feeding of an initial oligonucleotide pool using designed DNA sequences quantitatively validated our theoretical predictions, see Section 1.3.4.

A key characteristic of this selection mechanism is its strong sequence specificity, even in the presence of interacting sequences. Sequences showing robust interactions with others are enriched within the dense phase, while weakly interacting ones partition into the dilute phase and are expelled from the system. Remarkably, this mechanism proves effective even for very short oligonucleotides, as observed in the experiments involving 22 nt long sequences forming

Table 1.2: Sequences of the additional SPs. Sequence i is shaded in light grey and sequence ii is shaded in dark grey. For SP 4, the nucleotide 'mutation' is highlighted in red. SP 7 does not have spacer sequences.

Seq. pair	Sequence		a/a′	s	b/b′	s	c/ c′	
4	i	5′	GGAGGC	TT	GCGCGG	TT	GACCCG	3′
	ii	5′	GCCTCC	TT	CCGCGC	TT	CGGGTC	3′
5	i	5′	GGCGGC	TT	GCGCGG	TT	GGCCCG	3′
	ii	5′	GCCGCC	TT	CCGCGC	TT	CGGGCC	3′
6	i	5′	GCGGCC	TT	GGCGGG	TT	GGCGCG	3′
	ii	5′	GGCCGC	TT	CCCGCC	TT	CGCGCC	3′
7	i	5′	GGCGGC	-	GCGCGG	-	GGCCCG	3′
	ii	5′	GCCGCC	-	CCGCGC	-	CGGGCC	3′
8	i	5′	CCGCCC	TT	CGCGCC	TT	CCGGGC	3′
	ii	5′	GGGCGG	TT	GGCGCG	TT	GCCCGG	3′

cooperative base-pairing networks and undergoing phase separation at room temperature, see Section 1.3.3.

The robustness of this selection mechanism, particularly with short oligonucleotides, hints at its relevance for the molecular origin of life. This mechanism could have been behind the selection of other biomolecules that also undergo LLPS such as short-chained peptides or RNA sequences, effectively enriching them from the prebiotic pools where they were assembled. The cyclic removal of weakly interacting sequences can guide the selection of longer sequences which face dilution by the exponentially growing size of sequence space. Moreover, the dense phase could have provided enhanced stability against degrading chemical reactions such as catalytic cleavage [117] or hydrolysis due to the duplex formation [156]. In fact, there is a correlation between catalytic sequences and phase separation in functional ribozyme polymerases [115]. Ultimately, an enhanced selection propensity when combing self-replication with our selection mechanism relying on base-pairing interactions is to be expected.

1.5 Experimental Realization

1.5.1 Oligomer stocks

DNA oligonucleotides were purchased in dry form with HPLC purification from biomers.net and then adjusted to a stock concentration of approximately 200 µM with nuclease-free water (Ambion nuclease-free water from Invitrogen). All the strands were stored at -20°C. Before every experiment, the stock solutions of the strands were incubated at 95°C for 2 min to promote denaturation.

1.5.2 Sample preparation

For the phase separation with subsequent sedimentation, initial pools of 15 µL were prepared with 25 µM of each respective DNA strand, 10 mM Tris Buffer-HCl pH 7, 5X SYBR Green I (intercalating dye; excitation 450-490 nm, emission 510-530 nm), 125 mM NaCl and 10 mM

$MgCl_2$. The mixtures were heated to 95 °C for 2 min to ensure full de-hybridization of the strands. The temperature protocol that allows hybridization and consequent phase separation, was *i*. 95°C for 2 min, *ii*. 65°C for 10s, *iii*. cooling to 15°C (ramp rate: 6°C per minute), *iv*. 15°C for at least 3 h. Temperature protocols were performed in a standard thermocycler (BioRad CFX96 Real-Time System). The melting curves were measured in triplicates using the same reaction mixture and temperature profile as for the sedimentation experiments. Baseline correction using a reference measurement with only SYBR Green I. In the case of feeding cycle experiments, after sedimentation, 7.5µL of the supernatant, corresponding to 50% of the initial volume, was removed by carefully pipetting only at the center of the meniscus to avoid removing material from the sediment. Afterwards, 7.5 µL of the initial pool stock was added to the remaining bottom fraction. The aforementioned temperature protocol was then repeated, completing one feeding cycle.

1.5.3 Melting curve analysis

The analysis of the melting curves was done with a self-written Labview script and based on the baseline adjustment described in [82]. First the signal from the background fluorescence was subtracted from the fluorescence of the sample. Afterwards, the lower and higher baseline (linear) functions were determined and used for the baseline adjustment. These correspond to fully bound and fully unbound duplex states, respectively. The corrected data were then exported to "Igor Pro 6.37" and fitted with a sigmoidal function, where the midpoint fitting parameter corresponds to the T_m. The T_m of all the SPs was below the denaturing temperature used in the thermal protocol (95°C), so we ensure the mixture is homogeneous before triggering phase separation through cooling.

1.5.4 Sedimentation imaging and analysis

The imaging experiments were performed in a microfluidic chamber containing multiple wells, cut out of 500 µm Teflon foil and sandwiched between two sapphire plates (see Figure 1.12). The sample volume (about 15 mL per well) was loaded by using microloader pipette tips. The temperature of the chamber was controlled using three Peltier elements. To remove the waste heat from the Peltier elements, a Julabo 300F waterbath (JULABO GmbH) was used to cool the back of the chamber. The entire chamber is held in place by screwing a steel frame on top using a homogenous torque of 0.2 Nm. After loading, the wells were sealed with Parafilm to avoid evaporation. Monitoring of the edimentation was performed using a self-built fluorescence microscope composed of a 490 nm LED (M490L4, Thorlabs), a 2.5x Fluar objective (Zeiss) and the FITC/Cy5 H Dualband Filterset (AHF). Multiple wells could be imaged by moving the chamber perpendicularly to the light axis with two NEMA23 Stepper Motors and a CBeam Linear Actuator (Ooznest Limited). Images were taken using a StingrayF145B CCD camera (ALLIED Vision Technologies) connected via FireWire to a computer running a self-written Labview code operating camera, motors, LED's and Peltier elements.

In order to perform the sedimentation analysis, the time lapse stack of micrographs for each sample was loaded into a self-written Labview script (Fig. 1.13). Since SyBR Green I fluorescent intensity scales linearly with the amount of dsDNA in solution, it can be used for quantification [161]. The first micrograph was acquired before the sedimentation started. It was used as a reference image after the temperature had reached 15°C. All the remaining micrographs were divided by this one to obtain relative concentration c/c_0, where c_0

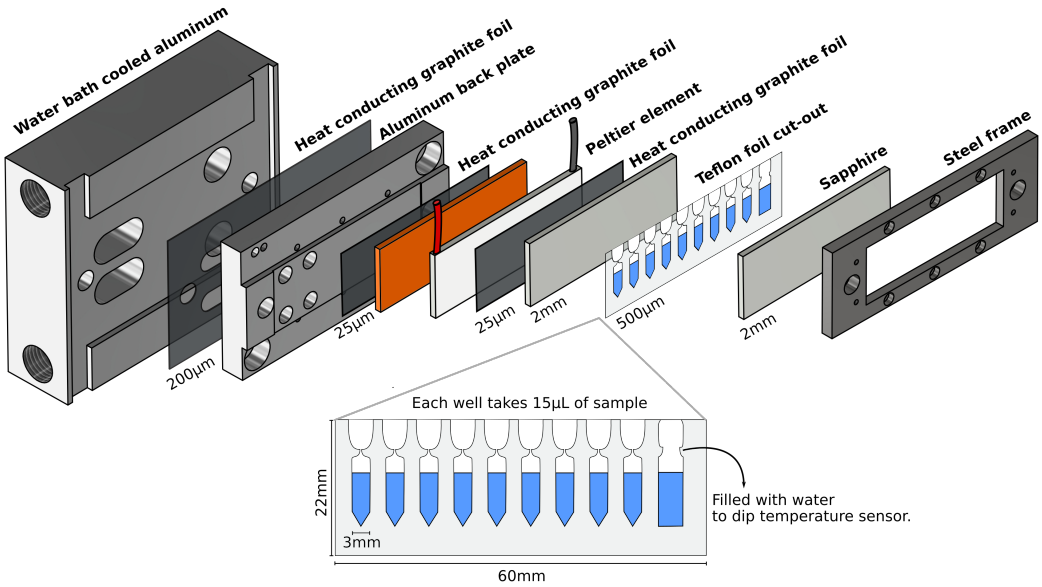


Figure 1.12: Schematic of sedimentation wells. The chamber is mounted from left to right. The temperature is controlled with the Peltier element, heat conducting graphite foil is used between all the parts to enhance heat conductivity. The water bath is used to dissipate the heat generated by the rear panel of the Peltier element back. The microfluidic Teflon cut-out is sandwiched between two sapphires which provide high heat conductivity when coupled to visible light transparency. The thickness is shown below each of the relevant parts. The Teflon cut-out, shown below as a zoom-in, has 9 each 3mm-wide wells and a temperature sensor is connected to the right-most well, which is wider. The shape of the remaining wells was designed to prevent the evaporation of the sample upon heating. The wide area above the narrow bottleneck is filled with parafilm, which seals the well and further prevents evaporation.

was the sequence concentration in the chamber immediately after flushing, i.e. the pool concentration.

When a sediment is present, the relative concentration through the sediment is obtained by measuring the concentration along a defined line perpendicular to the wall of the well, which we parametrize by x. The maximum of relative concentration occurs at the center of the sediment. Sediment height is determined along the x-direction from the center concentration $c_{\max}$ until the value has reached a $0.5\,c_{\max}$. The sediment height is then the distance where $c > 0.5\,c_{\max}$. The relative average sediment concentration is calculated by averaging over all points for which $c > 0.5\,c_{\max}$. Using the sediment height h_{sed} and average relative concentration $\bar{c}/c_0$, the total amount of sedimented material N_{sed} was calculated through Equation (1.5), where c_0 is the initial concentration, L is the length of the chamber and d is the depth of the chamber:

$$N_{\text{sed}} = \left(\frac{\bar{c}}{c_0}\right) c_0\, h_{\text{sed}}\, L\, d\,. \tag{1.5}$$

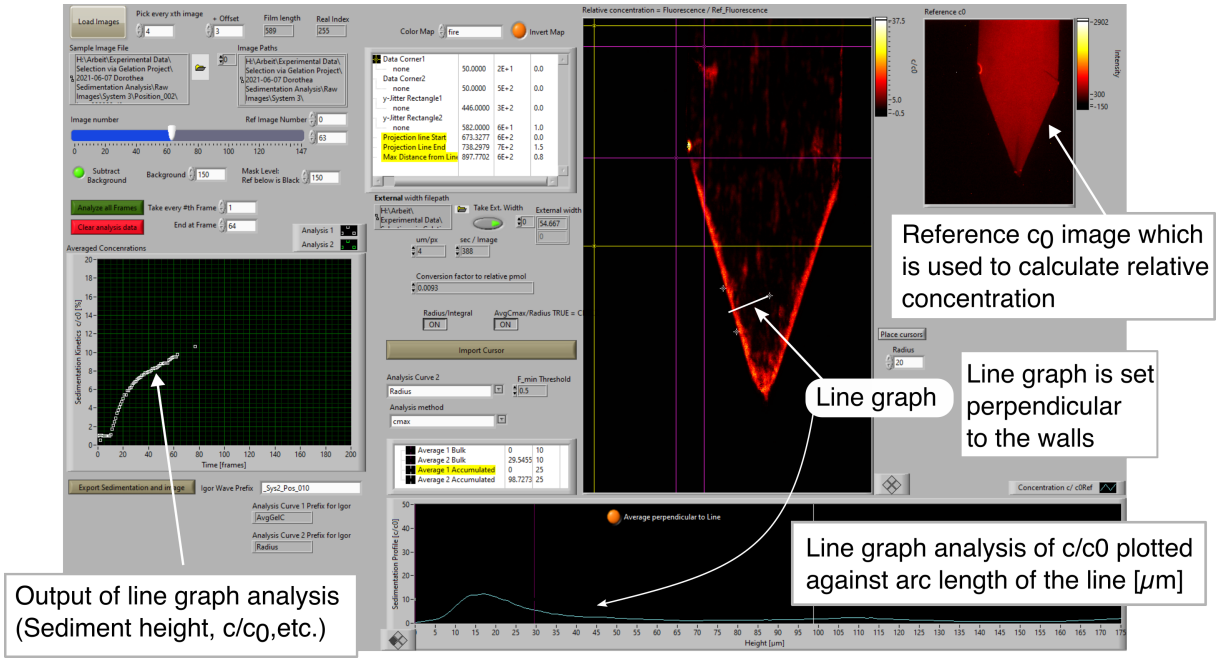


Figure 1.13: Screenshot of the self-written sedimentation analysis software in LabVIEW. A stack of images is analysed by defining a line perpendicular to the chamber wall (x-direction), along which the relative concentration c/c_0 is plotted for each image.

1.5.5 HPLC-UV absorbance

Ion-pairing reverse phase HPLC experiments were carried out on an Agilent 1260 Infinity II LC System coupled to Agilent 1260 Infinity II DAD WR detector and an Agilent 6230B ESI TOF Mass spectrometer. A C18 capillary column (AdvanceBio Oligonucleotide 4.6x150 mm with particle size 2.7 µm, Agilent) was used to perform reverse phase liquid chromatography. The temperature of the autosampler was set to 4°C. The mobile phases consisted of two eluents. Eluent A was HPLC water (Sigma-Aldrich), 200 mM HFIP[5] (Carl Roth GmbH), 8 mM TEA[6] (Carl Roth GmbH). Eluent B was a 50:50 (v/v) mixture of water and methanol (HPLC grade, Sigma Aldrich, Germany), 200 mM HFIP, 8 mM TEA. The injection volume for each measurement was 100 µL.

The samples were eluted with a gradient of 1% B to 58.6% B over the course of 45 min with a flowrate of 1 mL/min. Prior to the gradient, the column was flushed with 1% B for 5 min.Retention times were analyzed via a UV Diode Array Detector (Agilent 1260 Infinity II Diode Array Detector WR G7115A) at 260 nm with a bandwidth of 4 nm. Samples were diluted for HPLC loading in the following manner: 7.5 µL of sample, 105 µL nuclease free water and 75 µL of a 5 M urea solution. They were heated to 95°C for 2 min afterwards to ensure de-hybridization of the strands and dissolution of any sediment. Then, 105 µL of the diluted samples were transferred into N9 glass vials (MachereyNagel GmbH) and stored at 4°C in the auto-sampler of the HPLC system until injection.

[5]short for 1,1,1,3,3,3, -Hexafluoro-2-propanol
[6]short for triethylamine

Appendix

1.A Exemplary NUPACK design code

```
temperature[C] = 37.0
material = dna

# domains
domain a1 = S2
domain a2 = W2
domain a3 = S2
domain b1 = S2
domain b2 = W2
domain b3 = S2
domain c1 = S2
domain c2 = W2
domain c3 = S2
domain s = T2

# strands
strand s1 = a1 a2 a3 s b1 b2 b3 s c1 c2 c3
strand s2 = a3* a2* a1* s b3* b2* b1* s c3* c2* c1*

# complexes
complex ca = s1 s2
complex cb = s1 s2
complex cc = s1 s2

# target structures
ca.structure = D6(U16 +) U16
cb.structure = U8 D6( U8 + U8) U8
cc.structure = U16 D6(+ U16)

# tubes
tube tub = ca cb cc
tub.ca.conc[M] = 1e-6
tub.cb.conc[M] = 1e-6
tub.cc.conc[M] = 1e-6

prevent = AAAA,CCCC,GGGG,UUUU,AA,GGG
stop[%] = 10
```

Figure 1.14: Exemplary Nupack design code. For DNA at a temperature of 37°C, two strands were designed by segmentation into smaller domains (a, b and c). In the # domains section, the a, b and c are split into even smaller domains of similar nucleotide content (**S** being **G** or **C**, and **W** being **A** or **T**). In this example code, **TT** spacers were defined via "s = T2". In "# strands", the sequences of the system are defined, reading from 5′ to 3′. The * denotes complementarity. Hybridization constraints are shown in "# target structures", where the 3 possible binding interactions from "# complexes" between strand 1 and strand 2 are defined further. "D6U16" for example denotes a 6 base pair region followed by a 16 unpaired region. Preventing specific sequence-patterns, such as **AAAA** or **GGGG**, above a certain relative amount, can be useful to avoid for example **G**-quadruplex structures or unwanted stacking.

1.B Multi-component phase separation subject to cyclic material ex- changes

Here, concentration changes in a mixture of volume V that is composed of M different oligonucleotide sequences subjected to periodic exchange of material with a pool with fixed composition $\boldsymbol{c}_{\text{pool}}$ are described. This theory was developed within the collaboration with Dr. Giacomo Bartolucci and Prof. Christoph Weber from the University of Augsburg. The concentration of each sequences are stored in an M-dimensional vector $\bar{\boldsymbol{c}}(t)$. Starting from the initial state $\bar{\boldsymbol{c}}(t_0)$, and performing N exchange cycles, where each cycle is labeled with $n = 1, \ldots, N$ and composed of the two following steps:

Phase separation step: The homogeneous mixture of concentration $\bar{\boldsymbol{c}}(t_n)$ phase-separates into two coexisting phases. The concentrations of the dense and dilute phase are denoted as $\boldsymbol{c}^{\text{I}}(t_n)$ and $\boldsymbol{c}^{\text{II}}(t_n)$, respectively. The volume of the dense phase is $V^{\text{I}}(t_n)$ and thus the dense phase occupies the volume $(V - V^{\text{I}}(t_n))$. Mass and particle numbers are conserved during the phase separation step:

$$\bar{\boldsymbol{c}}(t_n) = \left[\frac{V^{\text{I}}(t_n)}{V} \boldsymbol{c}^{\text{I}}(t_n) + \frac{V - V^{\text{I}}(t_n)}{V} \boldsymbol{c}^{\text{II}}(t_n) \right] . \tag{1.6}$$

Partial dense phase removal step: A constant fraction of the volume of the dense phase, $\alpha(V - V(t_{n-1}))$, with the relative fraction $0 < \alpha < 1$, is replaced by the same volume taken from the pool, $\boldsymbol{c}_{\text{pool}}$. The average composition thus changes according to:

$$\bar{\boldsymbol{c}}(t_{n+1}) = \left[\frac{V^{\text{I}}(t_n)}{V} \boldsymbol{c}^{\text{I}}(t_n) + \frac{V - V^{\text{I}}(t_n)}{V} \Big(\alpha \boldsymbol{c}_{\text{pool}} + (1 - \alpha) \boldsymbol{c}^{\text{II}}(t_n) \Big) \right] . \tag{1.7}$$

For the general case where the initial average concentration is not equal to the average concentration of the pool, $\bar{\boldsymbol{c}}(t_0) \neq \boldsymbol{c}_{\text{pool}}$ (full lines in Fig. 1.8 **a** and **b**), the phase compositions $\boldsymbol{c}^{\text{I}}(t_n)$ and $\boldsymbol{c}^{\text{II}}(t_n)$, and the phase volumes $V^{\text{I}}(t_n)$ at each cycle time t_n are determined. The details of this are in the Supplementary material of the Bartolucci et. al publication [8]. Note that during the selection kinetics, the average concentration $\bar{\boldsymbol{c}}(t_n)$ approaches the tie line passing through the pool concentration vector $\boldsymbol{c}_{\text{pool}}$.

For the special case where the initial average concentration is equal to the average concentration of the pool, $\bar{\boldsymbol{c}}(t_0) = \boldsymbol{c}_{\text{pool}}$ (dashed lines Fig. 1.8 **a** and **b**), an analytic solution can be obtained, even for an arbitrary number of different components M. Only in this case, $\boldsymbol{c}^{\text{I}}$ and $\boldsymbol{c}^{\text{II}}$, remain constant in time since the average concentration moves along the tie line defined by the pool. The iteration rule Equation (1.7) simplifies to:

$$\bar{\boldsymbol{c}}(t_{n+1}) = \Big[\lambda(t_n) \boldsymbol{c}^{\text{I}} + (1 - \lambda(t_n)) \Big(\alpha c_{\text{pool}} + (1 - \alpha) \boldsymbol{c}^{\text{II}} \Big) \Big] , \tag{1.8}$$

where $\lambda(t_n) = V^{\text{I}}(t_n)/V$ is the relative volume of the dense phase. Using particle conservation in Equation (1.6) for the pool concentration $\boldsymbol{c}_{\text{pool}}$, remaining within the special case $\boldsymbol{c}_{\text{pool}} = \bar{\boldsymbol{c}}(t_0)$, the interaction rule becomes:

$$\bar{\boldsymbol{c}}(t_{n+1}) = \Big[\alpha\lambda_0 + (1-\alpha\lambda_0)\lambda(t_n)\Big]\,\boldsymbol{c}^{\mathrm{I}} + \Big[1-\alpha\lambda_0 - (1-\alpha\lambda_0)\lambda(t_n)\Big]\,\boldsymbol{c}^{\mathrm{II}}\,, \tag{1.9}$$

where $\lambda_0 = \lambda(t_0)$.

Using particle conservation at time step t_{n+1},

$$\bar{\boldsymbol{c}}(t_{n+1}) = \lambda(t_{n+1})\boldsymbol{c}^{\mathrm{I}} + (1-\lambda(t_{n+1}))\,\boldsymbol{c}^{\mathrm{II}}\,, \tag{1.10}$$

the term in the first bracket of Equation (1.9), can be identified as $\lambda(t_{n+1})$, the size of the dense phase at t_{n+1}, and obtain a recursion relation:

$$\lambda(t_{n+1}) = a\,\lambda(t_n) + b\,, \tag{1.11}$$

where $a = 1-\alpha\lambda_0$ and $b = \alpha\lambda_0$. For large times, the system reaches a stationary state

$$\lambda(t_\infty) = \frac{b}{1-a} = 1\,. \tag{1.12}$$

This can be used to rewrite the recursion in terms of $\delta\lambda(t_n)$ with $\delta\lambda(t_n) = \lambda(t_\infty) - \lambda(t_n) = 1-\lambda(t_n)$. As a result, $\delta\lambda(t_{n+1}) = a\,\delta\lambda(t_n)$. Its solution reads $\delta\lambda(t_n) = \delta\lambda_0\,a^n$ that can be written as

$$\lambda(t_n) = 1 - (1-\lambda_0)\,(1-\alpha\lambda_0)^n\,. \tag{1.13}$$

This solution completely determines the evolution of the mean volume fraction:

$$\bar{\boldsymbol{c}}(t_n) = \Big[1-(1-\lambda_0)\,(1-\alpha\lambda_0)^n\Big]\boldsymbol{c}^{\mathrm{I}} + (1-\lambda_0)\,(1-\alpha\lambda_0)^n\,\boldsymbol{c}^{\mathrm{II}}\,. \tag{1.14}$$

The characteristic number of iterations required to converge to the stationary state $\lambda(t_\infty) = 1$ (and, consequently, $\bar{\boldsymbol{c}}_\infty = \boldsymbol{c}^{\mathrm{I}}$) is

$$n_{\mathrm{c}} = -\frac{1}{\log a} = -\frac{1}{\log[\alpha(1-\lambda_0)]}\,. \tag{1.15}$$

2 Replication elongates short DNA, reduces sequence bias and develops trimer structure

Summary

Prebiotic oligonucleotide pools, formed through non-enzymatic nucleotide condensation, have a highly biased nucleobase composition. Various factors contribute to this phenomenon, such as varying environment availability of nucleotides and their precursors and different incorporation kinetics. This bias is more pronounced than that observed in 'evolved' and functional oligonucleotides like ribozymes. The persistence and propagation of this initial bias during replication, and its potential impact on the replication process, were so far unexplored topics. To address these questions, the evolution of 12-mer short biased DNA pools was investigated using an enzymatic model system. Analysis using next-generation sequencing (NGS) at various time points post-replication revealed the disappearance of the initial nucleotide bias in the elongated pool. Conversely, the nucleotide composition at each position in the elongated sequences remained biased, dependent on the initial pool's composition. Essentially, while the overall pool homogenized in nucleotide composition, certain positions held memory of the initial pool's nucleotide bias. Moreover, sequences that replicated more rapidly contained highly periodic dimer and trimer motifs. These periodic motifs likely contributed to the increase of intra- and inter-strand binding sites. Since binding through base-pairing is a prerequisite for replication, these motifs conferred a replicative advantage to the fast replicator sequences. The shift in nucleotide composition and the emergence of structure through templated replication provide insights into how biased prebiotic pools undergo molecular evolution. This process may result in the preferential selection of specific sequences over others, ultimately leading to the development of complex functional nucleic acids. [1]

[1]This chapter was published by Serrão and Dänekamp et al. [19] in NAR and is here adapted and reprinted in part with permission from NAR.
Full article attached in List of Publications.

2.1 Motivation

Early pools of nucleic acids likely featured distinct ratios of each nucleobase. Prebiotic condensation of mononucleotides leads to pools of short oligomers with a sequence bias, namely with one nucleobase incorporated more into the product strands [27, 32, 33, 42]. This bias can be due to the compounded contribution of different factors. On the one hand, there may be an imbalanced abundance in the environment caused by varying rates of nucleotide formation and degradation with different geochemical conditions [**?**, 22, 70, 71, 97]. On the other hand, even when the environment has equimolar concentrations of all reacting nucleotides, the rate of the condensation reactions themselves may also vary for different nucleotides [27, 32, 83] further exacerbating the incorporation bias.

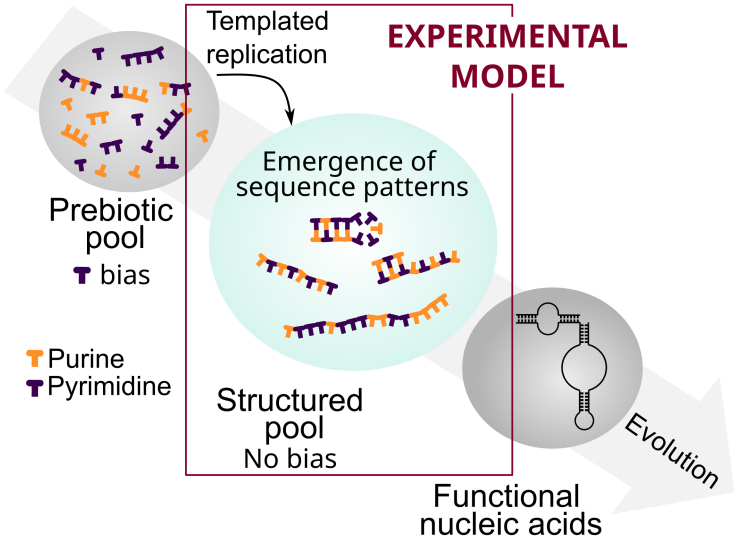


Figure 2.1: Templated replication as a dual mechanism for increase of compositional diversity and selection of sequence patterns. Prebiotic pools, resulting from non-enzymatic condensation reactions, would be highly biased pools of short oligomers. These, through templated replication could lead to a pool of longer oligomers with a homogenous nucleobase composition. Certain sequence and structural motifs may be enriched, as replication also acts as a selection mechanism. These pools can then be a starting point for the molecular evolution of functional nucleic acids.

Functional nucleic acids, whether from natural or artificial origin, were found to occupy only a subsection of the sequence space [63, 127]. Regardless of the displayed function, the nucleobase composition of functional RNA was found to be imbalanced towards purines [63]. These sequence biases however are milder than the ones obtained in nucleotide condensation studies [30]. Additionally, on the structural level, also referred to as the phenotype, pools of natural short RNA have a very heterogeneous distribution of secondary structure motifs [127]. In order for functional, and therefore complex, nucleic acids to evolve from such a pool, specific rare secondary structures should be represented. It was found that nucleotide composition plays a larger role than sequence length to bias a pool on the structural level.

While imbalances on the nucleotide composition and therefore structure level are important for molecular evolution to take place, the selection mechanisms that led to them are not yet fully understood. The highly biased short oligomer pools, resulting from non-enzymatic

mononucleotide condensation, would undergo selection and elongation, yielding pools with sequence and structural patterns [28, 55].

Templated replication is a potential mechanism through which both the compositional diversity and sequence length can increase to facilitate the exploration of sequence space while replicating sequence information [30]. While the overall nucleotide composition is expected to diversify, several studies have shown that templated replication can act as a selection mechanism in itself, enriching specific sequence motifs [41, 50, 67, 131]. More experimental investigations are needed to grasp the influence of the initial biases of the pool on the sequence level.

2.2 Scientific approach

The hypotheses behind these experiments were: *i)* Does a biased prebiotic pool of nucleic acids have its bias transferred or transformed by undergoing templated replication? and *ii)* Are the sequence patterns (i.e. enriched motifs) retrieved from the replicated pool dependent the bias of the initial pool?

To answer these questions, initial random pools of DNA sequences, with 12-mer length and a binary composition (**A**/**T** or **G**/**C**) were replicated with *Bst*, a strand displacing polymerase. Each of the four possible biased initial pools (**A**-rich, **T**-rich, **C**-rich, **C**-rich) were isothermally replicated and the resulting pools were sequenced at two different time points. An early time point pool was studied to learn about "fast replicators", and a late time point pool to understand the sequence distribution in the “left-behind” sequences. The initial pools used are shown in Table 2.1.

Table 2.1: Initial pools of random 12-mer DNA with the requested bias upon ordering. DNA sequences were ordered from biomers.net. **S** signifies **G** or **C** and **W** signifies **A** or **T**. DNA synthesis can have a significant shift from the intended bias and the actual nucleotide content of the binary initial pools was only assessed post-sequencing, as is shown and discussed in Section 2.5.1.

Pool		Sequence		Requested Bias
G-rich (G_0)	5′	**SSSSSSSSSSSS**	3′	(67% **G**:33% **C**)
C-rich (C_0)	5′	**SSSSSSSSSSSS**	3′	(33% **G**:67% **C**)
A-rich (A_0)	5′	**WWWWWWWWWWWW**	3′	(67% **A**:33% **T**)
T-rich (T_0)	5′	**WWWWWWWWWWWW**	3′	(33% **A**:67% **T**)

The templated nature of the replication mechanism is relevant because only when copying is present can the information of the template sequence be propagated to the nascent replicated strand. Due to the complementarity of Watson-Crick base-pairing, necessary for templation, a strong bias to one nucleobase leads to the complementary base being correspondingly more incorporated in the nascent strand. This in turn is expected to homogenize the average pool nucleotide fraction, as schematized in Figure 2.2. While this happens as an overall pool average, localized patterns or motifs may still arise, when replication exerts selective pressure.

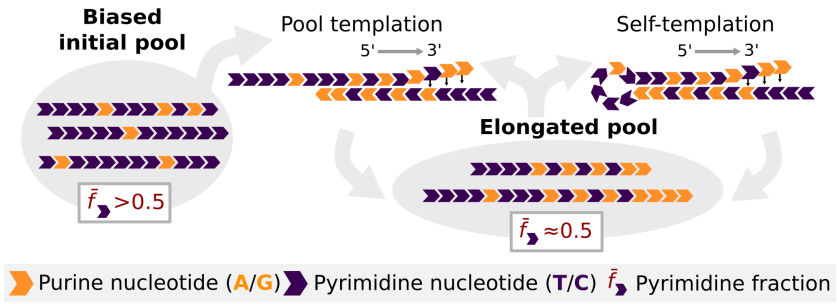


Figure 2.2: Polymerization starts from a binary initial pool (AT or GC) with a bias $\bar{f}$ of either purine (orange) or pyrimidine (purple) nucleotides. Distinct sequences from the pool base-pair to form short duplexes and are enzymatically extended (5′-3′) complementarily to the template ("pool templation"). Longer sequences may also self-template through hairpin-like secondary structures ("self-templation"). The biased pyrimidine fraction $\bar{f}_T$ or $\bar{f}_C$ in the initial pool is countered by complementary elongation.

This experimental approach implies a series of decisions and assumptions that facilitate researching the hypothesis in a laboratory setting and isolating the contribution of each of the parameters to strand selection. The topics of **DNA vs. RNA** and **reaction tube vs. rocky pore** have been discussed in Section 1.2 and a similar argumentation can be applied to the assumptions in this project.

- **Artificial strand design vs. pool generated in prebiotic conditions**

  The initial pools were synthesized by an external company to have a specific bias of each of the four nucleobases. Prebiotic pools synthesized through non-enzymatic condensation mostly yield sequences that are very short (less than 10 nt in length) with longer sequences being several orders of magnitude lower in concentration [27, 31, 72, 142]. For this reason, in order to isolate the contribution of the nucleotide composition bias from length and abundance, the initial pools are designed pools of 12-mer with binary composition. This length regime also allows for stable duplexes to form between sequences that could act as a substrate for templated polymerization. Additionally, with a designed pool the nucleotide bias can be controlled and kept on the same order of magnitude between different pools.

- **Binary pool vs. full-alphabet pool**

  Binary pools (with either **A**/**T** or **G**/**C**, henceforth named **AT** and **GC** pools) were chosen instead of full-alphabet pool (with **A**, **T**, **C** and **G**). The reasons for this choice are twofold: i) the sequencing analysis is far more challenging with a 4 nt pool, and ii) binary pools isolate the contributions of **A**:**T** and **G**:**C** base-pairs to the replication dynamics which facilitates the description of a more complex full-alphabet pool. It was assumed that the conclusions resulting from studying a binary pool would be applied to understanding replication of a full-alphabet pool. The validity of this assumption has been experimentally tested, with a proof-of-concept experiment for a biased full-alphabet pool and the analysis challenges are further discussed in Section 2.3.7.

- **Strand-displacement vs. environment-based strand-separation**

  In contemporary biology, strand separation and elongation occur in tandem [11, 122]. The displacement of any pre-hybridized strands is performed enzymatically. However, strand displacement can also be triggered by the hybridization of other sequences in the pool [135]. This non-enzymatic strand displacement has recently been described for a prebiotic RNA replication system [158]. When compared to other prebiotic mechanisms proposed for strand separation, such as pH [55, 80], heat and salt fluctuations [56], strand displacement has the advantage that it can also occur isothermally and with a constant chemical environment [135]. It thereby erases the need for a specific set of cycle conditions that are potentially more difficult to satisfy and isolates the impact of replication on sequence structure from other environment variables. Replication of sequence information was the main driver of evolution and to be able to generalize the conclusions obtained to several proposed prebiotic environments.

- **Enzymatic vs. non-enzymatic replication**

  Replication of the initial pools was done with *Bst*, which performs high-fidelity templated polymerization in a isothermal manner through a strand-displacement mechanism. *Bst* binds to double stranded regions and elongates the strand in the 5′ to 3′ direction with high-fidelity [4, 21, 102], displacing downstream bound strands and denaturing secondary structures present in the primer or template - it does not 'slip' [140]. See Appendix 2.A for a more detailed description of the specific *Bst* strain used. Note that all the sequences in the initial pool are 12-mer and that primer and template is a notation that solely depends on the direction of elongation, i.e. *Bst* adds nucleotides to the 3′ end of the primer, Figure 2.2. A single strand can therefore go through several replication rounds, even in isothermal conditions – first through pool templation and later, when a certain length threshold is crossed, through self-templation, Figure 2.2. This is therefore a robust model system for prebiotic primer extension through strand-displacement starting from a diverse pool. An enzyme was chosen, as an alternative to a more plausible non-enzymatic replication, due to its much faster kinetics. As the objective of this study is to describe molecular evolution in a pool undergoing several cycles of replication, evolutionary timescales can only be modeled experimentally with an enzyme.

2.3 Results and Discussion

The starting DNA pools A_0, T_0, G_0 and C_0, Table 2.1 were isothermally amplified with *Bst*. While the bias requested upon manufacturing was 2/3 vs 1/3, the actual initial pools were revealed to initially contain 60% **A** (A_0), 75% **T** (T_0), 70% **G** (G_0), 69% **C** (C_0) by NGS, see Sections 2.3.2 and 2.5.5. The possible sequence space was $2^{12} = 4096$, but sequences were not represented equally due to the bias. The incubation temperatures were 35°C for **AT** pools and 65°C for **GC** pools. Different incubation temperatures were necessary to observe replication due to the lower stability of the **A**:**T** base-pair in comparison to the **G**:**C** one. The initial mixtures contained 10µM of the initial pool, 1.4 mM of each dNTP, 10 mM $MgSO_4$, 320 U/mL of *Bst*, in 1X of the standard buffer provided from the manufacturer, Section 2.5.2.

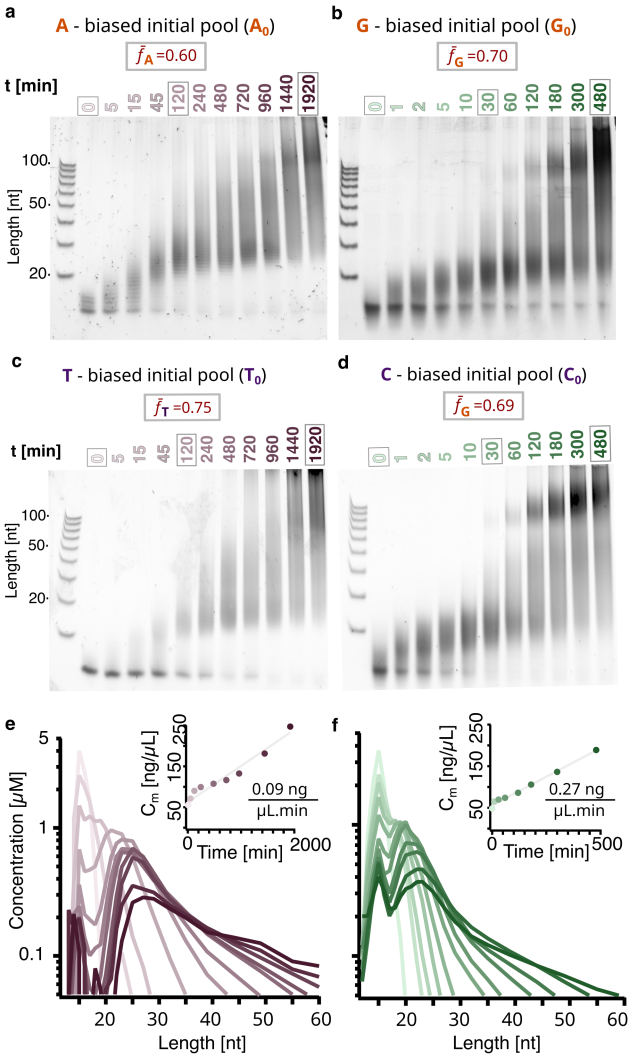


Figure 2.3: Initial pools polymerize producing products longer than 100-mer. PAGE analysis shows the length distribution of **a**, **A**-biased (A_0), **b**, **G**-biased (G_0), **c**, **T**-biased (T_0) and **d**, **C**-biased (C_0) pools over time. Aliquots of the initial mixtures were incubated at either 45°C (**AT** pools) or 65°C (**GC** pools) and removed at the different time points indicated. For the initial mixtures, the total DNA concentration was 10 μM, with 1.4 mM of each dNTP, 10 mM $MgSO_4$, 320 U/mL of *Bst*, in 1X of the standard buffer provided from the manufacturer, Section 2.5.2 for more methodological details. The fraction of nucleotide **T** (f_T) for **AT** pools and of nucleotide **C** (f_C) for **GC** reported in panels **a**-**d** was obtained through NGS sequencing, Section 2.5.5. The molar concentration of sequences was quantified through PAGE smear analysis detailed in Appendix 2.C and plotted over sequence length for each time point corresponding to individual lanes. A_0 corresponds to pink (**c**) and G_0 to green (**d**), with hue increasing over time. The total DNA mass concentration grows linearly with time (inlets) and was fitted in grey.

2.3.1 Kinetics of elongation

To understand the kinetics of nucleotide incorporation with *Bst* for each of the initial pool at the described conditions, the evolution of sequence lengths over time was analyzed through PAGE, Figure 2.3 **a**-**d**. Different time points were analyzed for **AT** and **GC** pools to account for the different kinetics observed with each of the data sets. The polymerization was stopped at 32 h for **AT** and 8 h for **GC** when the length distribution reached a state with an abundance of sequences well beyond 100 nt in the PAGE gel. Both the **A**-biased (A_0) and the **G**-biased (G_0) pools displayed replication to sequences longer than 100-mer within the first 2 h. For the case of the **AT** pools, most of the short initial sequences (<20 nt) were depleted after 2 h, whereas for **GC** pools, these remained detectable even for later time points.

The concentration profiles over strand length were obtained via ladder-calibrated SYBR Gold fluorescence intensity in PAGE gels and depicted for all time points of samples A_0 and G_0 in Figure 2.3 **e** and **f** respectively. The methodology for PAGE smear quantification through SYBR Gold is detailed in Appendix 2.C. The contribution of nucleotide composition to SYBR Gold intensity was ruled out by performing a screen with sequences of different compositions at known concentrations, Appendix 2.B. For both A_0 and G_0 pools, the molar concentration at later time points forms a double peaked length distribution with a long tail which continues to lengths longer than 300 nt. The first peak, around 12 nt, could be explained by the sequences of the initial pool that were not recruited for replication. The second peak, between 20 and 30 nt, could be due to fully hybridized duplexes that have a melting temperature above the incubation temperature [111]. These duplexes would be too stable at the working temperature to denature spontaneously, and therefore inaccessible to additional rounds of templation, whether by another sequence or itself (pool- vs. self-templation, respectively). *Bst* would also not denature a fully-bound duplex, as it needs a double stranded region with a single stranded region ahead to initiate polymerization, Appendix 2.A.

Through the elongation mechanism, free dNTPs are added to the already existing sequences. This means that while the total number of sequences does not change, the total DNA mass increases linearly with time as more nucleotides are incorporated, depending on the enzyme's average rate of incorporation, Figure 2.3 **e** and **f**, inlets. In the case of **AT**, this increase is 0.09 ng/µL min whereas for **GC** it is 0.27 ng/µL min. The difference in kinetics observed (about three times slower for **AT** experiments) can be explained by the compounded temperature dependent efficiency of *Bst* (10% at 35°C vs. 100% at 65°C) and the nucleotide-dependent differences in the rate of incorporation [101, 106].

2.3.2 Sequencing yield and read length distribution

To assess the sequence content of our product strands we used NGS. For each of the four initial pools three time point samples were sequenced (indicated in Figure 2.3 **a**-**d** by the grey outlines). The raw sequencing data obtained was pre-processed with a trimming step through which low quality reads or read segments were removed and followed by a Regular Expression (RegEx) filtering step. This second step ensured that only reads with binary sequence and with the proper adapter sequence at the 3′ end were selected. The NGS data processing is detailed in Section 2.5.5.

The three time points represent the initial pool, an early time point pool, from to retrieve "fast replicators", and a late time point pool to understand the sequence distribution in the "left-behind" pool. The maximum sequence length captured is 112 nt, corresponding to the

maximum read length of 120 nt minus the **CT** -tail of at least 4 nt and the **AGAT** adapter which are cut in the RegEx filtering step of the pre-processing, Section 2.5.5. This effectively isolates the left-behind sequences from the fast replicators in the sequenced late time point, as the fast replicators become much larger than 112 nt and are not captured in the late time point. In the case of A_0 and T_0, the samples sequenced were at 0 h, 2 h and 32 h, whereas in the case of G_0 and C_0, at 0 h, 0.5 h and 8 h. The resulting data sets allowed to characterize how the bias in the initial pools affects the pool evolution on short and long time scales.

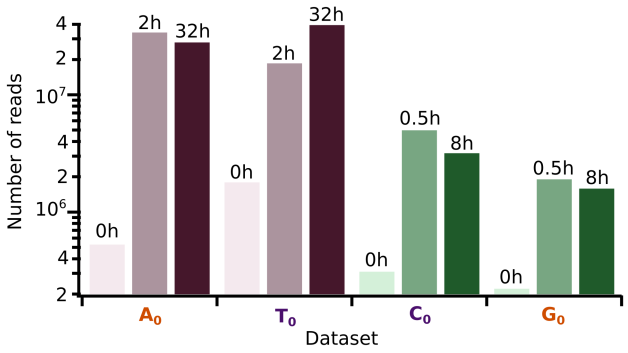


Figure 2.4: Total read counts obtained with NGS. The data was de-multiplexed and trimmed before this analysis (pre-processing), as described in Section 2.5.5. The total read number varies by more than one order of magnitude between samples, while it was expected to be constant, as the total amount of DNA strands does not change with polymerization (10 µM total DNA). Sequencing has both sequence biases and length biases. This can be seen, on the one hand, by the **GC** data sets systematically having less read counts than **AT** ones, which likely is due to differences in sequencing yield. On the other hand, initial pools consisting of 12-mer sequences also have less reads than their corresponding later time points, revealing that sequencing of longer strands is favoured.

The total read counts for all the samples are shown in Figure 2.4. Since the amount of strands should stay the same (10 µM total DNA strands), the read counts should be constant across all data sets, which is not the case. Firstly, the read counts observed were about two orders of magnitude lower for the initial pool in comparison to the early and late time points. Since the initial pool consists exclusively of short 12 nt sequences, this lower sequencing yield could be due to ineffective library preparation for short DNA fragments, both for the adapter ligation step or the PCR step.

Additionally, **GC** data sets have less read counts than **AT** data sets. As this is already the case for the initial 12-mer pools, likely is due to differences in sequencing yield possibly related to challenges in the amplification step of the library preparation necessary for NGS. For instance, both very high and low **GC** contents (both the case in the work as the systems were binary) have been shown to be a challenge to NGS due to difficulties in the PCR step [15].

The length distribution of the reads obtained after pre-processing is plotted in Figure 2.5. For all of the samples, the initial pool is almost entirely composed of 12-mer strands. Shortly after the onset of elongation, in the early time points (2 h for **AT** data sets and 0.5 h for **GC** data sets), the amount of 12-mer strands decreases, as they have been recruited for replication. The early time point curves peak at lengths between 20 and 30 nt, similar to the length distributions obtained via PAGE smear quantification, Figure 2.3. The end time points

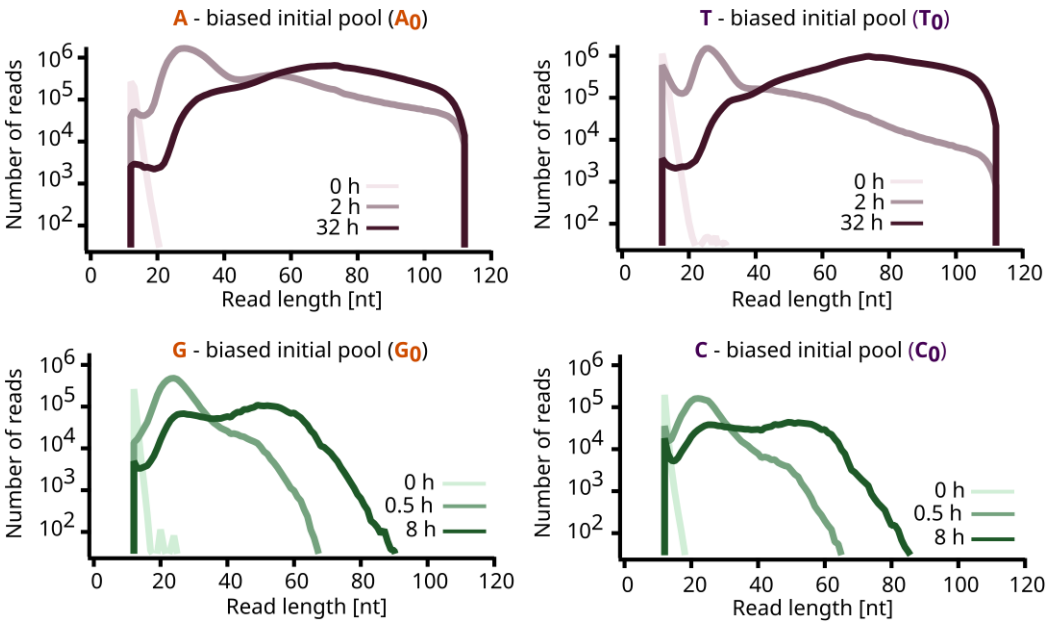


Figure 2.5: Length distribution of the NGS reads obtained after pre-processing. The initial pools are composed almost entirely of 12-mer strands. Those strands are depleted as sequences get recruited for replication. Between the early and the late time point, strands in the length regime 12-40 nt are depleted, supporting the idea that one strand can go through several rounds of replication. Strands longer than the maximum read length of 112 nt are only partially sequenced and were discarded during the RegEx step, see Section 2.5.5.

reveal a depletion both of the initial 12-mer and of the 20-30 nt peak, indicating that both 12-mer and longer strands get recruited in the later stages of replication. This supports the idea that one strand can go through several rounds of replication.

The maximum read length obtained with Illumina NGS was 112 nt which is the reason for the sharp wall in the **AT** data sets. There are likely longer products only partially sequenced, which are discarded by RegEx filtering. The longer strands are visible in the PAGE images for the **AT** data sets, Figure 2.3. The most striking difference between the length distributions obtained for the two methods (PAGE and NGS) is the higher abundance of long (up to 112 nt) products compared to short ($\approx$ 12-mer) in the case of NGS.

2.3.3 Global nucleotide fraction

The nucleotide fraction for all sequences in each of the datasets was calculated and is shown in Figure 2.6. Even though the initial pools are biased, replicated pools exhibited a shift towards a distribution centered around 0.5 for both **AT** and **GC** pools ($\bar{f}_{T/C} = 0.5$), indicative of an uniform average pool nucleotide fraction. Notably, there was no significant difference between the early and late time points, suggesting a rapid reduction in bias.

This phenomenon occurs since in pools where a majority of sequences are biased toward one nucleotide, the sequences are likely to encounter similarly biased templates. During template-directed polymerization, complementary nucleotides are incorporated into the templates, causing an inversion of bias in the newly forming strand segments, as shown in Figure 2.2.

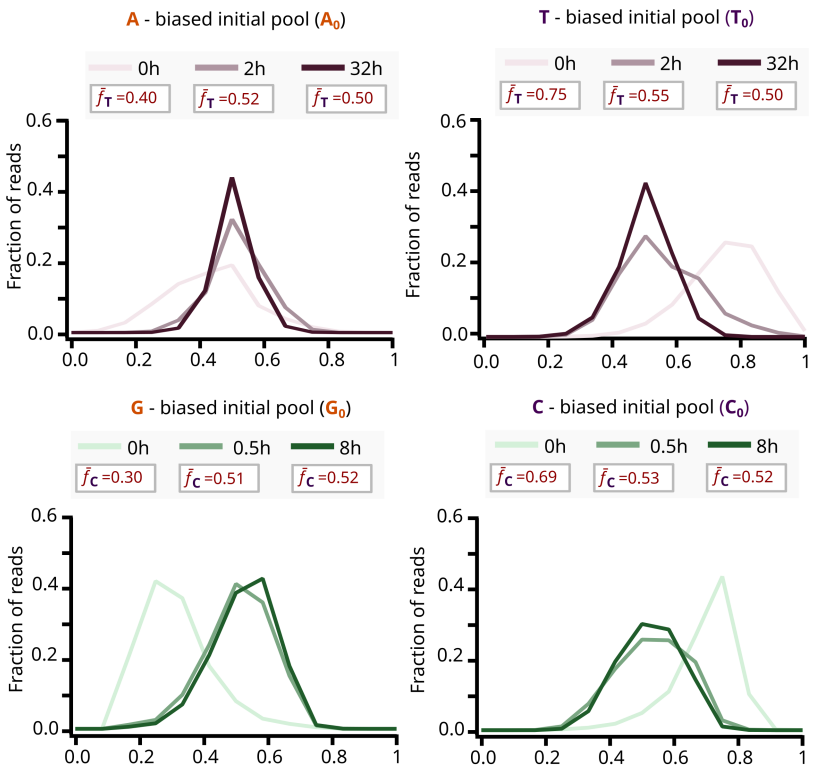


Figure 2.6: Distribution of nucleotide fraction among the obtained reads after pre-processing. The initially biased pools shift towards a distribution centered around 0.5, corresponding to a homogeneous average pool nucleotide fraction. Early and late time point distributions are very similar, indicating a rapid reduction of bias.

2.3.4 Positional nucleotide fraction

The overall bias of the pool was rapidly homogenized with polymerization with *Bst*, Section 2.3.3. However, to investigate whether specific positions or regions of the elongated sequences are enriched or depleted in certain nucleotides, the positional nucleotide fraction must be computed. To assess this, the average fraction $f_N(i)$ of nucleotide **N** at each position i was plotted. As polymerization leads to all possible integer lengths from the initial 12-mer sequences (to the maximum range detectable by sequencing of 112 nt), $f_N(i)$ was computed for each possible sequence length. The graphs were then stacked so that the positions align across lengths to facilitate visualization of position dependent.

The analysis of the $f_T(i)$ for **AT** data sets (A_0 and T_0) is depicted in Figure 2.7. The position is plotted in the direction 5′ to 3′ end, the direction in which *Bst* elongates the sequences. This way, for every sequence length, the probability of finding the nucleotide **T** at each position can be read. As the system is binary, a divergent color scale was used to facilitate interpretation: a positional enrichment in the **T** ($f_T(i) > 0.5$) corresponds to a purple shade, whereas an enrichment in **A** corresponds to an orange shade ($f_T(i) < 0.5$). A position that has a homogenous nucleotide fraction ($f_T(i) = 0.5$), i.e. no positional bias, is plotted in white. Evidently, $f_T(i) = 1 - f_A(i)$ for a binary system.

The initial pools of A_0 and T_0 show an overall homogeneous bias across position for the 12-mer length. The lack of structure across positions for the initial pools supports the assumption that structure in early and late time points develops through replication. In the case of the

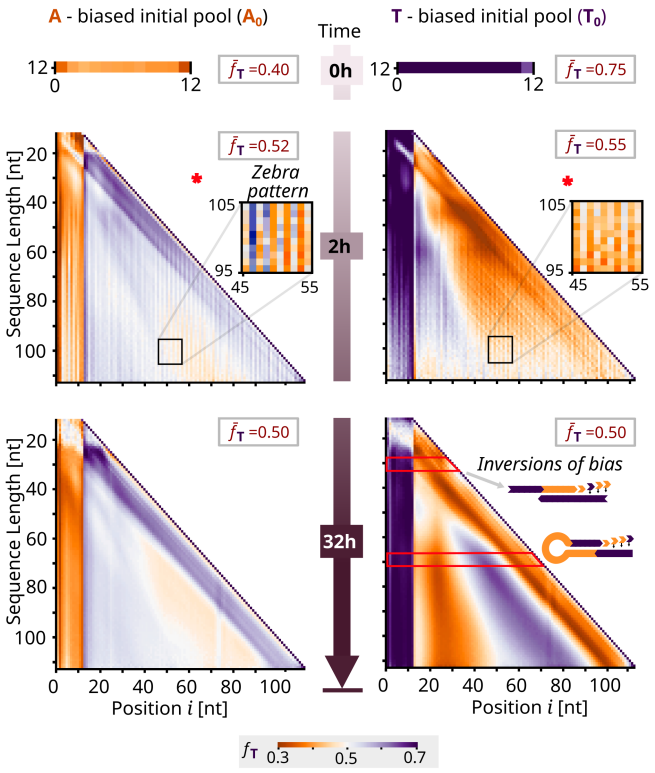


Figure 2.7: Effects of initial bias and elongation on sequence composition for AT (A_0 and T_0 experiments). Evolution of nucleotide fraction f_T across sequence lengths and positions in sequences for the initial pool, an early time point and a late time point (0 h, 2 h, 32 h). **A** -rich regions ($f_T \ll 0.5$) represented in orange and **T** -rich regions ($f_T \gg 0.5$) in purple. The initially biased average pool nucleotide fraction $\bar{f}_T$ is countered as the pool undergoes polymerization, homogenizing to 0.5 at later time points. The first 12 nucleotides at the 5′ terminus retain the initial sequence bias for all graphs, due to the directionality of the polymerization mechanism (5′-3′). In addition, an inverse bias at the 3′ is explained explained by pool templation from the biased pool. For the 2 h time point, horizontally alternating "zebra" patterns of f_T are visible, illustrated by the insets with increased contrast. At 32 h, gradients of alternating nucleotide fraction suggest self-complementarity, possibly a consequence of self-templation.

replicated sequences, through the 2 h and 32 h time points, while the pool-averaged bias was homogenized, in-strand positional biases were amplified. Due to the 5′-3′ direction of *Bst*, nucleotides do not get added to the 5′ end of sequences. This means, whichever bias the initial pool has, the initial 12-mer segment at the 5′ terminus will be preserved over the complete reaction period. Additionally, since the nucleotides added are mostly complementary, the nascent segment will be inversely biased at the 3′ terminus. This is observed in A_0 and T_0 as 12-mer columns with inverse biases stretching inwards from both termini of the sequences, both for the 2 h and 32 h time points.

Fast replicators, corresponding to sequences observed at the early time point, feature patterned structure. They display a zebra pattern, visible through the vertical stripes indicating alternating average nucleotide fractions, 2 h segments in Figure 2.7, insets. After 32 h of polymerization, the zebra patterns in the fraction of **T** have been replaced by smooth gradients.

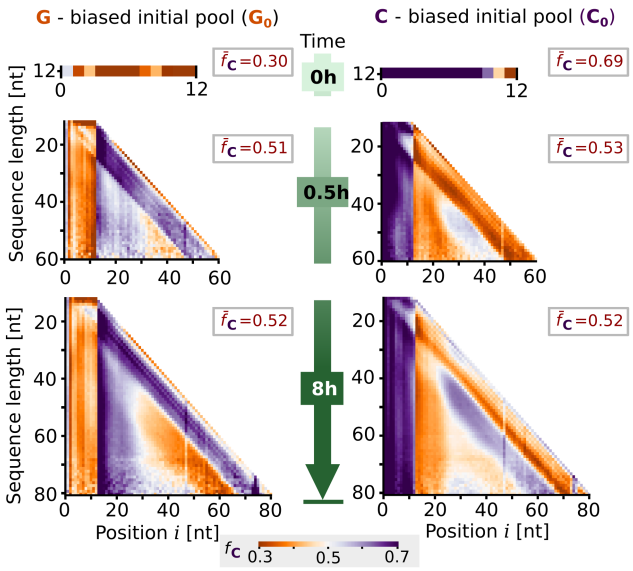


Figure 2.8: Effects of initial bias and elongation on sequence composition for GC (G_0 and C_0 experiments). The nucleotide fraction f_C for the 0 h, 0.5 h and 8 h time points was decomposed by length and position, which leads to graphs similar to those for **AT**. Again, the initially biased average pool nucleotide fraction $\bar{f}_C$ is homogenized with time and the first 12 nt retain the initial bias while the following segment is inversely biased due to pool templation. However, no zebra patterns are visible in the early 0.5 h time points.

A reason for this may be that the fast replicators have elongated even more and are no longer captured by sequencing analysis. This supports that such zebra patterns are a feature of fast replicators only. These gradients are antisymmetric around the center, corresponding to alternating inversions of bias. This indicates self-complementarity, suggesting self-templation through the formation of hairpins as a mechanism of elongation. Self-templation is favoured over pool templation when possible since it is kinetically more likely to find a complementary region within the proximity of the same molecule than within another molecule of the pool, Figure 2.2. Furthermore, the emergence of self-complementarity at the late time point suggests its possible adverse effect on replication, causing certain sequences to be left-behind, as these sequences form stable, fully bound duplexes.

Similarly to the **AT** pools, the positional nucleotide fraction was computed for the **GC** pools, G_0 and C_0 and are shown in Figure 2.8. In the case of the polymerized pools, the sequences obtained were overall shorter than in the case of the **AT** data sets, even for the later time points. This may be due to a combination of the different polymerization dynamics and a lower sequencing efficiency for **GC** samples which yields fewer and lower quality reads for a similar initial concentration, see Section 2.4. For this reason, the **GC** graphs are noisier and have shorter maximum length.

The initial pool for the case of G_0 and C_0 is not entirely homogeneous, and the specific positions that differ from the mean bias seem to retain that bias in the first 12 nt reinforcing the directionality as a reason for the preservation of bias on the 5′ terminus nucleotides. For the earlier time point, at 0.5 h incubation time, the alternating vertical stripes that indicated zebra patterns in the AT graphs are not present, indicating that the fast replicators for the

case of GC pools do not contain high zebraness. The inversion of bias both on the 5′ to 3′ terminus and in the intermediate region is evident for both the 0.5 h and 8 h time points as in AT. These can be explained with the pool and self-templation mechanisms in addition to the directionality of polymerization, Figure 2.2.

2.3.5 Intra-sequence periodicity and patterns

2.3.5.1 Conditional nucleotide probability

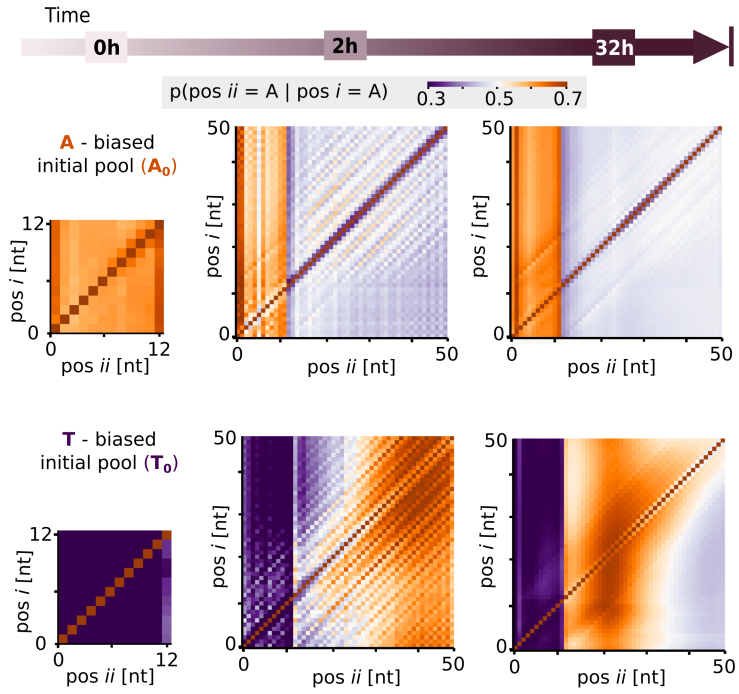


Figure 2.9: Conditional probability of finding base A at position ii given A at position i: $p(pos\ ii = \mathbf{A} \mid pos\ i = \mathbf{A})$ for the A_0 and T_0 data sets, top and bottom row respectively. The initial pools have a mostly homogeneous bias across positions and did not reveal any particular periodic structure. For the early time points (2 h), diagonal structure indicating periodicity was visible for all the samples. This periodicity, however, was not present in the late time points. These had only vertical regions of bias, corresponding to structures that are present in the average of the pool, as for instance for the initial 12-mer.

The $\bar{f}_T$ plots shown in Section 2.3.4, Figure 2.7 and 2.8 do not allow conclusions about in-sequence periodicity since the nucleotide fraction is averaged over all sequences of the same length. For this reason the conditional probability of finding a certain nucleotide N on position ii, knowing that position i has an N $p(ii =$ N $|i =$ N$)$ was investigated. In other words, what is the average nucleotide fraction in position ii for the sequences that have N in position i? This way the dependency of positional biases within a single sequence can be observed.

For the both AT and GC pools sequences of a single length were selected from the pool (50 nt), and the probability of finding one of the two possible bases at a specific position

depending on the base found at another position in a square graph, Figure 2.9 and 2.10 respectively. In these plots the diagonal always displays a probability of 100% as $i = ii$ for that case. Two main patterns propose themselves.

On the one hand the vertical patterns, that is, positional biases that are present independently of the nucleotide fraction on other positions of the sequence ($p(i =$ **N**$)$ does not affect $p(ii =$ **N**$)$) are pool-average patterns. For instance, the first 12 nt of a replicated pool (early and late time points) are on average preserving the bias of the initial pool, regardless of the remaining sequence. This can be seen for both **AT** and **GC** pools. Averaging all positions ii in the conditional probability plots yields the positional nucleotide fraction plots presented in Section 2.3.4, for this case, the line for sequence length 50 nt. Other vertical patterns, that not the first 12 nt can also be traced to the plots in Section Section 2.3.4. On the other hand, there are also diagonal patterns which are presented the early time points but not for initial pools or late time points for both **AT** and **GC**. Diagonal lines parallel to the main diagonal correspond to periodic structures present in individual sequences. This supports the idea that fast replicators are highly periodic. This was already suggested for **AT** pools through the zebraness observed in the pool averaged positional nucleotide fraction plots. However, here this 3 nt periodicity is revealed for fast replicators of both **AT** and **GC** which puts forward a characteristic of rapid replication.

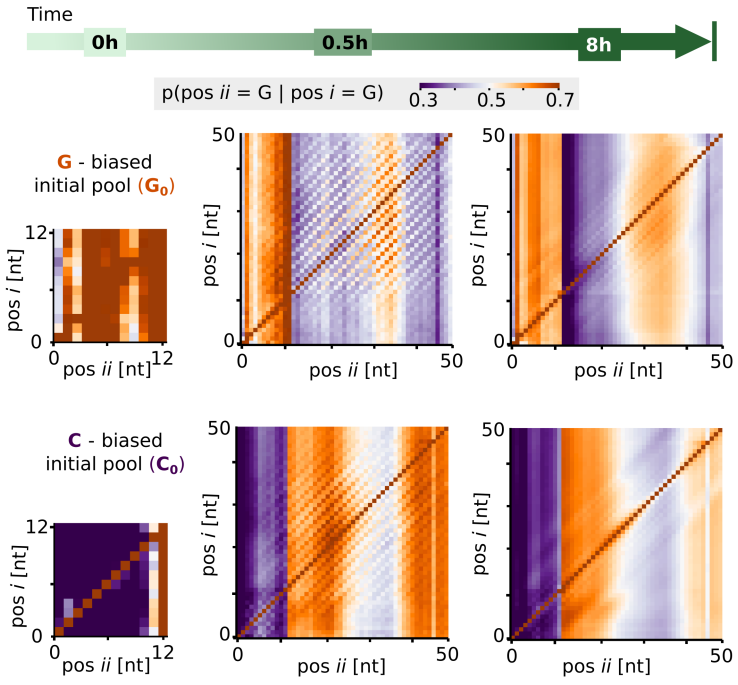


Figure 2.10: Conditional probability of finding base G at position ii given G at position i: $p(pos\ ii =$ **G** $|\ pos\ i =$ **G**$)$ for the G_0 and C_0 data sets, top and bottom row respectively. Similarly to the **AT** data sets shown in Figure 2.9, only the early time point (0.5 h) showed diagonal stripes which indicates periodicity.

2.3.5.2 Fourier transform

The presence of sequence periodicity, with onsets or shifts varying across the pool, visible as diagonal lines parallel to the main diagonal in conditional probability plots in Figures 2.9 and 2.10 suggests Fourier analysis as a method of quantification. To obtain the dominant period of the patterns, a discrete Fourier transform was applied to every row of the conditional probability correlation matrices (corresponding to a fixed position i) and averaged across all rows and sequences to obtain the Fourier amplitude for **AT** and **GC** data sets shown in Figure 2.11.

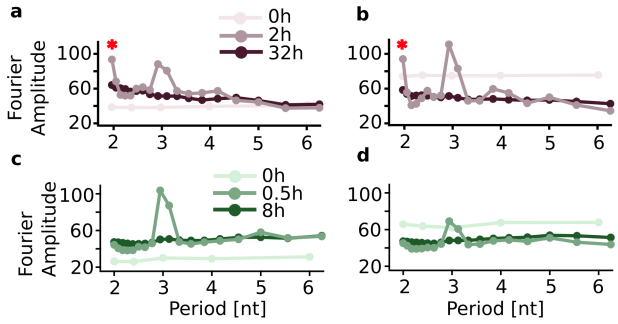


Figure 2.11: Periodicity obtained through the amplitude of the Fourier modes of a discrete Fourier transform. Fourier transform was performed on the position-dependent conditional probability of **A** (for A_0 and T_0, shown in **a** and **b** respectively), or **G** (for G_0 and C_0, shown in **c** and **d**, respectively), discussed in Section 2.3.5.1. For the **AT** data sets, the fast replicator sequences from the 2 h early time point display patterns with period 2 nt, matching the zebra patterns of the nucleotide fraction graphs, as well as period 3 nt. In contrast, for the **GC** data points the 2 nt periodicity is not present.

For the case of A_0 and T_0 the early time point graphs spike at period 2 nt and 3 nt above the baseline Fourier amplitude of 50 which random sequences would display (the baseline equals the average pool nucleotide fraction in percent). Initial and late time point pools do not present any dominant period as was expected from the conditional probability plots from Figures 2.9 and 2.10. The 2 nt periodicity zebra patterns (**ATATATATA** is an example of a zebra like sequence) were suggested from the positional nucleotide fraction plots in Figure 2.7. The enriched 3 nt periodicity, which stems from sequences or sequence segments that have a 3-mer motif repeated (e.g. **AATAATAAT** or **TATTATTAT**), was only retrieved from the Fourier transform analysis.

Similarly for G_0 and C_0 samples, only the early time points display sequences with increased periodicity, reinforcing its role for fast replication. However, unlike the **AT** samples, these lack 2 nt periodicity while still displaying an increased periodicity of 3 nt. This indicates that 3 nt periodicity is a feature of fast replicators independent of the initial pool type whereas

2.3.5.3 Zebraness

The fast replicators in the case of **AT** pools showed increased 2-mer periodicity (zebra patterns), in comparison with the initial and late time point pools. Zebraness has been found to be a useful measure of pool structure in a theoretical study simulating elongation of random RNA strands, permitting an investigation of evolutionary dynamics in sequence space [50].

The zebraness $\zeta(S)$ of a strand S of length L_S is defined as the number of alternating "zebra" motifs (**XY** or **YX** , being **Y** and **X** any two nucleobases) within the sequence divided by the total number of 2-mer motifs, given by $L_S - 1$. Correspondingly, 2-mer bulky motifs are defined as homodimers (**XX** or **YY**). A random sequence is expected to have a zebraness of 0.5, a fully alternating strand one of 1 and a homogenous strand one of 0.

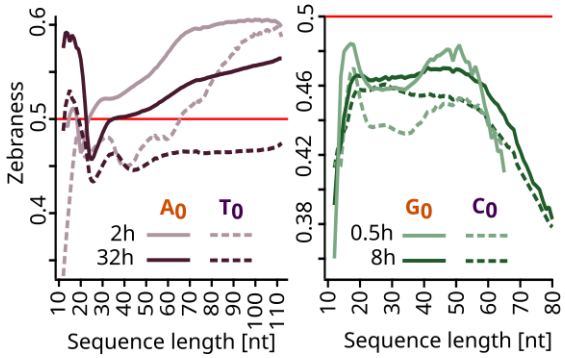


Figure 2.12: Zebraness by sequence length, defined as the fraction of alternating 2-mer motifs (XY , YX). In the case of **AT** data sets (left panel) the 2 h early time point sequences possessed a higher average zebraness than their 32 h late time point counterparts. Additionally, the zebraness was higher for longer sequences, suggesting that 2 nt periodicity is present in fast **AT** replicators. In constrast, **GC** samples (right panel) had a generally lower zebraness, consistently below 0.5. Furthermore, the zebraness decreased for longer strands, indicating that the bulky non-alternating 2-mer motifs (**XX** , **YY**) are favoured for the fast **GC** replicators.

While 50 nt long sequences in the **AT** early time point pools have an increased 2-mer periodicity on average, in order to obtain more mechanistic insights about how sequences become fast replicators, it would be important to understand the evolution of zebraness across sequence lengths for all the pools. Zebraness-per-lenght is defined Z_L as the average of all $\zeta(S_L)$ for all the strands S_L of the same length L and plot it across sequence length in Figure 2.12 for both **AT** and **GC** early and late time point pools. The zebraness that a random pool would have (0.50) is indicated in red. The findings reveal that in **AT** sequences, zebraness of fast replicators is higher than 0.5 and consistently exceeds that of left-behind sequences. For both early and late time points, zebraness increases with length. In contrast, for **GC** sequences, zebraness consistently falls below 0.5 and decreases with sequence length. Thus zebraness appears to confer a replicative advantage to **AT** sequences, while not benefiting **GC** sequences.

2.3.5.4 4-mer motifs

The analysis of periodicity through conditional probability, Fourier transform and 2-mer motifs, as well as indirectly through the concept of zebraness-per-length, provides insight about differences in sequences with replicative advantages between **AT** and **GC** pools, Section 2.3.6. However, certain small motifs may be enriched in the overall pool without manifesting as periodic within sequences. For this reason the distribution of the 4-mer motifs for all the data sets was computed. A sliding window algorithm was used to count the occurrences of every possible 4-mer (out of the possible 16 different 4-mer motifs) in each of the sequences of either pool, and the normalized by the total number of motifs counted. Randomly generated

sequences would yield almost perfectly evenly distributed motifs, as deviations quickly disappear through statistical narrowing. This is shown in Figure 2.13, where the motif-axis is ordered from **A**-rich motifs to **T**-rich motifs, such that reverse complementary motifs are in symmetric positions on the axis.

In the 2h timepoint for **AT** and the 0.5h timepoint for **GC**, only sequences with a length of at least 40 nt are analyzed whereas the 32 h (**AT**) and 8 h (**GC**) pool is split into two subsets, of which the 12-40 nt and the 70-120 nt (**AT**) or 40-120nt (**GC**) are displayed in Figure 2.13. The symmetry in the 4-mer motif graphs reveals the enrichment in reverse complementary motifs for both pools after replication. The increase in overall pool complementarity leads to the convergence of the pool average nucleotide fraction to about 0.5 as discussed in Section 2.3.3. The main difference between **AT** and **GC** is that the favoured motifs are more zebra-like – with 2 or especially 3 zebra 2-mer submotifs (e.g. **AATA** and **ATAT**, respectively) – for **AT** pools and more bulky – with a majority of bulky 2-mer submotifs (e.g. **CCCC** or **CCGG**) – for **GC** pools.

2.3.5.5 Self-complementarity

Self-complementarity within single sequences has been suggested through two of the analysis performed: 1) the anti-symmetric inversions of nucleotide fraction bias near the 3′ and 5′ end of sequences, when averaged across the whole pool, indicating that the terminal segments fold back and sequences may self-template, Section 2.3.4 and 2) the symmetric distribution of 4-mer motifs, for which reverse complementary motifs have similar relative abundance in the overall pool, Section 2.3.5.4.

In order to investigate self-complementarity, the longest possible complementary region was computed for each sequence. This is done by comparing the sequence starting at the 3′-end to its sequence from the 5′-end and looking for the largest possible complementary overlap. In case more than one region has the same maximum self-complementary length, the region with the maximum area between the two complementary parts of a strand is kept. The length of the longest region found was then averaged for all the sequences of the same length and is shown in Figure 2.14. To establish a reference point, a random pool was generated with a nucleotide fraction of $\bar{f}_{T/C,pool} = 0.50$. This reference provides a baseline for the maximum length of self-complementary regions in the absence of pool or sequence level patterns.

It is essential to clarify that this study exclusively examined hypothetical perfect hairpins, and no exploration into potential secondary structure folds was conducted. The objective was to comprehend the relationship between the length of self-complementary regions and the overall length of the elongated sequence. Future research exploring other more complex secondary structures, distinct from hairpins and potentially independent of sequence length, could provide valuable insights into the mechanism of replication. The enrichment of complex secondary structures in a pool has been shown to be a metric on whether molecular evolution can start within that pool which would be a next step in the time line of nucleic acid emergence [127]. While secondary structure remains unknown plotting the average length of self-complementarity against the sequence length allowed to discover at which strand lengths self-complementarity becomes a prominent phenomenon. For **AT** the fast replicators possess the longest average longest self-complementary regions, Figure 2.14. Left-behind sequences also have longer self-complementary regions than the random sample, but only approximate the region length of left-behind sequences for very long sequence length. Both

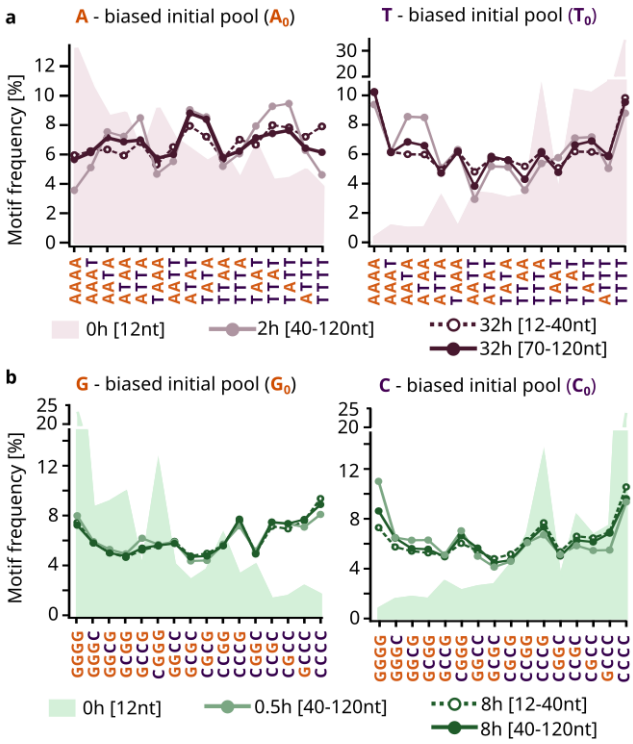


Figure 2.13: 4-motif distributions of different time points for A_0 and T_0 (a) and G_0 and C_0 (b) data sets. The motif distribution of the initial pool, in a solid color in the background, reveals a highly skewed distribution, due to the initial bias of the pool. In the case of **AT** data sets (**a**) for the 2 h time point, the motifs were plotted for the sequences that are above 40 nt in length in order to characterize the fast replicators. These showed an enrichment in zebra motifs. For the 32 h pool, two different length frames were analysed: 12-40 nt and 70-120 nt. The longer sequences had a similar motif distribution to that of the fast replicators, whereas the shorter sequences had a flatter distribution. The symmetry of the distribution indicates that reverse complementary motifs are enriched. For the **GC** pools (**b**), in the case of the 0.5 h time point, the motifs were plotted for the sequences that are above 40 nt in length in order to characterize the fast replicators. Similarly to **AT** pools, for the later time pool, 8 h, two different length frames were analysed: 12-40 nt and 40-120 nt. In this case, all of the replicated pools had a homogeneous distribution of motifs with a slight preference towards bulky motifs.

the fast replicators and the left-behind sequences deviate from a random sample at around 40 nt.

The graphs for **GC** strands do not exhibit longer self-complementary areas than a randomly generated sample with homogeneous average pool nucleotide fraction for either the fast replicators or the left-behind sequences indicating that self-complementarity does not confer a replicative advantage to **GC** sequences. This may be due to the stronger binding energy of the **G** : **C** base-pair in comparison with the **A** : **T**.

2.3.6 Mechanistic insights

For elongation to occur, two sequences need to form overlap duplexes or a sequence needs to self-template, Figure 2.2. However, if the resulting duplex is excessively stable after replication,

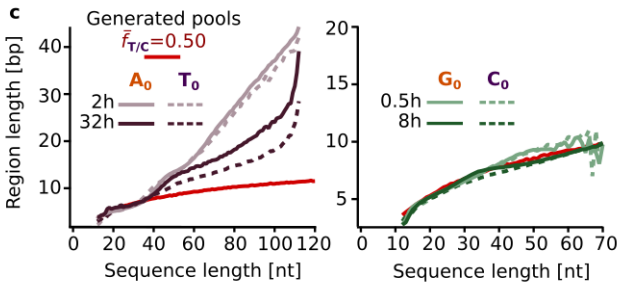


Figure 2.14: Analysis of length of self-complementary regions, decomposed by sequence length, for AT and GC experiments. AT pools (left panel) displayed higher self-complementarity compared to a randomly generated homogeneous pool, particularly for sequences longer than 40 nt. No significant deviations from a randomly generated pool were present in the GC pools (right panel).

it hinders strand separation and further replication, effectively leading to the sequences being left behind. To understand sequence evolution, both early and late time points were analyzed, aiming to distinguish characteristics of fast replicators from left-behind sequences. Late time points reveal anti-symmetric bias inversion regions (Figure 2.3.4) indicative of fully self-bound sequences, which are too stable to replicate and therefore remain in the left-behind pool.

To gain insights into how self-complementarity evolves during replication, we analyzed the longest potentially self-complementary region in each sequence, Figure 2.14. Notably, the longer AT self-complementary regions of fast replicators coincide with an increased 2 nt periodicity. To understand this characteristic, we investigated zebraness, and found that in AT sequences, zebraness of fast replicators is higher than 0.5 and consistently exceeds that of left-behind sequences. In contrast, for GC sequences, zebraness consistently falls below 0.5. Thus zebraness appears to confer a replicative advantage to AT sequences, while not benefiting GC sequences. The difference between the AT and GC fast replicators can be explained by the intrinsic differences in the stacking energies ΔG of zebra – averaged from the motifs XY and YX – and bulky XX / YY motifs which have been determined in [118] (literature values ΔG^{SH}, all in mol^{-1}):

$$\begin{aligned}
\Delta G^{zebra}_{AT} &= (\Delta G^{SH}_{AT/TA} + \Delta G^{SH}_{TA/AT})/2 &&= -0.73 \\
\Delta G^{bulky}_{AT} &= \Delta G^{SH}_{AA/TT} &&= -1.00 \\
\Delta G^{zebra}_{GC} &= (\Delta G^{SH}_{GC/CG} + \Delta G^{SH}_{CG/GC})/2 &&= -2.20 \\
\Delta G^{bulky}_{GC} &= \Delta G^{SH}_{GG/CC} &&= -1.84
\end{aligned}$$

Thus for AT, bulky motifs are more stabilizing than zebra motifs, whereas for GC the opposite is true, with the stacking energy difference $\Delta G^{bulky} - \Delta G^{zebra}$ equaling 0.27 mol^{-1} for AT versus -0.36 mol^{-1} for GC. Stacking energy of neighbouring nucleotide pairs is the main contributor for duplex stability [151], explaining its strong effect on sequence evolution. The sequences rich in the most destabilizing motif type replicate the fastest into very long strands. This prevents them from being stuck in very stable secondary structures and renders them more accessible for several rounds of priming. Additionally, long zebra regions are fully self-complementary, allowing a single strand to have many possible transient fold-back conformations and undergo several rounds of self-templation, which could be a replication

mode of AT fast replicators. Due to the elevated stability of the G : C base-pair in comparison to the A : T base-pair, in addition to stacking, for GC this mode of replication might lead to overly stable self-folded conformations impeding their status as fast replicators.

Table 2.2: Possible motifs of length 2 to 5-mer for AT pools, categorized into pool-templating or self-templating. The periodicity of a n-mer motif increases the complementarity of the pool and provides sequences (or sequence subsets) with a replicative advantage. Sequences rich in motifs that are self-complementary to themselves may undergo self-templation whereas the periodicity of other short motifs may lead to pool templating networks. Pool-templating motifs are grouped with their corresponding reverse complement. In red are homopolymeric motifs, which if repeated would not correspond to a defined periodicity.

2-mer motifs	3-mer motifs	4-mer motifs	5-mer motifs
Pool-templating	Pool-templating	Pool-templating	Pool-templating
AA / TT	AAA / TTT	AAAA / TTTT	AAAAA / TTTTT
Self-templating	AAT / ATT	AAAT / ATTT	AAAAT / ATTTT
AT	ATA / TAT	ATAA / TTAT	AAATA / TATTT
TA	TTA / TAA	AATA / TATT	AAATT / AATTT
		TAAA / TTTA	ATAAA / TTTAT
		TAAT / ATTA	ATAAT / ATTAT
		Self-templating	AATAA / TTATT
		AATT	AATAT / ATATT
		ATAT	TAAAA / TTTTA
		TATA	TAAAT / ATTTA
		TTAA	TAATA / TATTA
			TAATT / AATTA
			TAATT / AATTA
			ATATA / TATAT
			TTAAA / TTTAA
			TATAA / TTATA

The enhanced 3 nt periodicity is a distinct characteristic of fast replicators both in AT and GC experiments (Figure 2.11). The base of the mechanism may however be the same: an increase of periodicity in the pool leads to more possibilities of finding a binding partner. Unlike zebra sequences, which are reverse complementary to themselves (i.e. XY is its own reverse complement), 3 nt periodic sequences cannot as easily self-template through hairpin formation unless they are composed of (at least) two regions with repeating reverse complementary 3-mer motifs. However, their periodic regions offer an increased amount of potential binding sites for reverse complementary periodic regions of other sequences, allowing for the formation of duplex regions for elongation to start. For this reason 3 nt periodicity balances the formation of duplexes for elongation with the avoidance of overly stable self-binding, enabling fast replication. Two subpopulations of sequences with reverse complementary periodic 3-mer motifs may form efficient primer-template pairs that rapidly bind, elongate, and separate again, effectively cooperating to achieve fast replication.

In fact, any pool periodicity would be enriched as motif repetition leads to an increase of binding partners either intra- or inter-sequence (leading to fast replication through pool- or

self-templation, respectively). However, only periodic motifs with even length can be their own reverse complement (such as **XY** for 2-mer and **XYXY** for 4-mer). This means, sequences rich in uneven periodic motifs need to form cooperative replication networks in order to replicate (unless for the special case where one sequence is rich in both one motive and its reverse complement). The advantage of 3 nt periodicity over longer 4 nt or 5 nt periodicities is not only the higher amount of potential binding sites, but more importantly the small sequence space associated with 3-mers. This results in only four "3 nt periodic partner" subpopulations (containing periodic motifs **AAT** / **ATT** , **ATA** / **TAT** , **TAA** / **TTA** , **AAA** / **TTT**) instead of the combination of six pool-templating plus four self-templating subpopulations for 4 nt periodicities or a total of 16 "5 nt periodic partner" subpopulations in the pool. The example for 3- 4- and 5-mer motifs is given in Table 2.2.

2.3.7 4-letter alphabet: a special case

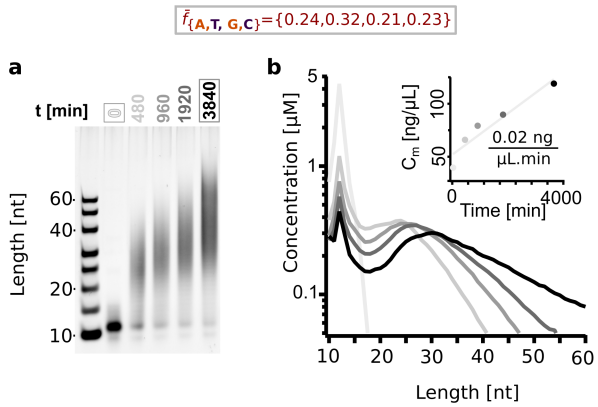


Figure 2.15: Kinetics of polymerization for ATGC pools. **a** Similar as for the binary data sets, initial pools were polymerized for different time points at 45°C. **b**, The molar concentration of sequences was quantified (Section 2.5.3 and Appendix 2.C) and plotted over sequence length for each time point corresponding to individual lanes, with hue increasing over time. The inlet shows the total DNA mass concentration, which was fitted linearly in grey.

One of the initial assumption made was that studying binary pools would provide insights about more realistic 4-letter pools, through the simplification of the methodology and isolation of the different contributions, see Section 2.2 for further explanation about the assumptions of the approach and Appendix 2.D for a breakdown of the difficulties of sequencing 4-letter pools. To challenge this assumption a supplementary experiment that analyzes the replication of a 4 nt data set ($ATGC_0$) at two time points (0 and 64 h) was perfmoerd. An initial 4-letter alphabet pool was replicated, applying the *Bst* experiment protocol to a random pool with 12 nt long single stranded sequences with a **T** -bias ($ATGC_0$). The 10µM initial pool was incubated at 45°C for up to 64 h (3840min), as these conditions led to the most extensive elongation. The $ATGC_0$ samples were incubated for longer than the binary samples to account for the slower kinetics of nucleotide incorporation resulting from the much larger sequence space, which decreases the probability of primer to template attachment creating a suitable double-stranded region for polymerization with *Bst* to start. The sequence space was $4^{12} = 16777216$,

though sequences were not equally represented in the initial pool due to the bias. Assuming the equal occurrence of every possible sequence at our concentration of 10μM and volume of 15μL, each sequence would be represented approximately 5 million times, averting effects of undersampling.

The length distribution of sequences over time was analyzed with PAGE, Figure 2.15 **a**. The $ATGC_0$ pool displays replication to sequences more than 60 nt long after 3840 min, with most of the initial 12-mer depleted. The concentration profiles over strand length were obtained via ladder-calibrated SYBR Gold fluorescence intensity in PAGE gels, Figure 2.15 **b**. Similarly to the binary pools, replication of the $ATGC_0$ pool displays a double peaked length distribution with a long tail. The first peak at around 12 nt corresponds to the 12-mers of the initial pool not yet recruited for replication. The second peak, between 20 and 30 nt, could be due to fully bound duplexes that cannot melt at the incubation temperatures. The total DNA mass concentration also grows approximately linearly, with an incorporation rate about 4.5 times slower than for the **AT** experiments and 14 times slower than for the **GC** experiments. The nucleotide fraction determined by NGS was 24% **A**, 32% **T**, 21% **G** and 23% **C**. The recovery of sequencing signal from the reads, discussed in Appendix 2.D, required an additional adapter filtering step to partly meet the challenges associated with sequencing a random pool of all possible discrete lengths and all 4 bases.

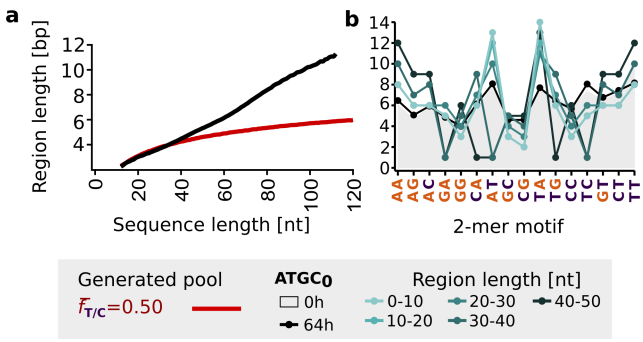


Figure 2.16: Longest self-complementary regions that were found for each sequence, plotted averaged per sequence length. a The polymerized $ATGC_0$ sample (64 h) displayed higher self-complementarity compared to a randomly generated homogeneous pool, particularly for sequences longer than 40 nt. This behaviour is similar to that of **AT** pools shown in Figure 2.14 **b**, for the longest self-complementary regions found with the analysis in **a**, the 2-mer motifs where plotted, in length subsets. The longer the self-complementary regions are, the richer in **AT** and **TA** motifs, when compared to the average of the whole $ATGC_{0,64h}$ pool.

Self-complementarity, that is the longest possible complementary overlap in each sequence over sequence length is shown in Figure 2.16 **a**. For binary systems, this parameter revealed an important difference between **AT** and **GC**. While for **AT** sequences longer than 40 nt, the longest self-complementary regions were longer than ones from a generated random pool, the **GC** sequences did not show any deviations. For the $ATGC_0$ pool, the sequences longer than 40 nt displayed increased self-complementarity similar to that of the **AT** ones. To understand whether the self-complementary regions were richer in specific motifs, particularly motifs containing **A** and **T**. The **ATGC** 2-mer motifs of self-complementary regions were computed, dividing the sequences into subsets of different lengths and comparing that to the 2-mer motif

distribution for the whole ATGC$_0$ pool after 64h, Figure 2.16 **b**. We thereby found that increases in length of self-complementary regions corresponded to an enrichment in **AT** and **TA** motifs. Simultaneously, **CC** and **GG** motifs were depleted more in longer self-complementary regions. The replication with a four-base pool therefore has characteristics that are a combination of both **AT** and **GC** pools analyzed. This promotes the validity of analyzing binary pools to simplify the full alphabet system while still being able to isolate its important characteristics.

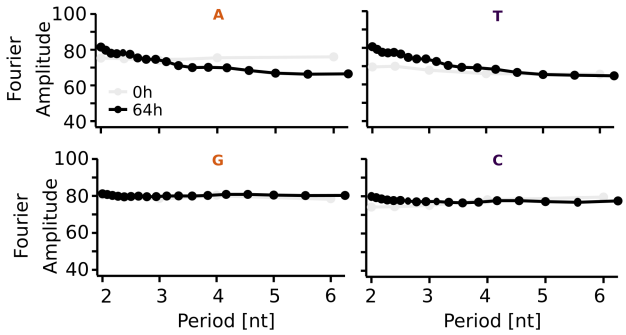


Figure 2.17: Fourier transforms for all four nucleotides, obtained from the conditional probability graphs for the 50 nt long sequences, $p(pos\ ii = A/T/G/C \mid pos\ i = A/T/G/C)$ as discussed in Section 2.3.5.1. As for the late time points of the binary systems, no specific periodicities are apparent, with the exception of **A** and **T** favoring shorter periodicities $\leq$3 nt and especially 2 nt, indicating **AT** zebraness in **ATGC** data as the 2-mer motif analysis did.

The analysis of periodicity for sequences of length 50 nt was conducted in a manner similar to binary systems, as shown in Section 2.3.5.1 and 2.3.5.2. The probability of a certain nucleotide in pos ii given the presence of that nucleotide in pos i, was computed for all four nucleotides. Subsequently, a Fourier transform of this conditional probability was applied to each position and the results were averaged across positions, Figure 2.17. In the case of ATGC$_0$, the potential periodicities of the early replicators could not be retrieved since only the initial pool and the late time point were sequenced. Nevertheless, the Fourier transforms of the late-time point concur with those of **AT** and **GC**, as no enhanced periodicities were observed.

In an analysis similar to the approach with binary pools performed in Section 2.3.4, the nucleotide fraction $f_{N(i)}$ of nucleotide N at each position i for sequences of the same length was also plotted. This is visualized in the 5′ to 3′ end direction. The initial pool consisted of 12-mer sequences with a bias towards **T**, Figure 2.18. For the initial pool, a divergent color scale ranging from grey to the color corresponding to the nucleotide (orange for purines and purple for pyrimidines) is used, centered around a homogeneous nucleotide fraction of 0.25 in white. Grey indicates the absence of the nucleotide in question. In contrast to the **AT** and **GC** samples, the ATGC$_0$ pool, being non-binary, required the plotting of the fraction for all four nucleotides. Some inhomogeneities were observed in the initial pool, particularly in the **T**-fraction, which was especially rich in **T** within the first 6 nt.

For the late time point, 64 h, the color scale was centered around the average nucleotide fraction for each nucleotide ($\bar{f}_N$) to highlight regions where each specific nucleotide deviated from the pool's average for that nucleotide. In this case, the $\bar{f}_N$ did not consistently correspond to 0.25, as the pool did not tend to a homogeneous nucleotide fraction through replication,

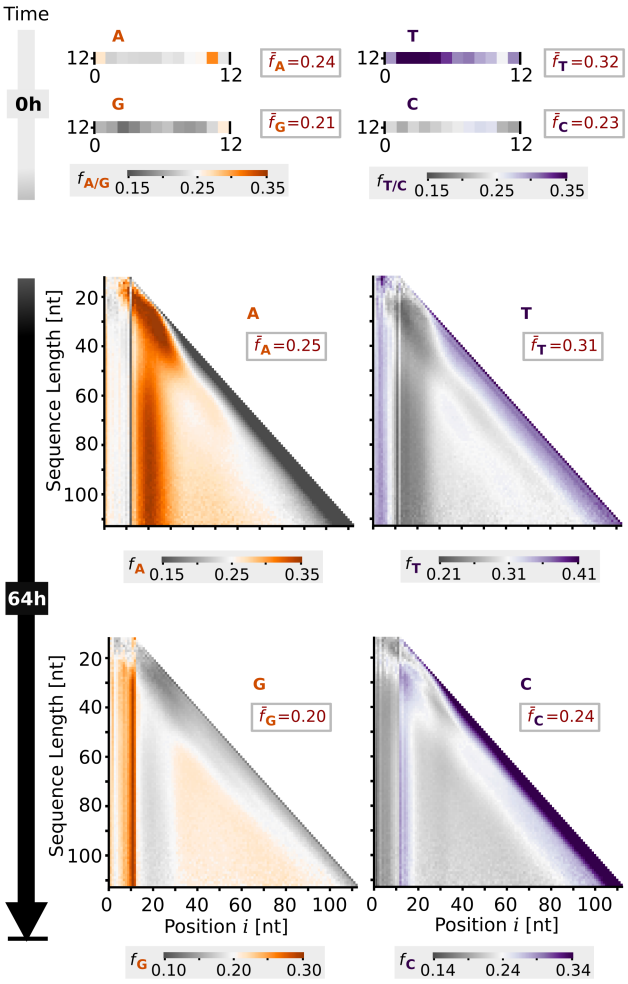


Figure 2.18: Nucleotide fraction for the 0 h and 64 h time point, for each of the four nucleotides. In the case of the 0 h time point, the divergent color scale was centered in 0.25, the homogeneous nucleotide fraction in a non-biased pool. The initial pool was biased towards **T** and slightly depleted in **A**, **G** and **C**. The nucleotide fraction for the 64 h time point was decomposed by length and position, for each of the four nucleotides. In this case, the divergent color scale was centered around the pool average nucleotide fraction, for each specific nucleotide. This way, it is possible to assess which regions of the pool have a higher than average nucleotide fraction. Grey represents absence of the nucleotide. The first 12 nucleotides at the 5′ end retain the initial sequence bias for all graphs, due to the directionality of the polymerization mechanism ($5' - 3'$) as has been seen for binary data sets. Gradients of alternating nucleotide fraction indicate the presence of long self-complementary regions.

likely due to the interplay of several more complex factors influencing the final nucleotide fraction in a full alphabet experiment. While the elongation mechanism is still the same compared to binary experiments – complementary nucleotides to the template are incorporated – duplexes with certain nucleotide biases are more stable than others. The stability of **ATGC** duplex sequences depends on the **GC** to **AT** ratio as well as on the contribution of the stacking energies from all the submotifs and the patterns present in fast replicators

and left-behind sequences would depend more on the replication temperature. To gain a comprehensive understanding of the intricate interplay among these and potentially other factors, conducting an **ATGC** screening experiment with various initial biases at both early and late time points and at different incubation temperatures would be required.

Nonetheless, the limited investigation of this 4-letter replication systems revealed patterns akin to those in binary systems. The initial 12-mer columns resulting from the directionality of polymerization were evident, along with clear regions of enrichment and depletion for all four nucleotides. Notably, the initial 12-mer sequences incorporated were richer in **G** and **T** compared to the overall average, while positions between 12 and 25 nt were notably **A**-enriched and slightly **C**-enriched. Unlike binary systems, not all nucleotide fraction gradients were antisymmetric; this was the case for **G** and less pronounced for **C**.

Overall, patterns resembled those of binary systems, validating the utility of binary pools as simpler, more controlled systems, isolating effects of replication on sequence pools. Detailed understanding of the mechanisms underlying a full 4-nucleotide data set however necessitates not only the sequencing of early time points to decipher the patterns of fast replicators, but especially the screening of different initial pool biases and incubation temperatures. Lastly, theoretical models would likely be required to tackle the interplay of the effects isolated in binary systems as well as new ones arising from 4 nucleotides.

2.3.8 Reproducibility

Reproducibility of the obtained length distributions with PAGE analysis was verified with different aliquots of the initial pools in similar experiments, leading to similar PAGE images. While reproducibility was not purposefully investigated by sequencing these aliquots, two independently synthesized initial pools similar in average base nucleotide fraction were obtained when attempting to perform experiments with a non-biased **AT** pool (50% **A**, 50% **T**). After sequencing, the base content was revealed to be similar to the A_0 pool, providing with a fully independent repeat, herein named A_0^*.

The positional nucleotide fraction is depicted in Figure 2.19, where the graphs for A_0 from Section 2.3.4 are reproduced alongside it. The prominent resemblance is a good indicator of general reproducibility not only for average pool nucleotide fractions, but also for the structure of sequences. The Fourier transform reveals the same periodicities of 2 and 3 nt. A slight enhancement in 4 nt periodicity for the A_0^* 2h data compared to the A_0 2 h is visible. As is explained in the Section 2.3.6, periodicity is a feature of fast replicators as they facilitate templation by increasing binding possibilities and possibly enabling catalytic network formation within the pool. While the two dominant periodicities are the shortest (2 and 3 nt), the enrichment of longer periodicities not to be precluded. Even though, A_0^* and A_0 have the same starting nucleotide bias, specific sequence composition varies, leading to slight variations in the outcome. This may explain the increased enrichment in 4 nt periodicity for A_0^*.

2.4 Conclusion

With this project it was demonstrated that, while biased pools that underwent templated replication display positional biases, the average pool nucleotide fractions became more homogeneous. Replication from two independently synthesized initial pools with the same

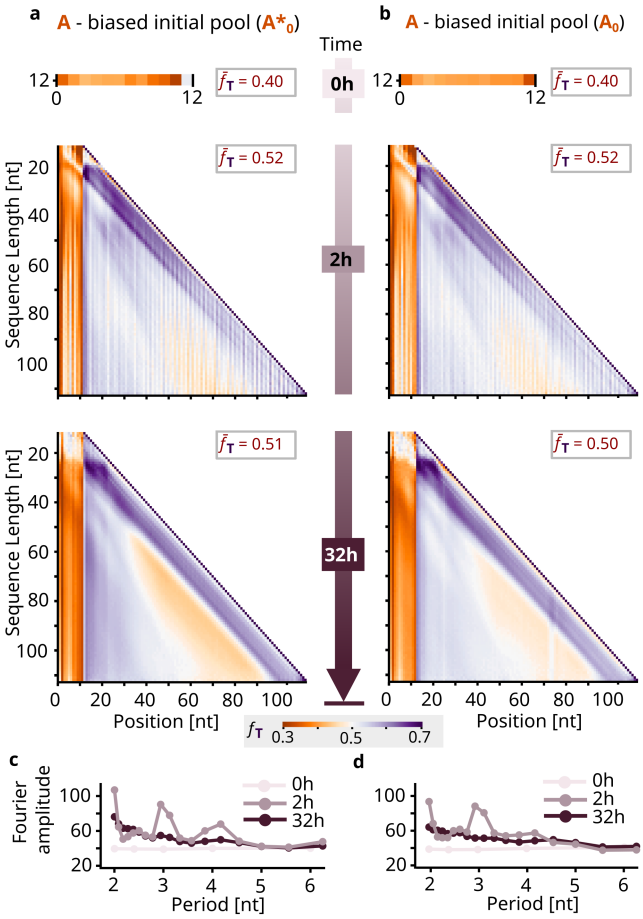


Figure 2.19: Reproducible sequence patterns formed from independent initial pools with the same nucleotide fraction bias a, positional nucleotide fraction over time for the A_0^* data set, which has the initial bias as A_0. The overall patterns observed are identical for both idential pools (A_0 is shown again in **b** as a comparison reference. **c** and **d**, The Fourier transform reveals the same periodicities of 2 and 3 nt for A_0^* and A_0, respectively.

bias resulted in reproducible length distributions, average pool nucleotide fractions and sequence structure, Section 2.3.8.

Compositional diversity, represented by the average pool nucleotide fraction, was shown to arise from biased binary pools via templated replication. This is a necessary characteristic for the exploration of sequence space with the possibility of generating a functional sequence. Similar conclusions have previously been described for binary DNA systems *in silico* [30], particularly for templated ligation.

Simultaneously, the replication of an initially biased pool resulted in regions in the replicated sequence that possess the same or the symmetric bias, which alternate and balance each other on average. This allows for a biased exploration of subsections of sequence space with structured sequences, without restricting the sequence space to a subset of similar sequences. Different nucleotide biases have been shown to correlate with enrichment of different secondary structures [127], implying that the sequences obtained from this templated

replication may exhibit a diverse range of secondary structure, which is in turn correlated with functionality.

Symmetry breaking, triggered by the selection for the reverse complement due to templation mechanisms, has been experimentally described for templated ligation. In a previous study [67] where binary **AT** pools were studied, two different sub-populations of sequences were found to contain a high amount of reverse complement sequences, with different nucleotide biases being enriched for each sub-population (an **A**-rich and a **T**-rich). Indeed, a comparable behaviour was observed also within single sequences.Highly periodic sequences were found to replicate faster, interestingly amplifying a periodic trimer structure in all studied pools. We attribute this to the potential emergence of cooperative sequence networks made up of subpopulations within the pools. These subpopulations would be characterized by reverse-complementary 3-mer periodic motif sequences that would cross-catalyze each other's rapid elongation.

Besides this agreement in 3 nt structure, the 2 nt periodicity differed for the two binary systems investigated. **AT** pools favored the 2 nt zebra motifs **AT** and **TA**, whereas **GC** pools preferred the bulky motifs **GG** and **CC**, likely due to intrinsic differences in stacking energies. Our findings, especially of the high self-complementarity in long **AT** sequences, support the mechanism of "hairpin elongation" for repetitive DNA, as previously suggested [98]. Repetitive DNA strands possess a high number of potential fold-back sites for hairpin formation. Repeated complete or partial melting, possibly induced by the strand displacing activity of *Bst*, alternating with hairpin formation and self-templation, would quickly elongate highly repetitive sequences.

In this study, an experimental model system was employed to provide insight into the role of replication as a mechanism of selection. Using a protein-based replication system with strand displacement (*Bst*), certain sequence patterns emerged as the fittest, found through the analysing of fast replicators. In addition, the dependency of the emergent structure on the initial pool was characterized. Overall, these findings contribute to elucidate the steps involved in the molecular evolution of short unstructured nucleic acids into long functional sequences.

2.5 Experimental Realization

2.5.1 Nucleic acid sequences

DNA oligonucleotides were purchased in dry form from biomers.net with HPLC purification and then adjusted to a stock concentration of approximately 200 µM with nuclease-free water (Ambion nuclease-free water from Invitrogen). All of the purchased DNA sequences with the corresponding nucleotide fraction bias measured by Illumina sequencing (Section 2.5.5 for method) are shown in Table 2.3. This bias differs from the requested bias upon ordering, which is shown in Table 2.1. This can be due both to synthesis and sequencing artifacts. The stocks solutions were stored at -20°C and denatured at 95°C for 2 min.

2.5.2 Reaction conditions

The polymerization reactions were performed with *Bst* 2.0 DNA Polymerase (New England BioLabs, #M0537S). The conditions were according to the protocol provided by the manufacturer: 1x Isothermal Amplification Buffer (containing 2 mM $MgSO_4$), 8 mM $MgSO_4$ (for a total

Table 2.3: List of all DNA sequences used in the experiments with the measured bias upon sequencing DNA sequences were ordered from biomers.net. **S** signifies **G** or **C** and **W** signifies **A** or **T**.

Name		Sequence		Measured Bias
G_0	5′	**SSSSSSSSSSS**	3′	(70% **G** :30% **C**)
C_0	5′	**SSSSSSSSSSS**	3′	(31% **G** :69% **C**)
A_0	5′	**WWWWWWWWWWW**	3′	(60% **A** :40% **T**)
A^*_0	5′	**WWWWWWWWWWW**	3′	(60% **A** :40% **T**)
T_0	5′	**WWWWWWWWWWW**	3′	(25% **A** :75% **T**)
$ATGC_0$	5′	**NNNNNNNNNNN**	3′	(24% **A** :32% **T** :21% **G**)

of 10 mM with the 2 mM $MgSO_4$ from the buffer), 320 U/mL *Bst* (all supplied when ordering the enzyme), 1.4 mM of each dNTP and 10µM DNA. **AT** samples were supplied with 1.4 mM dATP and dTTP and **GC** experiments with 1.4 mM dGTP and dCTP (all from Sigma-Aldrich) the **ATGC** experiments with all four nucleotides (1.4 mM of each). All experiments were conducted with initial DNA samples containing only random 12-mers provided by biomers.net, with binary base alphabets (**AT** , **GC**) in varying base content and for the **ATGC** experiment a full base alphabet, Table 2.3. The polymerization reactions were incubated in a standard thermocycler with the following protocol: 1. constant temperature (35°C for **AT** , 65°C for **GC** , 45°C for **ATGC**) for the reported time; 2. 90°C for 20 min to deactivate *Bst*. The incubation temperature was lower for **AT** than for **GC** due to differences in T_m, and based on a temperature screening performed with *Bst*.

2.5.3 Denaturing PAGE

PAGE was used to analyze and quantify the length distribution of the strands obtained at different time points of polymerization. The samples were run in a denaturing 15% polyacrylamide made from a 40% acrylamide/bis-acrylamide (19:1) stock solution (Carl Roth) and contained 50 wt% urea and 1x TBE (from 10x, Carl Roth) and polymerized with TEMED [2] and APS [3]. Each gel has a thickness of 0.75 mm and approximately 5 mL of the gel mixture. The gel mixture was prepared with 5 mL of the 15% PAA [4] mixture, 25 µL of APS and 2.5 µL of TEMED.

The gels were pre-heated in the electrophoretic chamber at 300 V for 27 min. The samples were then loaded, in a mixture with a ratio of 2:7 of sample to loading dye. Loading dye is prepared in-house (for 10 mL: 9.5 mL formamide, 0.5 mL glycerol, 1 µL EDTA [5] (0.5 M) and 100 µL Orange G dye (New England BioLabs). The samples were at 50 V for 5 min followed by 300 V for 25 min. After the run, the gels were stained with a 2x SYBR Gold (Thermo Fischer Scientific) dilution in TBE [6] buffer 1x. They were then rinsed with 1x TBE buffer twice and imaged using a Bio-Rad ChemiDoc MP imaging system. The 20-100 bp ladder (DNA oligo

[2] short for Tetramethylethylenediamine
[3] short for Ammonium persulphate
[4] short for polyacrylamide
[5] short for ethylenediaminetetraacetic acid
[6] short for Tris-Borate-EDTA

length standard 20/100 Ladder, IDT) was supplied in a final concentration of 2.04ng/µL (for each rung) and the 100-1517 bp ladder (100 bp DNA Ladder, New England BioLabs) in a final concentration of 71.4 ng/µL (for all rungs; concentrations vary by n-mer as described by the manufacturer). Finally, the obtained micrographs were loaded into and analyzed with a self-written LabVIEW program for quantification of the concentration-per-length, Section 2.5.4.

2.5.4 PAGE quantification

PAGE images were analyzed with a self-written (adapted from an existing program of the AG Braun) LabVIEW tool, which allowed to obtain the concentrations of DNA strands depending on length from smears in the gel lanes by using known total molar concentrations of each lane and the linear increase in fluorescence intensity of SYBR gold with strand length and concentration [66]. The total molar concentration is known since it stays constant throughout the experiment as no new strands may appear through polymerization. Effects of hydrolysis should be small. The analysis happened in three main steps: gel image to lane data conversion, ladder peak detection and concentration analysis.

Firstly, the PAGE image was converted to lane data by loading the image and extracting the intensities, matching a lane mask with the imaged lanes and performing a background correction. A uniform x-axis for all lanes was then calculated and the y-axis rescaled to make them comparable. In the second step, the approximate positions of the ladder n-mers in the leftmost lane of the gel (ladder peak positions) were detected. The program then fitted each peak with a Gaussian function individually to obtain the precise ladder peak position and interpolate the positions to arrive at a position vs. length function $\mu(m)$ as explained in more detail in Appendix 2.C.

In order to quantify the concentrations in the third step, an intensity to concentration relation was obtained for each gel lane individually by knowledge of the constant total molar concentration. This had the advantage of precisely covering the whole range when calculating concentrations for all n-mers and permitted to obtain a measure of total molar concentration for each lane allowing a normalization of each lane for correction of experimental variability and better comparability.

2.5.5 Illumina sequencing

Samples were sequenced by the Gene Center Munich (LMU) using the NGS Illumina NextSeq 1000 machine (flow cell type P2, 2 x 50bp with 138 cycles for 100bp single-end reads; at most 120 bp with 2 indexes were read, with declining quality towards the end). 50 million reads were ordered for each sample. Before sequencing, the samples were prepped using the ACCEL-NGS 1S Plus DNA Library Kit (Swift Biosciences) for library preparation. The raw sequencing data obtained, in FastQ format, was processed in this order by demultiplexing, quality score trimming, and regular expression filtering. Demultiplexing was performed with software from Galaxy servers [2], provided by the Gene Center Munich. During sequencing, each read base was assigned a Phred quality score $Q = -10 \log_{10} P$, where P is the probability of an incorrectly read base [39]. Using Trimmomatic [13] we trimmed low quality segments by running a sliding window of 4 nt in the 3′ to 5′ end direction over the sequence that allowed a minimum average Phred quality of 20, otherwise trimming at the leftmost base of the window, corresponding to an average accuracy of at least 99%. As the experimentally

obtained sequences were appended on the 3′ terminus with a **CT** -region followed by an **AGAT** during sequencing preparation, those needed to be found and cut, for which we employed the following regular expressions:

(^[AT]{12,})(?=([CT]{4,}AGAT)) for **AT**
(^[CG]{12,})(?=([CT]{4,}AGAT)) for **GC**
(^[ATGC]{12,})(?=([CT]{4,}AGAT)) for **ATGC**

This also ensured that only binary sequences were included in the analysis of binary pools. For the ATGC experiment, a further adapter filtering step was employed to recover the sequencing signal from the adapter contaminated reads, Appendix 2.D.

Appendix

2.A *Bst* enzyme specifications

Bacillus stereathermophilus polymerase I (*Bst*) is a thermophilic, strand displacing A family polymerase used in many isothermal amplification applications [4]. The enzyme binds to a double stranded segment and elongates it in the 5′-3′ direction, adding the dNTPs complementary to the template and displacing any other downstream bound strands. Specifically, in this work, Bst 2.0 DNA Polymerase was used. This is an *in silico* designed homologue of the large fragment of the *Bst* enzyme that contains the 5′-3′ polymerase activity, but lacks 5′-3′ exonuclease activity. It also has higher speed, yield, salt tolerance and thermostability than the wild-type. This enzyme should have an activity of about 10% for 35°C and 100% for 65°C, and has maximum performance at 4-10mM Mg^{2+} according to the manufacturer (New England Biolabs). The large fragment of *Bst* has proofreading activity which contributes to its high fidelity. This is not achieved through an exonuclease domain, but though a mechanism that checks the structure of the incorporated nucleotide at the active site [21]. The active site of the protein interacts with the minor groove of the double stranded DNA helix, particularly the 4 base pairs closest to the 3′-OH terminus. The proofreading is done at the last 3′ base. In case the base added is wrong (i.e. not complementary to the template), the 3′-OH terminus will not be oriented correctly and the elongation will not proceed [64]. In addition to the downstream bound strands, any secondary structure of the primers or template is denatured by *Bst* due to its strand displacing activity – it does not 'slip' [140].

2.B SYBR Gold staining efficiency

The staining efficiency of SYBR Gold exhibited significant variability based on nucleotide composition in previous literature. It was found that the efficacy of SYBR Gold staining is predominantly influenced by the composition when for the case of ssDNA, in opposition to ssDNA which displays consistent staining efficiency regardless of sequence [53]. Notably, ssDNA that is composed of a single base (homopolymers), with the exception of poly-`G`, exhibited no discernible staining. Furthermore, ssDNA lacking complementary bases demonstrated distinct binding efficiencies to SYBR Gold. Conversely, it was noted that ssDNA featuring some degree of complementary bases bound to SYBR Gold and displayed uniform staining, even under denaturing PAGE.

While the mechanism behind the differential staining is not fully understood in [53], it is proposed to occur through the transient formation of secondary structure. Despite these data set falling within the category of ssDNA formed by complementary bases, tests to assess whether the SYBR Gold binding efficiency was dependent on the composition were performed. To do this, 12-mer sequences corresponding to the initial pool were stained. These are the most heavily biased samples and the a priori bias is known through sequencing, allowing us to analyse a 'worst-case scenario'. As polymerization progresses, nucleotide

Table 2.4: Concentration of at least three independent stocks of biased initial pools obtained via absorbance at 260 nm A_{260}. The independent stocks were provided as separate aliquots for the same pool by the DNA manufacturer (biomers.net) and diluted to approximately 10 µM. For each of the pools, the composition of each nucleotide was assessed by NGS. The absorbance at 260 nm of each of these stocks was measured using Nanodrop and converted to concentration by the Lambert-Beer law. The extinction coefficient ϵ_{260} was calculated from the individual nucleotides' coefficients using the nucleotide composition data.

Initial pool	Composition	ϵ_{260} [$\mathrm{mM^{-1}\,cm^{-1}}$]	A_{260}	Concentration[µM]
A_0^*	60% **A** 40% **T**	142.4	0.128	9.0
			0.139	9.7
			0.143	10.0
			0.171	12.0
A_0	60% **A** 40% **T**	142.4	0.143	10.0
			0.127	8.9
			0.120	8.5
			0.127	8.9
T_0	25% **A** 75% **T**	108.8	0.142	13.1
			0.126	11.5
			0.118	10.8
			0.126	11.6
G_0^*	46% **G** 54% **C**	122.7	0.121	9.9
			0.124	10.1
			0.225	18.3
			0.128	10.4
G_0	70% **G** 30% **C**	133.1	0.131	9.8
			0.131	9.9
			0.150	11.3
			0.160	12.0
C_0	31% **G** 69% **C**	116.2	0.182	15.7
			0.166	14.3
			0.186	16.0
			0.144	12.4
$ATGC_0$	24% **A** 32% **T** 21% **G** 23% **C**	121.9	0.108	8.9
			0.100	8.2
			0.167	13.6

composition undergoes changes, diminishing the inherent bias. Consequently, the impact of nucleotide composition dependence primarily concerns the quantification of the initial pool concentration.

The concentration of three independent stock solutions was quantified for 12-mer pools of different composition by measuring the absorbance at 260 nm with Nanodrop. The absorbance of these replicates is shown in Table 2.4. The extinction coefficient of ssDNA depends on the nucleotide composition, as it varies for different deoxyribonucleotide 5′-monophosphates [20]. The extinction coefficient of a single DNA strand can be approximated by adding up the ex-

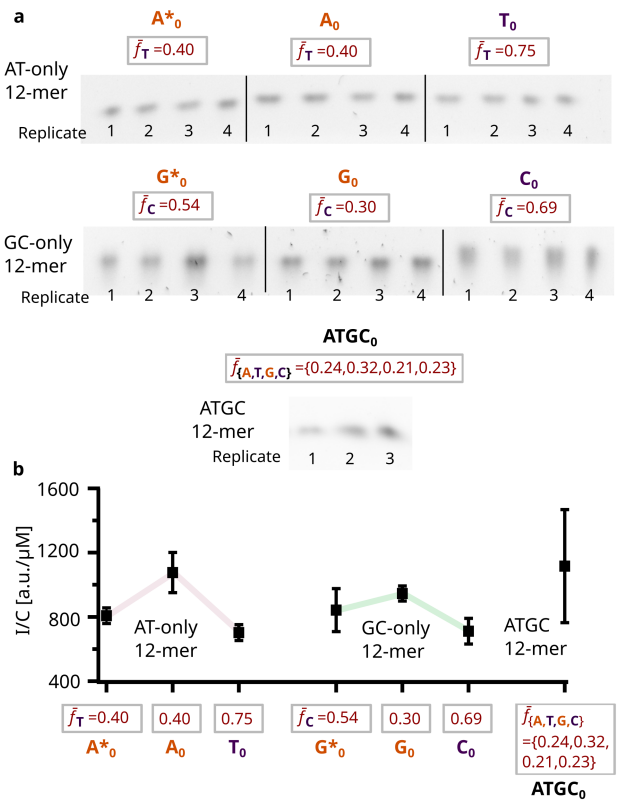


Figure 2.20: PAGE quantification of SYBR Gold signal for samples with different nucleotide composition. Samples with 12-mer DNA of known concentration consisting of **A** / **T** -only, **G** / **C** -only or with four bases and with varying nucleotide composition were loaded in a 15% PAGE, with at least three replicates per composition. The nucleotide composition was assessed with NGS. The fluorescence intensity of the bands was measured with a self-written Labview script, described in Section 2.5.4. **b** The SYBR Gold intensity obtained with the PAGE quantification in **a** was then normalized for the concentration for each sample to obtain a value of fluorescence-per-concentration (I/C). Data corresponds to average ± one standard deviation.

tinction coefficients of its individual bases. To do this, the composition of the initial pools obtained via NGS was used to perform a weighted average of the extinction coefficients provided in [20], see Table 2.4. Lastly, the concentration of DNA in each replicate was determined applying the Lambert-Beer law.

Besides nucleotide composition, SYBR Gold binding efficiency also depends on DNA length and concentration. Extracting the concentration of DNA of certain lengths from the gel image is thus difficult for two reasons. Firstly, the functional relation of imaged intensity of fluorescence to the concentration of DNA in the gel is usually unknown. However, for SYBR gold, fluorescence signal was found to increase linearly with stained DNA concentration as well as DNA strand length for a wide range of concentrations [66].

Secondly, DNA strands of the same length move through the gel with a slightly different speed due to diffusion and influence of sequence composition, leading to a smearing of signals through an overlap of n-mer bands (DNA strands with a length of n nt) with neighboring $n \pm \Delta n$ bands. As polymerization, unlike for example ligation, produces DNA strands of

all possible discrete lengths which result in continuous intensity smears in gels, a direct summation of the intensity signal for DNA strands of a specific length is impossible.

2.C Smear quantification from PAGE

In order to quantify individual strand lengths from a length distribution with all possible discrete lengths, yielding a smear in the PAGE, a model relying on small second order changes in intensities of neighboring DNA strands was developed. The central idea of the model is that for obtaining the area A of a centered symmetric peak, e.g. a Gaussian peak, instead of integrating the peak function from $-\infty$ to ∞, it is possible to integrate the sum of infinite evenly spaced equally shaped peaks from $x - \frac{1}{2}\Delta\mu$ to $x + \frac{1}{2}\Delta\mu$ with peak centers spaced with $\Delta\mu$ (2.1):

$$A = \int G_m(h_m, \sigma_m, \mu_m)\, x = \int_{\mu_m - \frac{1}{2}\Delta\mu}^{\mu_m + \frac{1}{2}\Delta\mu} \sum_{n=-\infty}^{\infty} G_n(h_m, \sigma_m, \mu_n)\, x \tag{2.1}$$

where h describes the height, σ the width and μ the position of the peak $G(h, \sigma, \mu)$.

As peaks are centered, one can then neglect contributions from the sides with little error and consider only the immediate neighboring peaks (2.2):

$$A \approx \int_{-2\sigma_m}^{2\sigma_m} G_m(h_m, \sigma_m, \mu_m)\, x \approx \int_{\mu_m - \frac{1}{2}\Delta\mu}^{\mu_m + \frac{1}{2}\Delta\mu} \sum_{n=-\Delta m}^{\Delta m} G_n(h_m, \sigma_m, \mu_n)\, x \tag{2.2}$$

with Δm such that $|\mu_{m\pm\Delta m} - \mu_m| \approx 2\sigma_m$.

The inner part of the integral resembles the measured gel intensity $I(x)$ at point x, which is described by (2.3):

$$I(x) = \sum_{n=-\Delta m}^{\Delta m} G_n(h_n, \sigma_n, \mu_n) \tag{2.3}$$

when gel bands are modeled as centered symmetric peaks. The only difference is that in the gel, neighboring peaks are not evenly spaced and are described by different h's and σ's (observe that h_m and σ_m were replaced by h_n and σ_n).

Small second order changes of intensities by length

But how large is the error, if neighboring gel peaks are assumed to be evenly spaced and equally shaped? – Very little error can be associated with the variability of peak spacing in the immediate vicinity of each peak. For the peak heights and widths, this doesn't hold. Deviations would lead to larger errors.

In most cases though, while intensities of neighboring gel bands vary, the change in intensities doesn't vary much. But when second order changes are small, peaks of same distance to the one that is to be measured can be described with symmetric deviations in h and σ (2.4) and (2.5):

$$G_{m-\Delta m} = G(h_m \pm \Delta h_m, \sigma_m \pm \Delta\sigma_m, \mu_{m-\Delta m}) \tag{2.4}$$

$$G_{m+\Delta m} = G(h_m \mp \Delta h_m, \sigma_m \mp \Delta\sigma_m, \mu_{m+\Delta m}) \tag{2.5}$$

In this case, the relative error in the area calculation (2.2) can expected to be smaller than $\frac{2\Delta h_m \Delta\sigma_m}{h_m \sigma_m}$ and should stay reasonably small unless intensities change abruptly with certain oligomer lengths.

Assuming a continuous smear with slow second order change in intensity, total intensities A_m of m-mer peaks are approximated well by (2.6):

$$A_m = \int_{\mu_m - \frac{1}{2}\Delta\mu}^{\mu_m + \frac{1}{2}\Delta\mu} I(x)\, x \tag{2.6}$$

Required measurements: intensities and ladder peak positions

The required values remaining besides the gel intensity $I(x)$ are then the position of the m-th peak, μ_m, and its spacing to neighboring peaks, $\Delta\mu$, which was assumed to be constant in its vicinity and can thus be calculated as (2.7):

$$\Delta\mu(m) = \frac{\mu_{m+1} - \mu_{m-1}}{2} = \frac{\Delta\mu_{\pm m}}{\Delta m} \approx \frac{\mu(m)}{m} \tag{2.7}$$

with the continuous interpolation $\mu(m)$ of μ_m's, leaving only the function $\mu(m)$ relating oligomer lengths to gel positions to be found.

To determine $\mu(m)$, DNA ladders can be used: by interpolating the ladder rung-mer peak positions of known oligomer lengths with a function relating in-gel-distances to product lengths, $\mu(m)$ is acquired. The simplest function yielding a good fit is a logarithm scaled by m, $\mu(m) = a\frac{\ln(bm)}{m}$, with fit parameters a and b. For fitting of long strands (>100nt), another logarithmic factor improved the fit even more, giving the final fit function (2.8):

$$\mu(m) = a\frac{\ln(m)}{m} + b\ln(m) + c \tag{2.8}$$

with peak position in gel μ, oligomer length m and fit parameters a, b and c. Fits were performed using the Levenberg-Marquardt algorithm implemented in LabVIEW.

Total molar concentrations

The accuracy of the PAGE gel quantification, should the total molar concentrations of DNA strands in each lane be known, can finally be verified by comparing calculated total molar concentrations to the known total molar concentration of the probes. The molecular weight of a single-stranded polymerized DNA oligomer can approximately be calculated as (2.9):

$$\text{molecular weight } [g/mol] = n * 308.95 - 61 \tag{2.9}$$

where n is the number of nucleotides in the n-mer. 308.95 is the average molecular weight of an incorporated nucleotide (**A**: 313.2, **T**: 304.2, **C**: 289.2, **G**: 329.2) and -61 is obtained by subtracting one phosphate (weight 79) for the hydroxyl-end and adding one water molecule (weight 18).

Reversing this line of thought, intensities can also be related to concentrations for each lane individually, as long as their molar concentrations are known. This normalization for each lane additionally helps to account for experimental variations in intensities introduced through pipetting of low volume / high viscosity samples and possible evaporation effects

in thermocyclers. For these reasons, each gel lane was indeed normalized to have the total molar concentration (the sum of molar concentrations for all oligomer lengths) equal the known molar concentration of the initial sample for gel analysis in this work – the total molar concentration was known to equal the initial molar concentration at each time because elongation of initial DNA strands does not change total molar concentration.

2.D Sequencing signal recovery in ATGC data

For **ATGC** data sequenced by NGS, the recovery of the actual experimental data from reads that are artifacts, such as those resulting from the sequencing of adapters, proves challenging since there is neither a genome to compare the reads to, nor a specific length (as for ligation) or a limited alphabet (as for binary data). In addition to the RegEx filtering step given in the methods section (to remove the **CT**-tail and **AGAT**), RegEx filtering the **ATGC** sequencing data for all possible 12 nt long adapter snippets is an effective way to recover the signal. With this in mind, a LabVIEW program was written to generate a list of RegEx filtering expressions and perform the adapter filtering.

The adapters in this case were:

GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCTCGCGCATCTCGTATGCCGTCTTCTGCTTG and
AATGATACGGCGACCACCGAGATCTACACTCAGAGCCACACTCTTTCCCTACACGACGCTCTTCCGATCT for 0 h,
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCGATAGATCTCGTATGCCGTCTTCTGCTTG and
AATGATACGGCGACCACCGAGATCTACACCTTCGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCT for 64 h

To check the quality of the adapter filtering, the same procedure was applied to the A_0 and the G_0 experiments. This allowed us to compare the results between RegEx filtering for **CT**-tail and **AGAT** plus adapter filtering with binary RegEx filtering. We assessed both the total read counts and the length distributions of the 8 data sets, Figure 2.21.

These graphs allow to conclude that the adapter filtering was, with some minor deviations, effective for the A_0 experiments, while strong artifacts remain for the G_0 ones. For the **ATGC** case, the length distribution for the 0h initial time point does not exhibit a specific artifact with a peak at 43 nt as seen in the G_0 ones, but still displays a long tail after the 12 nt long sequences, which might in part be due to adapter contamination. To address the long tail, sequences up to 16 nt were considered for the analysis of the **ATGC** 0h data set. In this range, the filtering works reasonably well for the two **ATGC** data sets.

The two individual adapters for each sequenced experiment in this study were each made up of three segments, of which the beginning and ending segments (Illumina TruSeq HT adapters D701-D712 and D501-D508) remained the same across all experiments. These were as follows:

D701-D712: **GATCGGAAGAGCACACGTCTGAACTCCAGTCAC** [i7] **ATCTCGTATGCCGTCTTCTGCTTG**,
D501-D508: **AATGATACGGCGACCACCGAGATCTACAC** [i5] **ACACTCTTTCCCTACACGACGCTCTTCCGATCT**

The two individual middle segments (i7 and i5) for each sequenced experiment are given in Table 2.5.

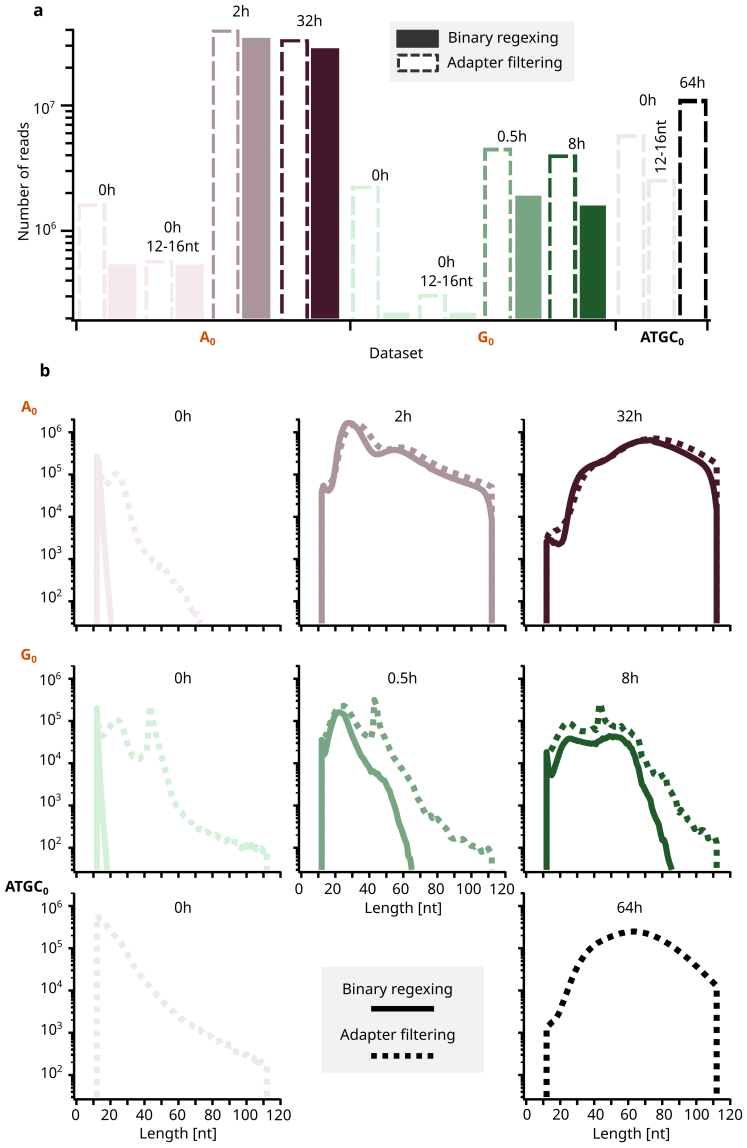


Figure 2.21: Assessment of adapter filtering efficiency. Comparison of sequencing read counts and length distributions for the A_0, G_0 and $ATGC_0$ experiments, after RegEx filtering for binary sequences or the combination of RegEx filtering for **CT** -tail and **AGAT** with adapter filtering. While artifacts are present in the 0 h initial time points and the G_0 ones with a peak at 43 nt for adapter filtering, the A_0 ones show good agreement with the binary RegEx filtering. No binary RegEx filtering being possible for the ATGC, artifacts have to be assessed in comparison to the other binary samples. For the 0h time point, sequences up to 16 nt were considered, and since there is no specific artifact with a peak at 43 nt, it is assumed that the late time point predominantly contains sequencing signal rather than adapter noise.

Experiment	i7 index	i7 sequence	i5 index	i5 sequence
$T_{0,0h}$	D706	GAATTCGT	D503	AGGATAGG
$T_{0,2h}$	D707	CTGAAGCT	D504	TCAGAGCC
$T_{0,32h}$	D708	TAATGCGC	D505	CTTCGCCT
$A^*_{0,0h}$	D701	ATTACTCG	D506	TAAGATTA
$A^*_{0,2h}$	D702	TCCGGAGA	D507	ACGTCCTG
$A^*_{0,32h}$	D707	CTGAAGCT	D508	GTCAGTAC
$A_{0,0h}$	D703	CGCTCATT	D508	GTCAGTAC
$A_{0,2h}$	D704	GAGATTCC	D501	AGGCTATA
$A_{0,32h}$	D705	ATTCAGAA	D502	GCCTCTAT
$C_{0,0h}$	D702	TCCGGAGA	D503	AGGATAGG
$C_{0,0.5h}$	D703	CGCTCATT	D504	TCAGAGCC
$C_{0,8h}$	D704	GAGATTCC	D505	CTTCGCCT
$G_{0,0h}$	D711	TCTCGCGC	D508	GTCAGTAC
$G_{0,0.5h}$	D712	AGCGATAG	D501	AGGCTATA
$G_{0,8h}$	D701	ATTACTCG	D502	GCCTCTAT
$ATGC_{0,0h}$	D711	TCTCGCGC	D504	TCAGAGCC
$ATGC_{0,64h}$	D712	AGCGATAG	D505	CTTCGCCT

Table 2.5: Middle segments of adapter sequences.

3 Sequence-specific ligation of short RNA with 2',3' cyclic phosphates

Summary

Templated ligation offers an efficient approach to replicate long strands in an RNA world with fewer condensation steps than a base-by-base replication mechanism. The 2′,3′-cyclic phosphate (>P) is a prebiotically available activation that forms during RNA hydrolysis, nucleobase polymerization and prebiotic phosphorylation. Using PAGE and HPLC, the impact of reaction conditions on yield, kinetics and sequence fidelity was investigated. Templated ligation of RNA with >P was shown to proceed in simple low-salt aqueous solutions with 1 mM $MgCl_2$ under, alkaline pH ranging from 9 to 11 and temperatures from -20 to 25°C yielding up to 40% in 7 days. No additional added organocatalysts were required. Moreover, an increased ratio of 50% of the canonical 3′-5′ linkages at the ligation site was observed, an improvement over previously reported aqueous condensations involving >P. The reaction proceeds in a sequence-specific manner, with an experimentally determined ligation fidelity of 82% at the 3′ end and 91% at the 5′ end of the ligation site. Multistep ligations within a splinted RNA system using >P were shown to generate long RNA molecules on the length scale of 100 nucleotides through cross-templation. Five ligations created a 96-mer strand, demonstrating a possible pathway for ribozyme assembly. Due to the low salt requirements, the ligation conditions will be compatible with strand separation. Templated ligation mediated by 2′,3′-cyclic phosphate in alkaline conditions therefore offers a performant replication and elongation reaction for RNA on early Earth. [1]

[1]This chapter was published by Serrão and Wunnava et. al [18] in the Journal of the American Chemical Society (JACS) and is here adapted and reprinted in part with permission from JACS. Full article attached in the List of Publications.

3.1 Motivation

Nucleic acid replication is essential for the propagation of genetic information and, therefore is a central step for the origin of life [48]. An early form of molecular evolution that preceded catalytic polymers, i.e. enzymes and ribozymes, required a non-enzymatic copying mechanism. While primer extension by the addition of mono-, di- or tri- nucleotides has been demonstrated to copy shorter sequences [72, 124, 142, 157], its processivity is limiting and the combination with strand separation is a considerable hurdle due to the high divalent salt concentration necessary. A simple way to overcome the issue of processivity would be templated ligation, which reduces the number of steps-per-length required to generate an oligonucleotide and if possible to operate at low magnesium concentration would offer an easier strand separation particularly when coupled to non-equilibrium microenvironments with salt and pH cycling [114]. Additionally, ligation chain reactions have been shown to offer exponential replication [37].

Due to its high processivity, templated ligation can bridge the gap between two length scales in the nucleic acid emergence timeline: the very short oligomers with 10 nucleotides or less and the very long, structurally complex functional replicators (see Figure 3.1). Base-by-base mechanisms likely still existed in earlier stages, preceding templated replication in order to generate strands long enough to form the duplexes necessary for templation. After ribozymatic replication emerges from molecular evolution, this more efficient mechanism would take over the slower non-enzymatic replication chemistries.

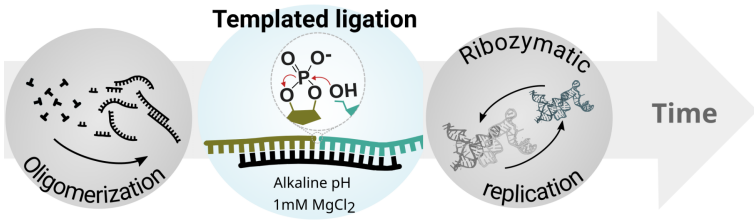


Figure 3.1: Templated ligation bridges the gaps between two length scales in the emergence of functional nucleic acids. First, short oligomers (< 10 nt) are formed by non-enzymatic likely non-templated prebiotic chemistries. These base-by-base chemistries are not effecient, as each condensation step only adds one nucleotide. When strands are long enough to form stable double strands, templated ligation may have taken over providing with a higher processivity. This would have allowed the formation of strands with tens or hundreds of base-pairs which are long enough to undergo molecular evolution and form ribozymes. Ribozymatic replication would then take over with potentially higher speed and sequence fidelity.

The state-of-the-art method for templated ligation makes use of phosphoramidates, such as phosphorimidazolides [143, 158]. The need for a separate (ex-situ) pre-synthesis step with condensing agents, coupled with their short half-life reduces their prebiotic likelihood. Furthermore, in situ activation with prebiotically plausible organocatalysts in ligation-compatible scenarios has not been shown. Disregarding the need for multiple synthesis steps required to make the phosphorimidazolides [61, 104], imidazole activated oligonucleotides are less reactive than their mononucleotide counterparts lowering the yield of templated ligation compared to that of polymerization [143]. Moreover, studies demonstrating the assembly of a long catalytic RNA by templated ligation of imidazole-activated oligonucleotides required a

high concentration of Mg^{2+} which leads to product inhibition [143, 159, 160]. This strongly supports the need for an efficient ligation system compatible with strand separation.

The quest for such a system led to cyclic phosphates since they generate short oligomers in the dry state [27, 138, 139] which retain the active >P ends and further undergo ligation. More importantly, they represent a simple and endogenous activated group, minimizing the need for complex multi-step synthesis and ex-situ activation with other organocatalysts. 2′,3′ cyclic phosphate (>P) endings were likely readily available in the prebiotic pool since they are the primary product of prebiotic nucleotide synthesis [103] and phosphorylation reactions [47, 75], Figure 3.2 **"Synthesis"**. Moreover, ribozymatic [69] and alkaline hydrolysis of RNA strands via transesterification [14, 17, 73] produce >P ends (Figure 3.2 **"Hydrolysis"**) which are substrates for ribozymes catalyzing phosphodiester bond formation [90, 123]. Hydrolysis of >P results in 2′- or 3′-monophosphate, the recycling of which, i.e. the re-cyclization to activated oligonucleotides with >P, under prebiotic conditions was also demonstrated [123] (Figure 3.2 **"Cyclization"**), suggesting a way for the in situ recycling of hydrolyzed substrates.

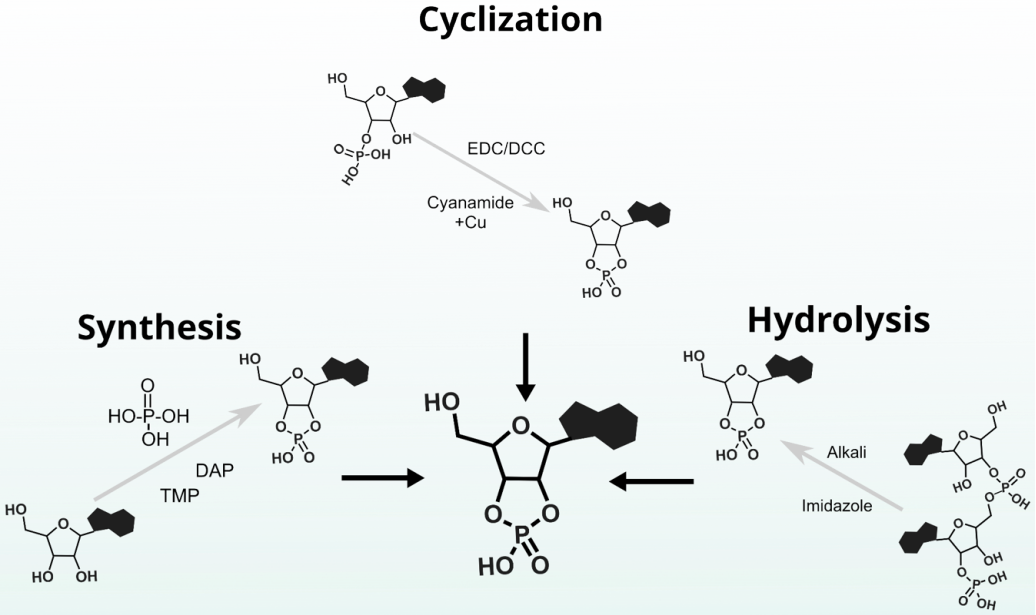


Figure 3.2: 2′,3′-cyclic phosphate were likely available in the prebiotic pool. Prebiotic nucleotide synthesis yields >P as a primary product whereas phosphorylation with TMP[2] and DAP[3] have been shown to phosphorylate nucleosides yielding >P. Alkaline RNA oligomer hydrolysis yields bot cyclic phoshpate or 2′- or 3′-monophosphate end, where the later can undergo cyclization, with the addition of EDC or Cyanamide to yields >P.

Biochemical protocols using EDC[4] at low temperatures are also common [76, 90, 92]. The widespread presence of these activated phosphate groups in established prebiotic pathways and their relatively longer half-life was the motivation to investigate their role in ligating RNA oligomers in a templated setting and their relevance for early RNA replication. Previous work with >P containing oligonucleotides has demonstrated template copying only through ligation with DNA:RNA chimeras, resulting in low yield and a predominance of 2′–5′ linkages [76, 77]. Non-templated ligation of random sequences with >P has also been shown to proceed in

[4]short for 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide

eutectic phase at low rates [91]. However, the potential of ligation of RNA sequences with >P ends for quantitative genetic copying remained largely unexplored.

3.2 Scientific approach

The hypotheses behind the experiments in this project were: *i*) Does a fully RNA-based system, with a >P end, ligate on a template? *ii*) What are the characteristics of this reaction in terms of both yield and kinetics? *iii*) Is the reaction sequence dependent, and, if so, with which fidelity? *iiii*) What type of linkages are formed? *iv* Can successive cooperative ligation of short oligomers yield long ribozyme-sized products?

To answer these questions a system of three short RNA strands was designed, with the sequences based on the work by Murayama et. *al* [89]. The system consists of two primers (*a* and *b*) that bind on a template *BA*, Figure 3.3 **a**. Primer *a* has a >P at the ligation site terminus that can suffer a nucleophilic attack from the 5′ hydroxil (Figure 3.3 **b**) and yield product *ab*. In case the nucleophilic attack is performed by water, the ring opens and a 3′- or a 2′-monophosphate forms, yielding an inactive primer (*a-P*) that cannot undergo ligation. The sequences of a, b, ab and BA are shown in Table 3.1.

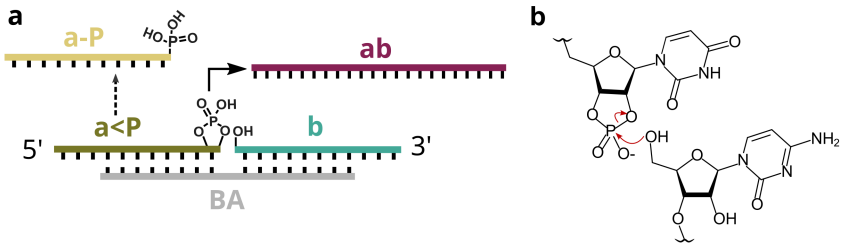


Figure 3.3: Non-enzymatic template directed ligation of short RNA strands. a, Schematics of reaction design. Both primers (*a* and *b*) bind on the complementary template, *BA*. The primer *a* has a 2′,3′-cyclic phosphate, while *b* contains a 5′-OH group. **b**, 5′-OH performs a nucleophilic attack on the cyclic phosphate group and forms a phosphodiester bond between the two primers, leading to the ligation product strand *ab*. As a side reaction, the cyclic phosphate in a can also hydrolyse, rendering a inactive.

Table 3.1: Sequences of primers, template and product used for the ligation reaction. RNA sequences were ordered from biomers.net.

Name		Sequence	
a	5′	AAAGCAUCAGU	3′
b	5′	CUCAUAGGAAA	3′
BA	5′	CCUAUGAGACUGAUGC	3′
ab	5′	AAAGCAUCAGUCUCAUAGGAAA	3′

This experimental approach implies a series of decisions and assumptions that facilitate researching the hypothesis in a laboratory setting and isolating the contribution of each of

the parameters to RNA ligation. The topic of **reaction tube vs. rocky pore**, has been discussed in Section 1.2 and a similar argumentation can be applied to the assumptions in this project. Other approximations explained in both 1.2 and 2.2 are now lifted. Specifically, because the actual non-enzymatic chemistry is at the focus of this project, in opposition to the impact of selective pressures on molecular evolution, some decisions were made so that the system is more prebiotically plausible. The three stands partaking in ligation are now RNA, with a full-alphabet pool and are ligated through non-enzymatic means.

- **Artificial strand design vs. pool generated in prebiotic conditions**

  The sequences were designed to have a balanced amount of all nucleotides (about 50% GC-content) and to form a stably bound duplex. Prebiotic pools generated through non-enzymatic oligomerization yield sequences that are very short (less than 10 nt in length) with longer sequences being several orders of magnitude lower in concentration [27, 31, 72, 142]. In order to be to characterize the ligation reaction yield, kinetics, and sequence dependence the starting point was therefore a simple three-sequence system with well-known sequence and concentration.

3.3 Results and Discussion

3.3.1 Condition screening

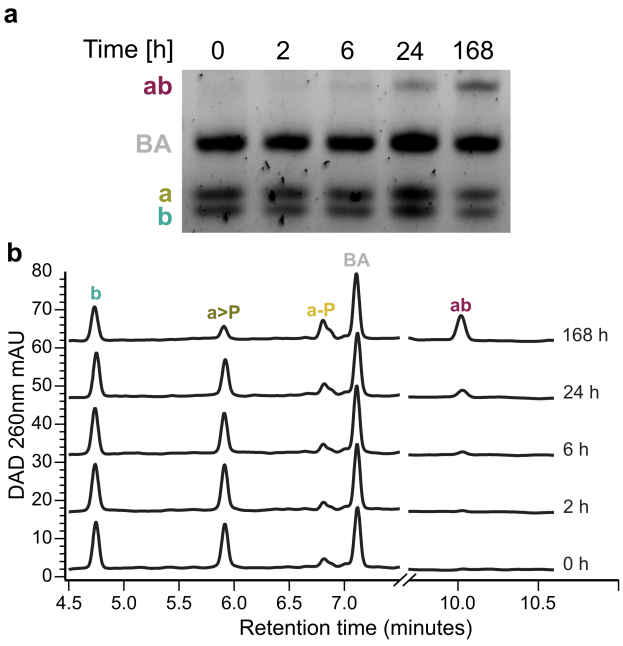


Figure 3.4: Methods to follow the ligation reaction over time **a**, Denaturing PAGE analysis of ligation reaction over time. Reaction contained 1 µM primers, 1 µM template, 50 mM CHES, pH 10 and 1 mM $MgCl_2$, at 5°C. **b**, Stacked HPLC chromatograms (absorbance at 260 nm) of the same reaction mixtures as in **a**. The product peak *ab* increases over time, as the primers get depleted.

In order to detect and quantify the ligation product formation over time, both denaturing PAGE and HPLC are used in parallel. Figure 3.4 shows an exemplary analysis for a ligation reaction at pH 10 and 5°C using these methods. The conditions and methodology for PAGE are described in Section 3.5.4. In panel **a** the time points over a week long reaction show that this method is able to differentiate between all the intervening strands, including the two primers *a* and *b*, which have the same length (11 nt). This is likely due to the different electrophoretic properties caused by the additional phosphate in primer *a*. The formation of product surpasses the limit of detection at about 6 h.

HPLC separation coupled with detection through absorbance at 260 nm is additionally able to differentiate between *a>P* and *a-P*, which allows to understand the kinetics of >P hydrolysis as it competes with the formation of product *ab*. Corresponding chromatograms are shown in Figure 3.4 **b**. However, the limit of detection of both methods is comparable. The details of HPLC detection and quantification are explained in Section 1.5.5 and 3.5.8. The quantification of the strands concentration will be mostly perfomed with HPLC UV absorbance as it introduces less artifacts when compared with SYBR Gold fluorescence.

Previous work with cyclic phosphates, both from the Braun lab on dry-state oligomerization [27] and other studies on ligation, has shown ligation performs in alkaline conditions [76–78]. Indeed a broad pH screen (pH 3-11) has shown ligation at mildly alkaline pH (Figure 3.5 **a**). For this an equimolar solution of all strands (1 µM *a*, *b* and *BA*, unbuffered) was incubated for a week at 25°C. The pH was adjusted with either NaOH (for pH 9 and 11) or HCl (for pH 3 and 5) and the reaction supplemented with 1 mM $MgCl_2$. The product *ab* is visible for pH 7 and 9, whereas for pH 11 there is almost complete hydrolysis of the backbone.

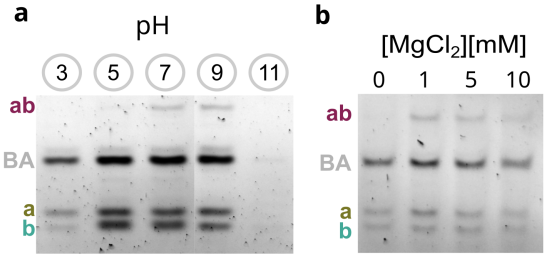


Figure 3.5: Dependence of ligation reaction on pH and $MgCl_2$ concentration. a, PAGE gel showing the pH screen for the ligation reaction. The reaction was done with 1 µM strands at 25°C for 7 days in the presence of 1 mM $MgCl_2$. The pH of the reactions was adjusted using NaOH (for pH 9 and 11) or HCl (for pH 3 and 5) and no buffer was added. The product *ab* is only visible at pH 7 and 9. Hydrolysis of all the strands is evident at pH 11. **b** PAGE gel showing the Mg^{2+} concentration screening. The Mg^{2+} was added in the form of $MgCl_2$. The reaction was done at 25°C for 7 days at pH 9 buffered with 50 mM Bis-Tris propane buffer. The product *ab* is visible only in the presence of Mg^{2+}, Addition of 1 mM $MgCl_2$ shows the most product as additional Mg^{2+} leads to hydrolysis of the strands as evident from the low intensity of the bands.

Divalent ions have been shown to be necessary for ligation with >P in previous work from Lutay et. al [77] and for other functional RNA motifs such as self-splicing group I and II introns, ribonuclease P, ribozymes etc. [105]. The formation of the phosphodiester bond, with or without ribozymatic catalysis mechanisms, forms through an S_N2 attack that has electrostatic and sterical requirements, depending strongly on the metal ion for the coordination of the intervening groups. The yield obtained with the divalent metal ions in the study by Lutay et.

al increased with increased pH, similar to what is seen for our system in Figure 3.4 **a**, and at mildly alkaline pH (8.8) was highest for Mg^{2+}. The high pH likely plays a role through deprotonation of the 5′-OH, making it a better nucleophile. Several different factors may affect which metal ions are better at catalysing a certain phosphodiester bond forming reaction, such as: reaction pH, metal ion pKa, metal ion hydroxide solubility, pka of 5′-OH group (which is dependent on sequence and microenvironment conditions [130]).

In this study we explored whether the presence of Mg^{2+} was necessary for ligation with >P and how it evolved with concentration 3.4 **b** as it seems to be ubiquitous for RNA catalysis and folding. With the same equimolar mixtures (1 µM of *a, b* and *BA*) now buffered at pH 9 with 50 mM Bis-Tris propane buffer and reacted at 25°C for several days, the influence of varying millimolar concentrations of Mg^{2+} was tested. The reaction was found to not proceed in the absence of Mg^{2+}, indicating the necessity of this divalent cation for catalysis. However, increasing the concentration of Mg^{2+} also did not lead to increased product formation, which also matches the observations by Lutay et. al [77]. This is likely due to the degradation of the phosphodiester backbone with is for RNA catalyzed at high Mg^{2+} concentrations [105]. All the reactions from here onwards were performed in the presence of 1 mM Mg^{2+} as it was found to be optimal.

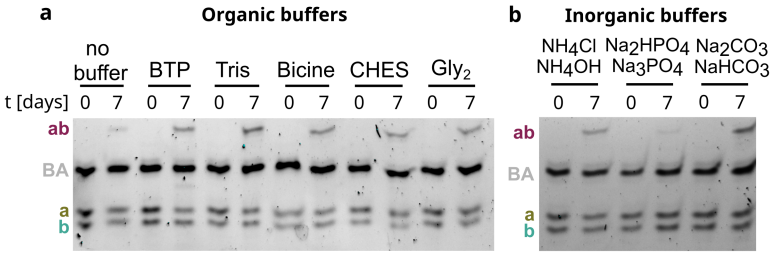


Figure 3.6: Buffer screening with organic and inorganic buffers. PAGE gels showing the ligation reaction done at pH 9 with different organic (**a**) and inorganic (**b**) buffers. The pH for each reaction conditions was set by 50 mM of the corresponding buffer. For each buffer tested, the reaction was done for 7 days at 25°C and the lane marked with 0 days denotes the control sample frozen at -80 °C. The ligation reaction proceeds irrespective of the buffer used, albeit with differing efficiencies. Phosphate buffer, especially, shows a reduction in the yield, possibly due to the high amount of multivalent phosphate ions, chelating the Mg^{2+} ions supporting the reaction. Additionally, the reaction also proceeds unbuffered, only adjusted with NaOH, with lower yield possibly due to a drop in pH upon ligation. This was also shown in Figure 3.5 at different pHs. The reaction was conducted in the presence of 1µM strands and 1mM $MgCl_2$.

A reaction pH of 9 corresponds to an OH^- concentration of 10 µM. Considering that the concentration of the RNA strands used in the reactions is in the same order of magnitude, RNA hydrolysis (of the cyclic phosphate and the phosphodiester bonds) would greatly affect the $OH^{\breve{}}$ ion concentrations thus rapidly changing the pH of the reaction. In the non-buffered pH 9 reaction (Figure 3.5 **a**) the initial pH 9.07 dropped to pH 7.72 after 7 days. As there was no significant hydrolysis of the backbone, this drop in pH is attributed to the opening of the cyclic phosphate. Furthermore, dissolution of CO_2 from ambient air could also have a partial effect on the drop in pH. Thus, to have a constant pH, the reaction was done in buffered solutions. However, in such cases, the interaction of the buffer molecules also must be considered.

To test this potential buffer interaction the ligation reaction was performed with the same equimolar concentration of RNA strands, 1 mM $MgCl_2$ and 50 mM of both organic and inorganic buffers, Figure 3.6 **a** and **b**. The reaction was conducted at pH 9 and 25°C for 7 d. We observed that the ligation reaction at pH 9 proceeds irrespective of the buffer used, albeit with slightly different efficiencies. It is worth noting that in the case of phosphate buffer, the ligation reaction is quite suppressed. The multivalent phosphate ions can chelate $MgCl_2$ ions may have thereby reduced the $MgCl_2$ ions available for the ligation reaction. In general, the yield is higher for the case of the buffered samples in comparison to non-buffered sample control. This is probably due to the aforementioned pH drop.

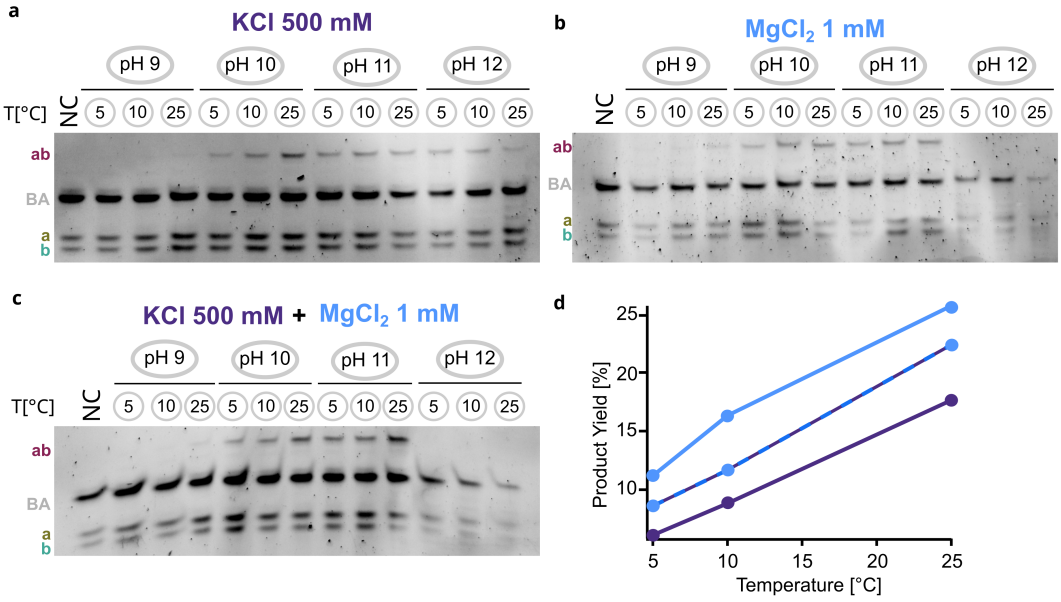


Figure 3.7: Role of Divalent (Mg^{2+}) vs Monovalent Cations (K^+) over a range of alkaline pH and temperature for templated ligation. PAGE gels comparing the ligation at different temperatures (5°C, 10°C and 25°C) and pH 9-12 in the presence of Mg^{2+} ions, K^+ ions or both (**a**, **b** and **c**, respectively). **d**, For pH 10, the ligation product yield was quantified via HPLC, at different temperatures. Mg^{2+} ions are important for the ligation reaction. Replacing Mg^{2+} with 500mM K^+ still results in 2 fold reduced ligation efficiency. However, a mixture of 1mM Mg^{2+} and 500mM K^+ shows intermediate efficiency. The reaction was done at with 50mM CHES buffer at pH 10 for 1 day.

Work with ribozymes has shown that monovalent cations in molar quantities can replace activity of Mg^{2+} [25, 100, 105]. For this reason, replacement of Mg^{2+} for K^+ but in much higher concentration (500 mM) was tested, coupled to a wider alkaline pH screening (9 to 12) and a wider temperature screening (5 to 25°C), shown in Figure 3.7 **a**, **b** and **c**. While pH 11 has shown degradation in Figure 3.5, it was hypothesized that lower temperatures and the absence of Mg^{2+} could prevent it. Supplanting the optimal low amount (1 mM) of Mg^{2+}, with 500 mM K^+ only recovers it to half the yield while a mixture of 500 mM K^+ and 1 mM Mg^{2+} rescues the reaction further, albeit moderately. Furthermore, the yield of the ligation increases linearly with temperature at pH 10 as evident from the plot in Figure 3.7 **d**. Generally higher yields are obtained to pH 10 and 11, which here still show minimal backbone hydrolysis as the reaction was stopped after 24 h (in contrast to the 7 days in Figure 3.5). At pH 12, only the sample without Mg^{2+} revealed product formation, Figure 3.7 **a**.

3.3.2 Kinetics

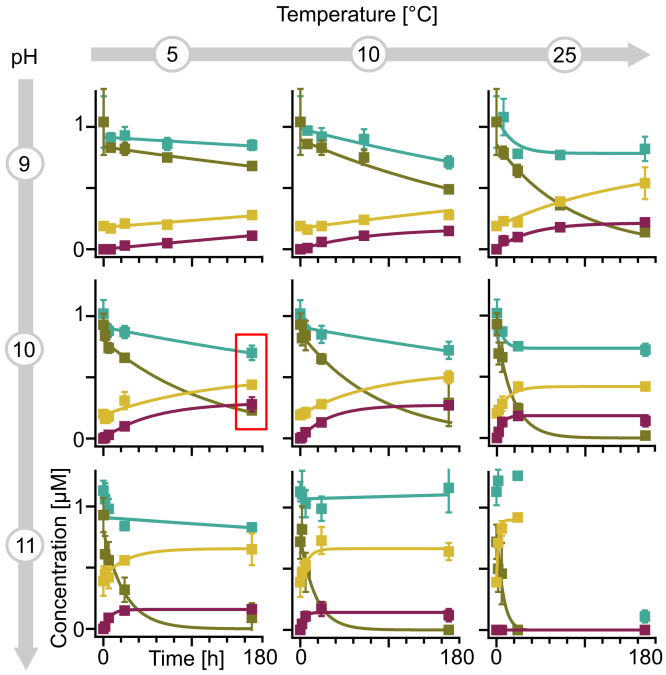


Figure 3.8: Concentration of all the strands over reaction time, for different pH and temperature conditions. Concentration of primers and product over a 7 days' period for varied pH (9-11, vertical axis) and temperature (5, 10, 25 °C, horizontal axis) measured through HPLC with UV absorbance at 260nm. Highest yield at 7 days are highlighted in red, for pH 10 and 5 °C. At both high temperature and pH (25 °C, pH 11) there is additional hydrolysis of the backbone, particularly after 7 days (bottom, right-most graph). The full lines correspond to the exponential fit of the data as a guide to the eye. Initial concentration of all strands (*a*, *b* and *BA* was 1 μM, in CHES Buffer 50 mM at the reported pH and with 1 mM $MgCl_2$). Data are represented as mean ± standard deviation of three independent replicates.

Concentration the strands *a>P*, *b*, *ab*, *a-P* was followed pver for different temperature-pH combinations and are plotted in Figure 3.8 **a** in green, blue, dark red and yellow, respectively to understand how the kinetics of both the ligation and the hydrolysis reaction develop with pH and temperature. The temperature was either 5, 15 or 25°C and the pH 9, 10 or 11, with a 1 μM concentration of each strand, 50 mM concentration of CHES Buffer, adjusted with NaOH to the corresponding pH and 1 mM $MgCl_2$. It is important to note that in all the experiments, the initial concentration of *a<P* is on average 26% of total *a*, suggesting that a part of *a>P* was already hydrolyzed in the stock solutions. This capped the maximum concentration of *ab*at 0.74 μM, the concentration of the initial *a>P*. The yield of *ab* depends on the temperature and pH combination of the reaction. The highest concentration of *ab* at 7 days (0.28

The reaction to produce *ab* proceeds in three steps: i) the annealing of primers *a<P* and *b* on the template *BA*; ii) the ligation through the nucleophilic attack; iii) the melting of product *ab* and *BA*. Each of these steps has a rate constant (k_{ann}, k_{lig} and k_{melt} as indicated in Equation 3.1.

$$a_{>P} + BA + b \xrightarrow{k_{\text{ann}}} [a_{>P} : BA : b] \xrightarrow{k_{lig}} [ab : BA] \xrightarrow{k_{\text{melt}}} ab + BA \tag{3.1}$$

The hydrolysis of the cyclic phosphate occurs in parallel with the rate constant k_{hyd} (Equation 3.2)

$$a_{>P} \xrightarrow{k_{hyd}} a_P \tag{3.2}$$

The timescales of RNA base-pairing and melting are in the order of magnitude of seconds [150]. This is a much faster timescale than the nucleophilic attack of the 5′-OH on the 2′,3′-cyclic phosphate which is in the order of hours to days. For this reason, the step from the primer:primer:template complex (C) to the product:template complex (P) is the rate determining step of the reaction in aqueous solution and it can be considered a pseudo-first order. These reactions have orders higher than 1 but behave experimentally as first order.

Thus, Equation 3.1 is simplified to Equation 3.3 .

$$C \xrightarrow{k} P \tag{3.3}$$

The reaction rate is given by Equation 3.4.

$$\frac{d[C]}{dt} = -k[C] \tag{3.4}$$

Equation 3.4 can be integrated to provide the concentration of C over time shown in Equation 3.5.

$$\int \frac{d[C]}{[C]} = \int -kdt \equiv C = C_0 e^{-kt} \tag{3.5}$$

Similarly, the product reaction rate and P concentration profile is given by Equation 3.6 and 3.7, respectively.

$$\frac{d[P]}{dt} = k[C] = k \cdot C_0 \cdot e^{-kt} \tag{3.6}$$

$$\int dP = k \cdot C_0 \int e^{-kt} dt \equiv P = P_0 + C_0 \left(1 - e^{-kt}\right) \tag{3.7}$$

Because the hydrolysis of the cyclic phosphate is also a first-order reaction, these expressions (3.4) and (3.6) are valid for the reagents (*a*<*P* and *b*) and both the ligation product *ab* and the linear phosphate side product *a-P*. The concentration profiles for all participating molecules, except the template strand *BA*, which is not consumed, are given by Equation 3.8-4.11.

$$a_P = a_{P_0} + a_{>P_0} \left(1 - e^{-k_{a_P} t}\right) \tag{3.8}$$

$$ab = a_{>P_0} \left(1 - e^{-k_{ab} t}\right) \tag{3.9}$$

$$b = a_{>P_0} \left(e^{-k_{ab} t}\right) + (b_0 - a_{>P_0}) \tag{3.10}$$

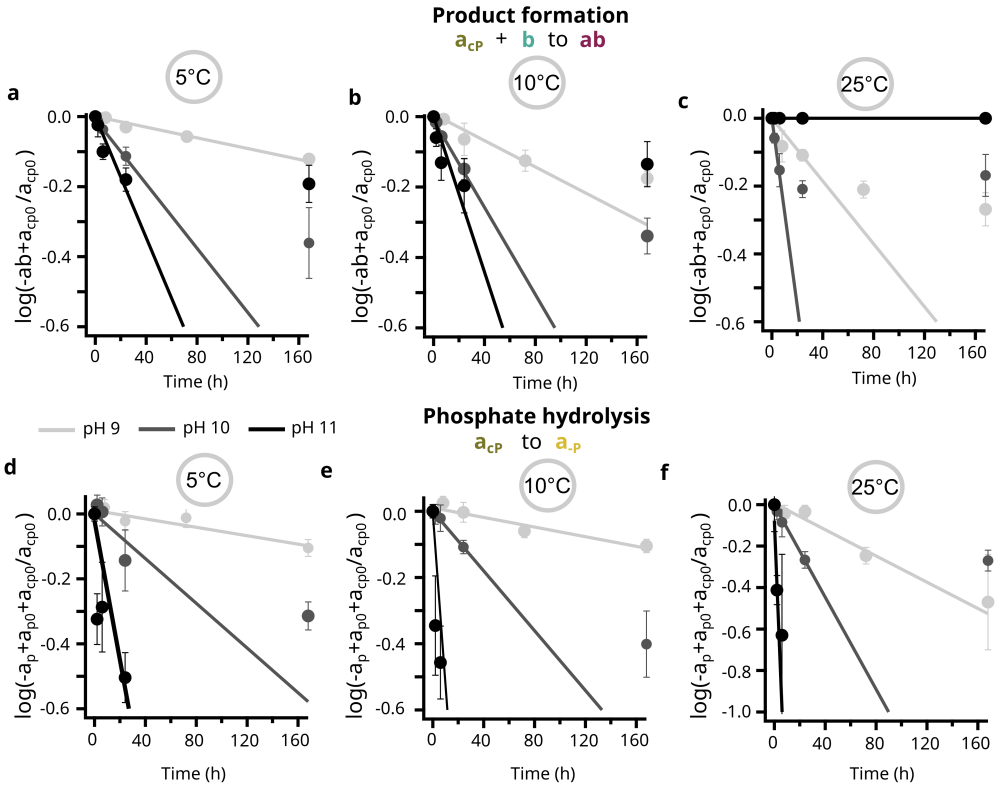


Figure 3.9: Linear regression fits to obtain the observed fitting constants for product formation and cyclic phosphate hydrolysis. For all the conditions presented in Figure 3.8 the linearized rate equations (Equation 3.12 and 3.13) were applied to the independent triplicate average and the corresponding propagated standard deviation was computed. The linear fits were performed with Igor 6.27 using Levenberg-Marquardt fitting algorithm. Only the time points corresponding to the initial rate (before saturation) were used for the fit computation, however all time-points are plotted in the figure.

$$a_{>P} = a_{>P_0}\left(e^{-k_{ab}t}\right) \tag{3.11}$$

Some considerations need to be addressed regarding Equations 3.8 - 3.11. Firstly, the k_{a_P} considered in equation (3.8) is the rate in the presence of all the other strands. The presence of the template means that, at the low temperatures below T_m, at which we are working, most of a<P will be bound to the template, and thus affecting the cyclic phosphate hydrolysis rate. Secondly, it is assumed that the complex concentration is equal to a<P. This is based on the presumptions that:

- a<P is the limiting reagent (as part of primer a<P is already hydrolyzed, a priori)
- All the primers are bound to the complex at the tested temperatures. The calculated T_m of primer-template is about 40C, so for temperatures between 5C and 25C this can be assumed to be true.
- concentration of complex P (product:template) has the same concentration as ab.

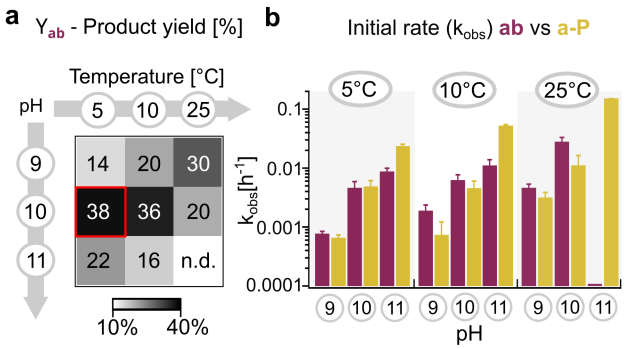


Figure 3.10: Product yield and rate constant for ligation and cyclic phosphate hydrolysis. **a**, Product yield obtained (%) at 7 days for all the conditions tested in Figure 3.8. Maximum obtained yield of 38% was for pH 10 at 5 °C (red square). This reported yield has been corrected for the limiting concentration of *a>P*. **b**, Observed initial pseudo-first order rate of product and inactive primer formation for all the conditions, obtained from the fitting of the data presented in Figure 3.9.

To obtain the rate of product formation and cyclic phosphate hydrolysis we re-arranged Equations 3.8 and 3.9 to Equations 3.12 and 3.13, respectively.

$$\log\left(\frac{a_{P_0} + a_{cP_0} - a_P}{a_{cP_0}}\right) = k_{a_P} \cdot t \tag{3.12}$$

$$\log\left(\frac{a_{cP_0} - ab}{a_{cP_0}}\right) = k_{ab} \cdot t \tag{3.13}$$

The equation 3.12 and 3.13 could then be applied to fit the data from the time point screening shown in Figure 3.8. The linear regressions for all the conditions are presented in Figure 3.9. All concentration data points correspond to mean ± one standard deviation of the three independent replicates. For Figure 3.9 the error bars correspond to the error propagation according to the linearized formula. The linear regression was performed with Igor 6.37 which uses the Levenberg-Marquardt fitting algorithm. The standard deviation calculated through error propagation, was used as weighting parameters for the fit, which allows a more accurate fit parameter error estimation. Additionally, to obtain the observed rate constants, only the points corresponding to the initial reaction rate, before saturation, were fitted, which depends on the condition (see Figure 3.8).

The yield, calculating for the real initial amount of *a>P*, is shown in Figure 3.10 **a**. For pH 10 and 5°C this corresponds to a yield of 38%. The obtained yield was measured under limiting concentration of the primer with a cyclic phosphate end (*a>P*), meaning that the formation of hydrolyzed *a-P* contributed to the observed partial conversion. However, the addition of excess primer *a* did not improve the yield (data not shown) indicating efficient product-inhibition, confirmed by our estimate that the product strands have an off-rate of about 40 days at 5°C. Comparable yields have been reported for ligation with phosphorimidazolide activation under similarly low Mg^{2+} concentrations [143].

The first order kinetic rate constants (k_{obs}) for product formation and hydrolysis of cyclic phosphate were fitted to the data in Figure 3.9 and are shown for all the samples in Figure 3.10. The obtained rates are between 0.001 and 0.03 h^{-1} which are in the same order of magnitude for the ligation of native RNA at 20mM $MgCl_2$ with 2-Me imidazolide chemistry [159]. A

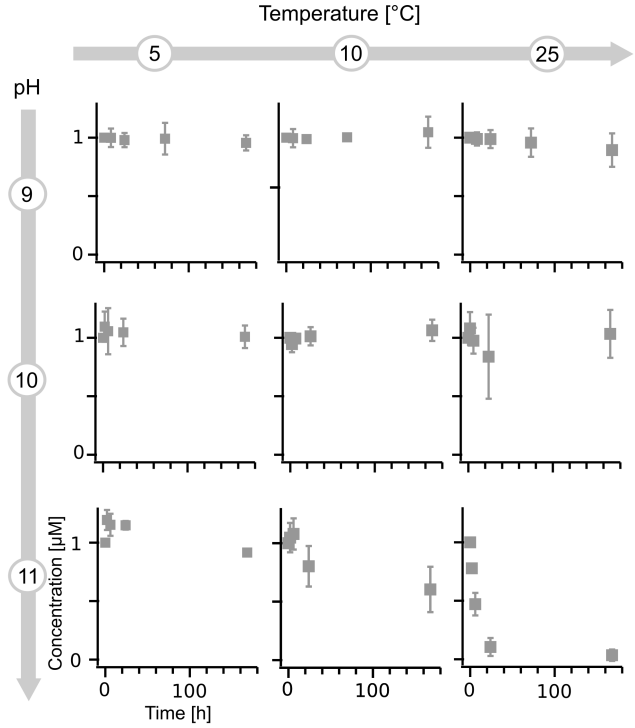


Figure 3.11: Template *BA* concentration over time at different pH and temperature. Screening the concentration template *BA* over a 7 days' period for varied pH (9-11) and temperature (5, 10, 25°C), for the same samples presented in 3.8 of the main. The concentrations were measured with HPLC UV detection 260 nm. Data are represented as mean ± standard deviation of three independent replicates.

few salient features of the plots in Figure 3.10 are the rates and yields of ligation (*ab*) and hydrolysis (*a-P*). Both the rates of ligation and hydrolysis increase with an increase in either pH or temperature, keeping the other parameter constant. Figure 3.10 **b** shows that while the observed ligation rate mostly increases from pH 9 to 10, the inactivation rate increases from pH 10 to 11. At higher pH, the ligation kinetics is slightly faster, however the final yield drops due to the competing inactivation rate. The rate of ligation is in general higher than the rate of the hydrolysis at 5 and 10°C while 25°C favor the inactivation. It has been reported that low temperatures reduce the entropic cost of the ligation reaction and shift the reaction equilibrium from hydrolysis to ligation [123]. The maximum 7 day yield obtained is at pH 10 and 5°C. Significant RNA backbone hydrolysis is observed when both temperature and pH are maximal (25°C and pH 11).

In order to test whether the template normalization used for Figure 3.8 is appropriate, as possible backbone hydrolysis would increase the adjusted concentration and introduce artifacts in the obtained yield, the average concentration of *BA* was plotted for each of the conditions in Figure 3.11. The template concentration varies slightly across different samples. For pH 11 at 25°C rapid hydrolysis of the template is observed, whereas for pH 11 at 10 °C, this is only significant for later time points. The remaining samples do not reveal hydrolysis. For the former since there is not product detected at 7 d this does not affect the yield reported. For the latter, the concentration at 7 d (the only time point with visible degradation) is similar

to that of 3 d indicating the normalization does not introduce a considerable shift. This is likely due to the fact that not only the template *BA* degrades but also the remaining strands in solution in similar rates. The shift introduced due to backbone hydrolysis for *ab* and *BA* is therefore similar not affecting much the normalized concentration and yield.

3.3.3 Loop-closing ligation

Functional RNAs possess folded secondary structures, which are essential for the molecular evolution of early oligomers into functional RNA, such as ribozymes [9, 16, 34, 40]. However, there exists an incompatibility between the folded structures involved in catalysis and the unstructured RNA that serves as a template in copying chemistry [147]. Therefore, a mechanism introducing non-copied regions, like loops and bulges, is considered a crucial step for nucleic acid catalysis to emerge, not only diversifying information but also, more importantly, introducing structure. To address this, a loop-closing reaction—where the template folds back, attacking the cyclic phosphate of the primer attached to it—was tested. This reaction yields a product with a hairpin as a secondary structure when using primer *a<P* and template *BA*, as depicted in the schematics shown in Figure 3.12.

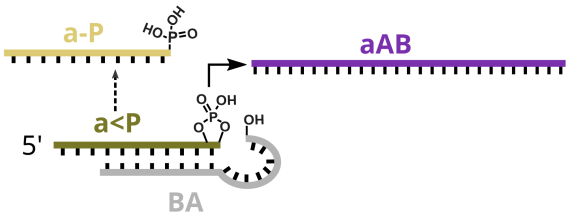


Figure 3.12: Scheme of loop-closing reaction, in the absence of primer *b*. Loop-closing reaction occurs when the 5′-OH group of the template *BA* attacks the cyclic phosphate on primer *a* forming a product herein named *aAB*. The cyclic phosphate can also hydrolyse in a competing side-reaction yielding the inactive primer *a-P*.

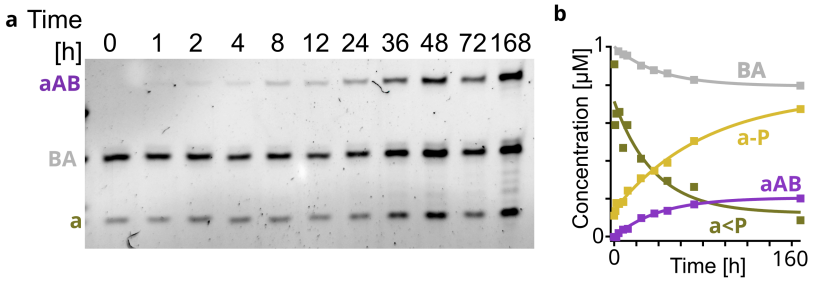


Figure 3.13: Loop-closing reaction product *aAB* formation kinetics. **a**, PAGE gel showing the loop-closing ligation reaction done over the course of one week, at pH 9 (buffered with Bis-Tris propane adjusted with NaOH) and 25° C. **b**, Concentration of all intervening strands measured with HPLC UV detection 260 nm. Reaction contained 1 μM primers, 1 μM template, 50 mM Bis-Tris propane and 1 mM $MgCl_2$.

To test the feasibility of loop-closing ligation, an equimolar mixture of the RNA strands was prepared (1 μM of *a<P* and *BA*), with 1mM $MgCl_2$ at pH 9 and 25°C and followed the reaction over the course of 7 days both through PAGE and HPLC with the data shown in Figure 3.13 **a**

and **b** respectively. The loop-closing product *aAB* is detectable after about 2 h and reaches a plateau around 0.25 µM after about 3 d. Both the kinetics and yield of this reaction are comparable to the ligation of primer *a<P* and *b* on the template discussed in Section 3.3.2. This also supports the idea that the ligation step is the limiting rate-determining step. Annealing should be fast enough such that one-strand on a template vs. two-strand on a template does not significantly affect the ligation rate.

3.3.4 Reverse system

To complete a full replication cycle, it is imperative that both strand separation occurs and that the ligated product can serve as a template to ligate two shorter segments, thereby producing an exact copy of the original template (as the product of the initial step is merely the reverse complement). To assess the feasibility of the second step in this process, a 'reverse system' was devised. Here, the product sequence *ab* now acts as a template, with primers *A>P* and *B* ligating to form *BA*, effectively replenishing the initial template, Figure 3.14 **a**. Additionally, to distinguish the strand ab_T from *ab*, it bears **AAA** overhangs both at the 3′ and 5′ ends. Similarly, the primers *B* and *A>P* feature **AAAAA** sequences to differentiate them from *a>P* and *b* in a gel. This also ensures that the product from the ligation, BA_p, is distinguishable from *BA*.

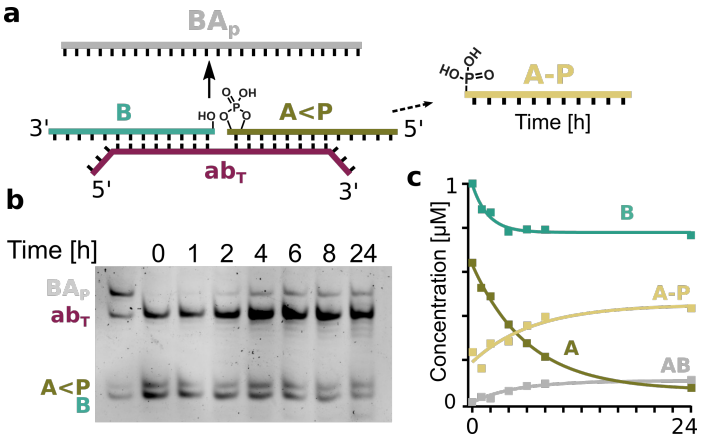


Figure 3.14: Reverse ligation that produces the template *BA*. **a**, Scheme of reverse ligation reaction, where primers *A* and *B* bind to the product *ba*, here acting as a template, in order to yield *BA*, the template of the forward reaction shown in Figure 3.3. **b**, PAGE gel showing the reverse ligation reaction done over the course of one day at pH 10 and 5°C. **b**, Concentration of all intervening strands measured with HPLC UV detection 260nm. Reaction contained 1µM primers, 1µM template, 50mM CHES buffer pH 10 and 1mM $MgCl_2$.

In order to investigate ligation with the reverse system, an equimolar mixture of all the intervening RNA strands was prepared (1 µM of *A*, *B* and ab_T, at pH 10 (50 mM CHES buffer) and 5°C with 1ũnit $MgCl_2$ and the reaction was follwed over the course of 24 h both through PAGE and HPLC with the data shown in Figure 3.13 **b** and **c** respectively. The ligation kinetics and yield are similar to those of the 'forward system' in 3.8 and the loop-closing system in 3.13. The implementation of a full cycle may still pose some challenges. Firstly, strand separation has to be compatible in these conditions. Even though the $MgCl_2$ is low, the temperatures are also cold, still potentially above T_m. Non-equilibrium systems with air-water interfaces

have shown however that Mg_2+ concentration in this same order of magnitude is compatible with both strand separation and ribozymatic activity [114]. Additionally, continuous feeding of primers should occur concomitantly to the strand separation such that the cycles can proceed. An air-water interface with continuous feeding would allow for both the denaturation and replenishment of activated primers, without necessarily high temperature which could degrade the cyclic phosphate and the RNA backbone.

The type of linkages formed on the product strand may also affect the viability of the full replication cycle. On the one hand, a 2′-5′ linkage affects the structure of the RNA strand [121] which may affect sterical orientation of the 5′-OH and >P in the attached primers. It has however been shown that the structural effect can be compensated for on the global structure, in case the relative amount of 2′-5′ linkages would be small, which is the case here. On the other hand, the destabilising properties of 2′-5′ linkages could also play a role in facilitating strand separation [38, 46, 65, 144]. To test the regiospecificity of the reaction, a study on the type of phosphodiester linkage formed through ligation was performed and is presented in Section 3.3.5.

3.3.5 Phosphodiester linkage

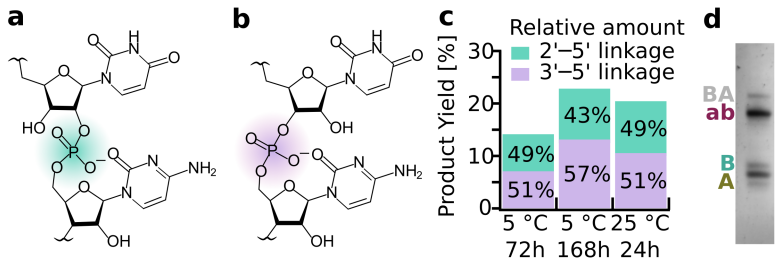


Figure 3.15: Linkage analysis of the reaction product *ab* through digestion with Nuclease P1. Scheme of the ligation site with a 2′–5′ linkage (**a**, blue) and a 3′–5′ linkage (**b**, purple). **c**, Total product yield obtained for three condition sets, with the corresponding relative amount of 2′–5′ and 3′–5′ linkage. The ligation with 2′,3′-cyclic phosphates does not exhibit regioselectivity, as both linkages are equally represented for the studied conditions. After the reactions with 10 µM primers, 10 µM template, 50 mM CHES pH 10 and 1 mM $MgCl_2$, they were digested with Nuclease P1 (Section 3.5.6). The concentrations of the samples before and after digestion were measured with HPLC UV detection at 260 nm, Section 3.5.6. Data represents the mean of independent duplicates. **d**, PAGE gel showing the reverse system ligation reaction, where the *ab* template strand has a 2′–5′ linkage. The product *BA* still forms showing that a template containing a 2′–5′ linkage does not hinder ligation. Reaction contained 1 µM primers, 1 µM template, 50 mM CHES buffer pH 10 and 1 mM $MgCl_2$.

The attack of the 5′-OH on the cyclic phosphate (Figure 3.3 **b**) can form either a 2′–5′ or a 3′–5′ phosphodiester bond (Figure 3.15 **a** and **b** respectively). This happens due to the similar nucleophilicity of both the 2′ and 3′-OH groups of the ribose [121]. Previous studies on the polymerization and ligation with cyclic phosphates have reported varying ratios of 3′–5′ to 2′–5′ linkages, depending largely on the experimental conditions. For example, dry state polymerization resulted in a natural linkage enrichment ratio of 2:1 [136,139], while an aqueous state (with 0.5 M diamine, pH 8 and 0°C) was reported to lead to at least 97% of 2′–5′ [107]. Templated cleavage and ligation at 25°C, pH 9 and 5 mM $MgCl_2$ in an aqueous solution was

reported to also show a predominance of 2′–5′ linkages (about 95%) [78]. Conversely, templated ligation in the eutectic phase resulted in an excess of 3′–5′ linkages [141]. For templated ligation reaction described here, no significant regioselectivity under the tested conditions was found (Figure 3.15 **c**). This difference in comparison to previous studies is potentially due to different systems and conditions tested. To investigate this, the reaction was quenched by ethanol precipitation and the samples were digested with Nuclease P1 following the manufacturer's protocol (3.3.5). Nuclease P1 specifically lyses the 3′–5′ linkages, which in this case would digest all the strands *a*, *b* and *BA* completely but digest *ab* either completely, or result in a **UC** dimer.

The concentration of total product pre-digestion and **UC** dimer post-digestion were determined using HPLC UV absorbance (3.3.5). Both types of linkages were formed equally (Figure 3.15 **c**) indicating that the reaction is not regioselective. However, a slight enrichment of the canonical linkage over time for the 5°C conditions can be seen, possibly due to the favored hydrolysis of 2′–5′ linkages, particularly in double-stranded RNA in alkaline solutions [110]. The presence of non-canonical linkages however, does not render the product strands obsolete. Such mixed backbone RNA have been demonstrated to still fold into functional structures [121]. The stability of the RNA duplex has been documented to be reduced for strands fully composed by 2′–5′ linkages in comparison for RNA with canonical linkages [46, 144]. For this system however, one single 2′–5′ linkage at the ligation site would likely not have a considerable destabilizing effect, as the duplex could accommodate for the structural disruption [121]. Furthermore, a reaction with *ab* containing a 2′–5′ linkage at the ligation site was performed (Figure 3.15 **d**) with equimolar amounts of primers *A*, *B* and template *ab* (1 μM)) in 50 mM CHES pH 10 and 1mM $MgCl_2$. It was shown that *ab* can still template the formation of *BA* through >P mediated ligation, highlighting the possibility of a replication cycle.

3.3.6 Sequence dependence

The sequence dependence of ligation was investigated through NGS sequencing, focusing on the effect of randomizing three nucleotides closer to the ligation site on primer *a*, while keeping primer *b* and template *ab* constant. The experiment with the randomized primer *a* conducted at pH 9 with equimolar amounts of primers and supplemented with $MgCl_2$, the reaction proceeded for 24 h at 25°C. With a fixed template sequence (5′ **ACU** 3′), the study anticipated an enrichment of the reverse complement at positions crucial for ligation, indicating a preference for specific nucleotides (**U** in N_a, **G** in N_c, and **A** in N_e) in the ligated product. Both the initial mixture (t_0) and the ligated mixture underwent sequencing, followed by quality trimming and RegEx filtering to isolate reads containing the exact sequence of non-random stretches in the primers or the template, Section 3.5.9.

Analysis of the relative abundance of nucleotides at t_0 for the randomized primer in positions N_a, N_c, and N_e revealed an initial synthesis bias. Notably, **G** and **U** were enriched by over 25%, while **A** was depleted at the ligation site, serving as a reference bias for comparison with the final bias observed. This provides an idea about the initial synthesis bias of the randomized stretch. For instance, **G** and **U** are enriched (>25%) in the position at the ligation site, whereas **A** is depleted. This serves as the reference bias to compare the final bias to. In Figure 3.16 **c** the bias observed in the three initially randomized position nucleotides of the ligated product (incorporation bias) shifted. The relative enrichment of the product *ab* through ligation, in comparison to the reference, primer *a* at t_0, is shown in Figure 3.16 **d**.

The closer the position is to the ligation site (N_a>N_b>N_e), the more enriched it is, relatively, in the reverse complementary nucleotide. While N_e did not show a relative enrichment to **A**, N_c has a 15% enrichment towards **G** and N_e has an enrichment of almost 40% towards **U**. This result highlights the importance of primer-template binding for successful ligation, particularly for the nucleotides at the ligation site.

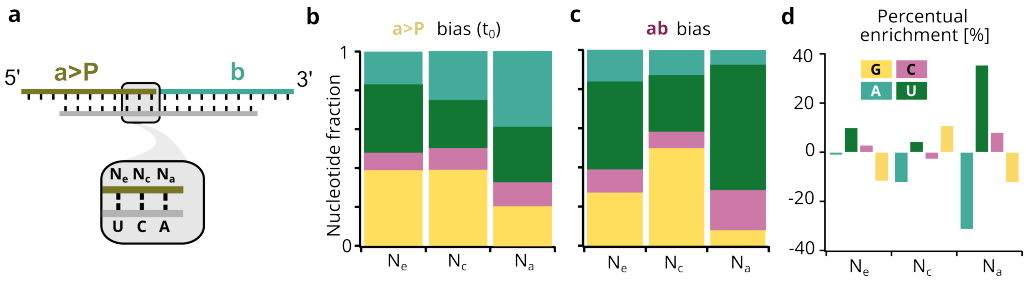


Figure 3.16: Sequence dependence in the segment of primer *a* close to the ligation site assessed by NGS. a Schematics showing the positions of primer *a* that were randomized. **b** Bias of randomized positions N_a, N_c and N_e of primer *a* assessed by NGS (Section 3.5.9) corresponding to the RNA synthesis bias. **c** Bias of the same positions in the ligated product *ab* corresponding to the incorporation bias. **d** Percentual enrichment in the same positions, normalizing the bias of the product *ab* in **c** with the initial synthesis bias of the primer *a* in **b**.

In order to investigate in detail how the ligation site nucleotides affect ligation, both the intervening nucleotides on primer *a* (N_a) and *b* (N_a) were varied. The reaction was conducted with each of the 16 different nucleotide combinations (four each on 3′ end of **a** and 5′ end of **b**) at the ligation site, while keeping the template sequence fixed (Figure 3.17 **a**). Additionally, two different templates were tested, one with **GA** and the other with **UA** at the position complementary to ligation site. Except for the ligation site, the remainder of the sequence was fully complementary to the template. These reactions were carried out at pH 10 and 5°C for 7 d. Figure 3.17 **b** and **c** show that of all the combination of the primers tested, the highest yield of *ab* is obtained for the sequence with the correct nucleotides at the ligation site (marked in red, **CU** for the template **GA** and **AU** for the template **UA**). However, mismatched ligations did occur, albeit with much lower relative yields, which were especially reduced for the template **UA**.

Interestingly, **G** : **G** wobble pairing of the 5′ **U** of primer *b* at the ligation site led to a high relative yield (78%) compared to the complementary primers for the template **GA**. When considering one single mutation at the ligation site either on *a* or *b*, the reaction yield drops on average by 91% or 82% respectively, relative to the non-mutated complex. This is considered to be the ligation fidelity for one mutation. If two mutations at the ligation site are considered (on both the 5′ and 3′ nucleotides), the average experimental yield drops to 12% (template **AG**) and 5% (template **AU**) and thus the ligation fidelity for the respective template is 88% and 95%.

3.3.7 Per-nucleotide fidelity

A prebiotic replication through ligation would take place from a diverse pool of oligonucleotides consisting of different sequences of varied lengths. In such a scenario, the likelihood of unstable primer template complexes is high. Two contributions of the nucleotides

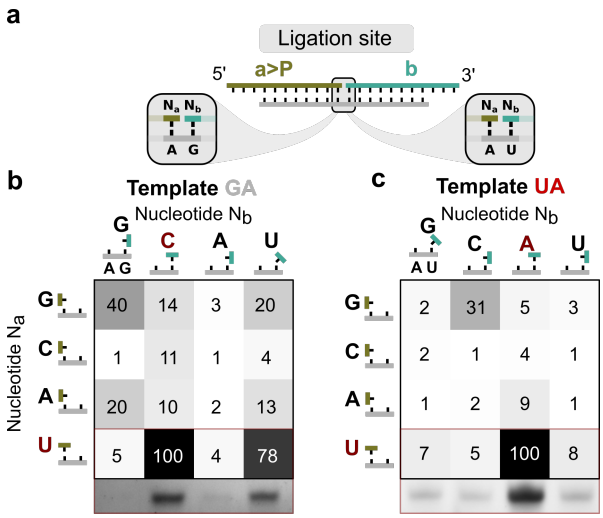


Figure 3.17: Ligation site specificity for two different templates sequences. Reaction yields at 7 days were quantified for reactions with primers containing each of the possible four nucleotides the 3′ end of *a* (N_a) and 5′ end of *b* (N_b) leading to 16 primer combinations. **a**, Schematics of the ligation sites and the two templates tested. The templates differed only at the dimer complementary to the ligation site, with either 5′ **GA** 3′ (**b**) or 5′ **UA** 3′ (**c**). The maximum yield obtained in both cases was for the correct combination of complementary primers (nucleotides highlighted in red). **G** : **U** wobble pairing is represented as a tilted nucleotide in the cartoon representation. For most combinations, one single mutation at the ligation site reduced its relative yield considerably or prevented ligation, even though the second primer is fully bound to the template. The snippet below the heat map in (**b** and **c**) corresponds to the PAGE of the bottom row, showing the ligated product *ab*. Reactions were performed with 10 µM primers, 10 µM template, 50 mM CHES pH 10 and 1 mM $MgCl_2$ for 7 days at 5°C. Data are represented as mean of independent triplicates.

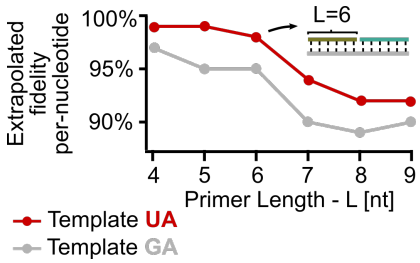


Figure 3.18: Per-nucleotide fidelity over primer length. The fidelity of ligation was extrapolated to a per-nucleotide replication fidelity using primers of varying length using single-mutation sensitive binding calculation of the primers with NUPACK (Subsection NUPACK Analysis). The fidelity dropped for longer primers. Also, a **G** at the ligation site lead to lower fidelity due to the **G** - **U** wobble pairs (Section 3.3.6).

surrounding the ligation site can be identified to have a significant effect on the stability of the complex, the amount of mismatches and the length of the binding region. To compare the performance of the ligation with a base-by-base replication, the per-nucleotide replication fidelity was calculated. This corresponds to the minimum fidelity that a base-by-base replicator would require, for each incorporation, to create the same number of errors within the ligated strand.

Figure 3.17 **b** and **c** suggests significant reduction in the ligation yield (80% or 94% on average) when the ligation site nucleotides are mutated. This in turn suggests that the positioning of the two nucleotides at the ligation site is crucial for the attack of the 5′-OH on the >P. Nevertheless, the role of the other nucleotides of the primer sequence cannot be disregarded, as the overall stability of the primer-template complex is also important to the templated ligation. Two contributions regarding the surrounding nucleotides can be identified to have a significant effect on the stability of the complex.

The first contribution regards the sequence mismatches between the primers and the template. Here, the destabilizing, and hence ligation-inhibiting effect scales with the number of mismatches and their proximity to the ligation site (not including the site itself). The second contribution regards the overall length (L) of the base paired region given no mismatches. At a given temperature, the stability of the RNA duplex would scale with the length. Summarized, the ligation reaction depends on both the probability of the duplex formation combined with the probability that the ligation site nucleotides are bound and in proper orientation. While the role of the ligation site nucleotides has been determined experimentally (Figure 3.17 **b** and **c**), the role of the complex formation and its dependence on the introduction of mismatches to the primer-template complex was determined by binding analysis on NUPACK [154]. The contribution of point mutations to duplex stability has been extensively studied in the context of PCR primer design [45, 74, 125].

NUPACK Analysis

The binding analysis of the ligation complex was performed using the NUPACK Python module [45, 154], which was accessed through a self-written LabView interface. Primary inputs for the analysis were the sequences of the three strands, the model parameter, and constraints for the introduction of mismatches to the primer-template complex, which we call mutations in the following text. For the complex analysis, only the nucleotides surrounding the ligation site, but not the two ligating nucleotides themselves were mutated to estimate the error rate of the ligation with respect to the overall sequence of the primers. Then both were combined to calculate the overall fidelity.

First, the equilibrium concentration of the primer-template complex without any mutation in the sequences is calculated (C_0). For the case of one mutation, each nucleotide is varied consecutively and equilibrium complex concentration (C_1) is calculated for each case. Then, the equilibrium concentrations of the structures in which the nucleotides involved in the ligation process are bound to the template are calculated ($C_{1,b}$). The binding error rate ($\varepsilon_{\text{duplex},1}$) is calculated as the ratio of the average of $C_{1,b}$ to C_0, as shown in Equation 3.14.

$$\varepsilon_{\text{duplex},1} = \frac{\left(\sum C_{1,b}\right)/N}{C_0} \tag{3.14}$$

where N is the number of possible mutants.

Identical analyses were done for two mutations (inclusive of all possible variants of the two mutations) and for the case of shorter strand lengths by successively removing one nucleotide from each side, furthest away from the binding site.

Fidelity calculation

The per-nucleotide replication fidelity (f) is defined as the probability of correct nucleotide incorporation by a base-by-base replicator. This corresponds to the inverse of the error rate (ε) associated with each nucleotide incorporation and is given by Equation 3.15.

$$f = 1 - \varepsilon \tag{3.15}$$

For the described replicative system based on templated ligation, mutation of one or more bases reduces the product yield or completely impairs reaction (Figure 3.17). This residual yield obtained upon mutation corresponds to a ligation error rate which when combined with the error rates in the complex formation, can be used to calculate the per-nucleotide fidelity i.e., the fidelity of a base-by-base replicator that would produce the same number of errors within the product strand.

Taking a symmetric system with two primers of length L bound to a template of length $2\,\mathrm{L}$, associated with a ligation error of $\varepsilon_{\text{ligation}}$ then the corresponding single-base f would be given by Equation 3.16.

$$f = \sqrt[2L]{1 - \varepsilon_{\text{ligation}}} \tag{3.16}$$

Mutating different nucleotides at different positions of the primers will have a different impact on the $\varepsilon_{\text{ligation}}$: It can be expected that mutations further away from the ligation site will mostly affect the duplex stability, while mutations at the ligation site will affect mostly the viability of the nucleophilic attack. In order to quantify this effect, we define the error rate $\varepsilon_{\text{site}}$ as the probability for a successful ligation despite a non matching nucleotide-pair at the ligation site, and the error rate $\varepsilon_{\text{duplex}}$ as the corresponding probability for successful ligation despite a mutation in the remaining positions within the primer. The overall average ligation error rate for one allowed mutation $\varepsilon_{\text{ligation},1}$ is then given by Equation 3.17.

$$\varepsilon_{\text{ligation},1} = \frac{3 \cdot \begin{pmatrix} 2L-2 \\ 1 \end{pmatrix} \varepsilon_{\text{duplex},1} + 3 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} \varepsilon_{\text{site},1}}{3 \cdot \begin{pmatrix} 2L \\ 1 \end{pmatrix}} \tag{3.17}$$

Where $3 \cdot \begin{pmatrix} 2L-2 \\ 1 \end{pmatrix}$ corresponds to the number of possible single mutations outside of the ligation site, since there are $2\,\mathrm{L}$ nucleotides in total, 2 nucleotides at the ligation site and there are 3 possible ways to mutate each position. Following the same logic $3 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ corresponds to the number of possible single mutations in the ligation site and $3 \cdot \begin{pmatrix} 2L \\ 2 \end{pmatrix}$ to the total number of possible single mutations.

Similarly, for the case of two mutations, there can be two mutations outside the ligation site, two in the ligation site or one in each. It follows that $\varepsilon_{\text{ligation},2}$ is then given by Equation 3.18.

$$\varepsilon_{\text{ligation},2} = \frac{3^2 \cdot \begin{pmatrix} 2L-2 \\ 2 \end{pmatrix} \varepsilon_{\text{duplex},2} + 3^2 \cdot \begin{pmatrix} 2 \\ 2 \end{pmatrix} \varepsilon_{\text{site},2} + 3^2 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} \varepsilon_{\text{site},1} \begin{pmatrix} 2L-2 \\ 1 \end{pmatrix} \varepsilon_{\text{duplex},1}}{3^2 \cdot \begin{pmatrix} 2L \\ 2 \end{pmatrix}} \tag{3.18}$$

For the particular system described in Figure 3.3 the $\varepsilon_{\text{site},1}$ and $\varepsilon_{\text{site}2}$ were obtained experimentally from the data in Figure 3.17 **b** and **c**, as it corresponds to the average yield obtained following one or two mutations at the ligation site, respectively. Through the NUPACK Analysis $\varepsilon_{\text{duplex},1}$ and $\varepsilon_{\text{duplex},2}$ were obtained by successively mutating the duplex regions excluding the ligation site. The ligation error rates and the corresponding per-nucleotide replication fidelity for each of the cases is summarized in Table 3.2.

Table 3.2: Average ligation error rate associated with mutations of the ligation site nucleotides ($\varepsilon_{\text{site}}$) and of the remaining nucleotides ($\varepsilon_{\text{duplex}}$) for 1 or 2 point mutations. $\varepsilon_{\text{site}}$ values were determined experimentally and correspond to the average relative yield obtained upon one mutation in the ligation site. $\varepsilon_{\text{duplex}}$ corresponds to the relative concentration of complexes with the ligation site nucleotides bound, upon one mutation not in the ligation site, in relation to in the absence of mutations and were computed with NUPACK.

	1 mutation			2 mutations		
	$\varepsilon_{\text{duplex1}}$	$\varepsilon_{\text{site},1}$	f_1	$\varepsilon_{\text{duplex},2}$	$\varepsilon_{\text{site},2}$	f_2
Template **AG**	0.93	0.20	0.89	0.85	0.12	0.93
Template **AU**	0.86	0.06	0.92	0.71	0.05	0.96

Fidelity for systems with primers of different lengths considering 1 mutation

As discussed, the length of the primers affects the stability of the duplex and single mutations on shorter primers are expected to have a more deleterious effect on the ligation. This would suggest an increased fidelity for replication of short primers.

Single mutations were performed to study how the per-nucleotide fidelity changes with length in systems composed of primers between 4 and 2 nt long. Starting with the system described in Figure 3.3 and Table 3.1 nucleotides were removed from the primer terminus farthest away from the ligation site to obtain sequences with length 4 to 7 nt. For the sequences between 9 and 12 nt, **U**, **UA**, **UAC**, **UACG** were added to the same terminus, respectively. Because the ligation site was kept constant, the experimental $\varepsilon_{\text{site},}$ was used for each of the templates. The $\varepsilon_{\text{duplex},1}$ was computed with NUPACK by performing all possible single mutations outside the ligation site. The obtained values are presented in Table 3.3.

Using Equation 3.17, the ligation error rate $(\varepsilon_{\text{ligation},1})$ for one point mutation is calculated by substituting the values of $(\varepsilon_{\text{duplex},1})$ for each length and the values of $(\varepsilon_{\text{site},1})$ from Table 3.2, depending on the template sequence. The per-nucleotide fidelity corresponding to one point mutation (f) was calculated using Equation 3.15. In general, fidelity decreases with primer length as each mutation becomes detrimental to the duplex stability.

Additionally, the template with G at the ligation site has a generally lower fidelity because both $\varepsilon_{\text{site},1}$ and $\varepsilon_{\text{duplex},1}$ are higher. This is a result of the stronger **G**:**C** binding, which

Table 3.3: Average ligation error rate associated with one mutation outside the ligation site ($\varepsilon_{\text{duplex},,1}$) for systems with primers of different lengths. Values correspond to the relative concentration of complexes with the ligation site nucleotides bound, upon one mutation, in relation to in the absence of mutations and were computed with NUPACK.

Length (L)	4	5	6	7	8	9	10	11	12
Template AG	0.37	0.22	0.48	0.88	0.93	0.94	0.95	0.95	0.96
Template AU	0.09	0.13	0.19	0.67	0.86	0.88	0.92	0.92	0.93

increases the stability of the duplex despite mutations and **G**:**U** wobble pair at the ligation site which increases the error $\varepsilon_{\text{site},1}$.

3.3.8 Shorter systems

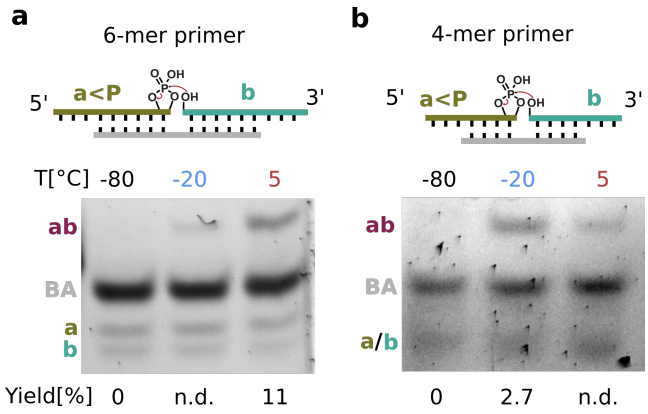


Figure 3.19: Shorter primer:primer:template systems with 6-mer or 4-mer base-paired region. Ligation product *ab* was detected for both systems through PAGE gel at both -20°C and 5°C, after reaction for 7 days with 100 µM of primers and templates each (CHES pH 10 with 10 mM $MgCl_2$). The primers for the 4-mer system are 7 nt long and therefore not well detected through PAGE analysis. Three independent replicates were analyzed with HPLC UV absorbance as described in Section 3.5.7 and 3.5.8 and the obtained yields are reported under the gel.

As most non-enzymatic oligomerization reactions yield mostly very short strands and the fidelity increases for shorter lengths as concluded in Section 3.3.7, the limit of primer length was tested by reducing the system in Figure 3.3 to 6-mer and 4-mer binding regions, removing nucleotides from the outer regions of the primers, Figure3.19 **a** and **b** respectively. This assured the **GC**-content stayed about the same as for the longer system. This length excludes the poly-adenosine overhangs added for PAGE detection. Due to the shorter binding region the duplex is less stable. Hence, the systems were assessed at lower temperatures (both 5°C and -20°C) at pH 10, 1 mM $MgCl_2$. Note that for the 4-mer primers had a length of 7 nt and thus were not discernible in the gel. However, the product was visible for both systems at all tested temperatures, except at -8 °C which served as a negative control. The yield of the reaction was quantified using HPLC UV absorbance for independent triplicates. For the 6-mer primer system, the yield was higher at 5°C than at -20°C, whereas for the 4-mer system, the

opposite was true. The ligation and dissociation rates are higher at higher temperatures, so the optimum ligation temperature differs depending on the duplex length and, therefore, its stability. Under the lowest yield conditions, the product was visible through PAGE but not detectable through UV absorbance, possibly reflecting the method's lower limit of detection. The obtained yields at 7 days were very low (about 3%) for the 4-mer primers. This suggests that to stabilize the duplex for short oligonucleotides, a compromise between the slower rate of ligation and the low probability of duplex formation must be made.

3.3.9 Ligation through concatenation

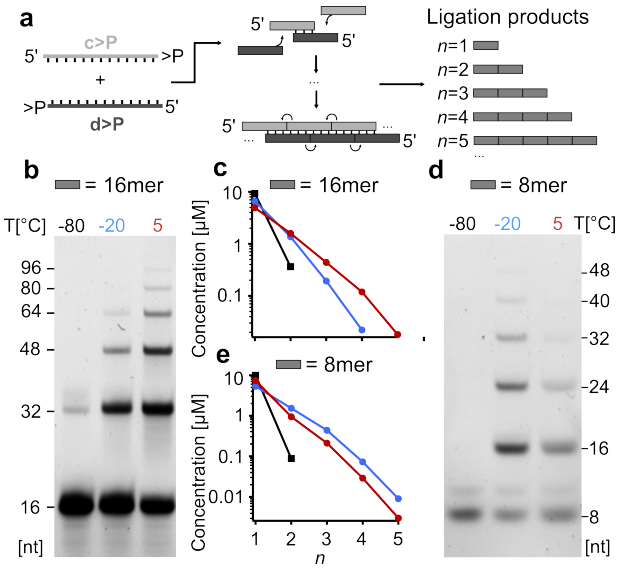


Figure 3.20: Assembly of long RNA via splinted ligation of 2′, 3′ cyclic phosphate containing oligonucleotides. a, Schematics of sequence design. Two strands (*c* and *d*) with complementary sub-regions and not corresponding to the complete reverse complement, bind to form a long chain with repeating units of each. Both *c* and *d* contain 2′, 3′-cyclic phosphate, and the ligation can yield all possible length multiples of the initial strands. For the case where **c** and **d** are 16-mer long, denaturing PAGE of ligation reaction at -80°C, -20°C(frozen) and 5°C (**b**) revealed products of up to five concatenations. The concentration of each ligation product (up to n=5), obtained from the PAGE quantification analysis, is plotted in **c**. Similarly, for an initial strand size of 8-mer, denaturing PAGE (**d**) and product concentration (**e**) are shown. The optimal temperature for splinted ligation depends on oligomer length, as 5°C yields higher concentration for 16-mer while -20°C is better for 8-mer. Reaction were performed with 10 µM primers, 10 µM template, 50 mM CHES pH 10 and 1 mM $MgCl_2$ for 7 days. Data are represented as mean of three independent replicates.

After understanding that the non-enzymatic ligation with 2′,3′ cyclic phosphates is a reliable copying mechanism, we aimed to investigate the potential for elongation. This would have been an important characteristic of a potential prebiotic replication mechanism, as it would establish a link between non-templated nucleotide condensation and the faster replication by long ribozymes, with tens or hundreds of nucleotides [54, 58, 120].

Splint strands with short (4- or 8-mer) binding regions that can both cross template and ligate (Figure 3.20 **a**) were designed to explore the possibility of bridging these two oligonuceotide length regimes. Each system has two strands (labelled *c* and *d*) with >P. The strands are

designed such that the 5′ half of the strand *d* is a reverse complement to the 5′ half of *c* and the same goes for the 3′ halves such that they bind and form a long network of *ccc*... bound to *ddd*.... The formed secondary structure allows for multiple ligations resulting in homopolymers of *c* and *d*. Figures 3.20 *b* and *d* show that up to 5 concatenations (n=6) could be detected for both the 16-mer system (Figure 3.20 **b**) and 8-mer system (Figure 3.20 **c**) resulting in 96- and 48-mer RNA respectively. Figures 3.20 **c** and **e** show the PAGE quantification of the respective gels in Figures 3.20 **b** and **d**. For the shorter system, the yield was reduced by about one order of magnitude for each additional concatenation, and was generally lower than for the 16-mer system, which was likely due to the slower kinetics in frozen state at -20 °C. Specifically for the case of the 16-mer at 5°C 5.1 μM of the strands were incorporated into the concatemers, and 4.9 μM remain. The remaining strands *c* and *d* contain either the active or inactive phosphate group, as the two species do not resolve through PAGE. For these conditions, as for the system in Figure 2.2, it is hypothesized that the main limitation to the yield is the hydrolysis of the cyclic phosphate. It is interesting to note that the maximum yield for the 16-mer system was obtained at 5°C whereas for the 8-mer system it was at -20 °C. This is likely resulting from the low duplex stability of the 4-mer duplex region at 5°C. This is supported by the results of ligation with shorter systems (Section 3.3.8).

3.4 Conclusion

A non-enzymatic replicator chemistry on the early Earth should have the capacity, under plausible conditions, to elongate strands and undergo further replication steps all while being highly accurate and processive. This work demonstrates that ligation with >P RNA fulfills these criteria.

Firstly, we show that the template-directed replication mechanism, only requires salts for stabilizing the duplex and alkaline pH making the ligation with >P RNA robust, reproducible and high-yielding in both aqueous and frozen solutions. For the aqueous case, 38% yield was observed in contrast to previously reported yield of 16% for similar reactions. It was found that a combination of high pH and low temperature promotes ligation over the hydrolysis of the cyclic phosphate moiety. Such conditions are thought to be plausible on early Earth, where the fainter sun contributed to cold surface temperature that would still allow liquid water [112]. Additionally, Hadean oceans were potentially alkaline due to the sequestering of CO_2 in carbonate minerals [62], and alkaline conditions present in freshwater volcanic lakes [133, 134], have been proposed to foster early metabolism.

However, a fraction of the >P still hydrolyzes, which contributes to its incomplete conversion. While the yield could be further improved by adding reagents that aid in the re-cyclization of the monophosphate moiety, such as diamidophosphate in combination with imidazole [123], this would increase the complexity of the system. Moreover, this reaction has low salt requirements (1 mM $MgCl_2$, Figure 3.5) ensuring RNA backbone integrity and compatibility with strand separation. It can also proceed in a wide pH range, even un-buffered (Figures 3.5 and 3.6).

The elongation of short RNA was demonstrated with splinted systems that yielded up to six-copy concatemers of short RNA strands of either 16- or 8-mer, resulting in long RNA on the scale of 100-mer (Figure 3.20). This length range approaches the average length of replicating ribozymes [54, 58, 120], representing a significant step toward assembling functional RNAs by plausible means. Even very short >P RNA fragments (with 4-mer base-pairing regions) ligate under frozen conditions (Figure 3.19), establishing a bridge from the single nucleotide condensation reactions, yielding very short RNA strands, to a regime where templated ligation reactions could dominate.

Furthermore, we evaluate the copying accuracy with two templates with varying nucleotides at the ligation site. One single mutation at either the 3′ or 5′ end nucleotide resulted in a reduction by more than 82% in yield (Figure 3.17), even when the remaining primer was entirely complementary. To compare with a base-by-base replication, such as primer extension, we calculated a yield of 89% that each nucleotide addition should have in order to obtain 38% yield for adding eight nucleotides. If we would have chosen primers with length 4 to 6 nucleotides, we estimated that a base-by-base replicator offers a fidelity between 95% and 98% depending on whether the ligation includes a **G**-base or a **U**-base on the template, respectively.

Contrary to previous studies on >P, we found that under the tested conditions, the reaction was not regioselective, producing equal amounts of 2′–5′ and 3′–5′ linkages at the ligation site (Figure 3.15). While approximately half of the linkages were non-canonical, we argue this does not diminish the applicability of the reaction in a prebiotic context. Strands with 2′–5′ linkages have been shown to fold into functional structures [121], and these non-canonical linkages have also been demonstrated to be more labile than 3′–5′ and have potential for interconversion [59, 62]. Furthermore, Figure 3.15 shows that the product *ab* with a 2′–5′ linkage

at the ligation site could still template the reverse ligation reaction. The non-canonical linkage in the template at the ligation site did not impede ligation, paving the way for exponential replication cycles.

Strand-separation, driven by non-equilibrium environments with thermal, salt or pH oscillations would allow for the implementation of a ligation chain reaction, similar to Edeleva et al., with the added benefit of not generating deleterious side-products by a prebiotically implausible EDC [37]. An air-water interface with continuous feeding would allow for both the denaturation and replenishment of activated primers, without necessarily high temperature which could degrade the cyclic phosphate and the RNA backbone6. This suggests that such scenarios could provide a niche where the ligation reactions by 2′, 3′ cyclic phosphate could evolve towards a ribozymatic replicator.

Considering these results, ligation with >P is an interesting framework to produce diverse pools of long RNA that could undergo molecular evolution. The system described in the current study enables the generation of long RNA, with high fidelity. This was demonstrated for a range of lengths, sequence combinations, reaction conditions and temperatures suggesting that ligation of RNA with >P holds a central position in the general conception of the RNA world.

3.5 Experimental Realization

3.5.1 Nucleic acid sequences

Table 3.4: Oligonucleotide sequences used in the experiments. All the sequences were ordered from biomers.net in lyophilized form and were diluted to a 200 µM stock concentration in RNAse free water, and stored in the -80°C. All the primers with >P were ordered with a 2′, 3′-cyclic phosphate. The polyadenosine overhangs were added to the primer strands to be able to differentiate the ligation product and the template through PAGE and HPLC.

Strand name	Sequence	Length (nt)	Remarks
Primer *a*	**AAAGCAUCAGU** >P	11	Sequences of the system presented in Figure 3.3. Template *BA* is also denominated template **GA** in Figure 3.17 of the main text, to differentiate the ligation site. Product *ab* is used for HPLC quantification.
Primer *b*	**CUCAUAGGAAA**	11	
Template *BA*	**CCUAUGAGACUGAUGC**	16	
Product *ab*	**AAAGCAUCAGUCUCAUAGGAAA**	22	
Primer *a*-**G**	**AAAGCAUCAGG** >P	11	
Primer *a*-**C**	**AAAGCAUCAGC** >P	11	
Primer *a*-**A**	**AAAGCAUCAGA** >P	11	
Primer *b*-**G**	**GUCAUAGGAAA**	11	Used for specificity of ligation site (Figure 3.17).

Strand name	Sequence	Length (nt)	Remarks
Primer *b*- **A**	**AUCAUAGGAAA**	11	
Primer *b*- **U**	**UUCAUAGGAAA**	11	
Template *BA*- **AU**	**CCUAUGAUACUGAUGC**	16	
ab $(2' - 5')$	**AAAGCAUCAGU(2′-5′)CUCAUAGGAAA**	16	The ligation site linkage is $2' - 5'$. Used for UV absorbance calibration of the linkage assay (Section 3.3.5).
Primer *a* - short 6	**AAAAUCAGU** >P	9	Used to test shorter primer-primer-template systems and their ligation yields (Figure 3.19). 6 and 4 correspond to the binding region excluding overhangs. The product was used for UV absorbance calibration.
Primer *b* - short 6	**CUCAUAAAA**	9	
Template *BA* - short 6	**UAUGAGACUGAU**	12	
Product *ab* - short 6	**AAAAUCAGUCUCAUAAAA**	18	
Primer *a* - short 4	**AAACAGU**	7	
Primer *b* - short 4	**CUCAAAA**	7	
Template *BA* - short 4	**UGAGACUG**	8	
Product *ab* - short 4	**AAACAGUCUCAAAA**	14	
Splint *c*	**GCAUCAGUCUCAUAGG** >P	16	Used for ligation through concatenation (Section 3.3.9).
Splint *d*	**ACUGAUGCCCUAUGAG** >P	16	
Splint *c* - short 8	**GCAUCAGU**	8	
Splint *d* - short 8	**AUGCACUG**	8	
Primer *A*	**AAAAACCUAUGAG**	13	Used for ligation of the reverse system (Section 3.3.4).
Primer *B*	**ACUGAUGCAAAAA**	13	
Template *BA*	**CCTATGAGACTGATGC**	16	

RNA oligonucleotides were purchased in dry form from biomers.net and then adjusted to a stock concentration of approximately 200 µM with nuclease-free water (Ambion nuclease-free

water from Invitrogen). The stocks solutions were stored at -80°C and thawed a maximum of five times. All the sequences used in this project are presented in Table 3.4.

3.5.2 Sample preparation

Non-enzymatic ligation reactions were performed with equimolar amount of primers and template (1 μM for the screening experiments and 10 μM for the sequence specificity, splinted ligation experiments and shorter systems and 20 μM for the Nuclease P1 digest analysis). The concentration was increased in order to facilitate analysis for lower yield samples. The salts were always kept at 1mM $MgCl_2$ (Ambion) unless stated otherwise. Time points were varied between 0 h and 168 h and temperature between -20°C and 25°C (-80 °C served as the negative control). The indicated pH for each experiment (pH 9-11) was maintained with 50mM of buffer, either CHES, Tris of Bis-Tris Propane.

3.5.3 Ethanol precipitation

After ligation, the samples were quenched and purified through ethanol precipitation. To each of the samples, 20 μg of glycogen (Sigma) and 500 mM ammonium acetate (Sigma) were added. To this, 3 volumes of cold 100% ethanol (Carl Roth) were added and the samples were incubated at -80°C for 30 min. The samples were then centrifuged at 15000 rpm for one hour at 4°C. The obtained pellets were washed with cold 70% ethanol and centrifuged at 15000 rpm for 10 min at 4 °C. The resulting pellets were air-dried and dissolved in the required volume of nuclease-free water (Ambion) for downstream analysis.

3.5.4 Denaturing PAGE

PAGE was used to analyze and quantify the length distribution of the strands obtained at different time points of polymerization. The samples were run in a denaturing 15% polyacrylamide made from a 40% acrylamide/bis-acrylamide (19:1) stock solution (Carl Roth) and contained 50 wt% urea and 1x TBE (from 10x, Carl Roth) and polymerized with TEMED and APS. Each gel has a thickness of 0.75 mm and approximately 5 mL of the gel mixture. The gel mixture was prepared with 5 mL of the 15% PAA mixture, 25 μL of APS and 2.5 μL of TEMED.

The gels were pre-heated in the electrophoretic chamber at 300 V for 27 min. The samples were then loaded, in a mixture with a ratio of 2:7 of sample to loading dye. Loading dye is prepared in-house (for 10 mL: 9.5 mL formamide, 0.5 mL glycerol, 1 μL EDTA [5] (0.5 M) and 100 μL Orange G dye (New England BioLabs). The samples were at 50 V for 5 min followed by 300 V for 25 min. After the run, the gels were stained with a 2x SYBR Gold (Thermo Fischer Scientific) dilution in TBE buffer 1x. They were then rinsed with 1x TBE buffer twice and imaged using a Bio-Rad ChemiDoc MP imaging system.

3.5.5 PAGE quantification

The ligation of splinted strands (Figure 3.20) yields long RNA strands (up to 100-mer). To estimate the concentration of these strands, SYBR gold stained gel images were analyzed with a self-written LabVIEW tool which extracts the peaks intensity for all the lanes, subtracted from

[5] short for ethylenediaminetetraacetic acid

background. To convert this data to a concentration distribution, the following presumptions were taken into account:

- SYBR Gold fluorescence intensity is linear with oligonucleotide concentration at the working conditions (i.e. 2X) as the SYBR Gold binding sites are saturated [66].
- For a given concentration of oligonucleotides, SYBR Gold fluorescence increases linearly with strand length [67].

The initial concentration of the strands was 10 μM concentration (for example, Figure 3.4) and 22 and its length was either 16-mer or 8-mer, depending on the splinted system. Knowing this, we could determine the concentration of the elongated fragments through normalization. For the normalization, the total intensity of the SYBR gold for the negative control (in the same gel) was used as a calibration point for 10 μM, from which the concentrations of other strands were derived. This was done for three independent triplicates.

3.5.6 Nuclease P1 digestion

Nuclease P1 digests the canonical 3′–5′ linkages of both RNA and DNA into acid soluble nucleotides, but it does not cleave 2′–5′ linkages. While the Nuclease P1 is regarded as a single-strand specific nuclease, it has been demonstrated to cleave long structured RNA such as tRNA, rRNA and viroid RNA at pH 5.3. The product of ligation reaction in our experiments can have both 3′–5′ and 2′–5′ linkages at the ligation site, while the rest of the backbone is 3′–5′ linked. Thus, upon digestion, the strand with 3′–5′ linkage at the ligation site would be completely cleaved as well as all the phosphodiester bonds of the template and the unreacted primers, and of the 2′–5′ linked product except the dimer with the 2′–5′ linkage.

The ligation reactions were prepared as described above (20 μM each strands in 20 μL) and incubated for 24 h at 25°C and for 72 h and 168 h at 5°C. A sample kept at -80°C served as the negative control for ligation. After ligation, the samples were purified through ethanol precipitation and then redissolved in 20 μL of nuclease-free water. A 2 μL aliquot of the precipitated sample was taken and diluted to 20 μL for the HPLC quantification of the product. Additionally, for calibration, a synthetic product *ab* strand was purchased to contain 2′-5′ linkage at the ligation site (Product *ab* 2′-5′). A dilution series (20 μM to 0.625 μM) of the Product *ab* 2′-5′ was made in 20 μL. More details about peak identification and quantification of the **UC** dimer shown are shown in Appendix 3.A). For the digestion by nuclease P1, the remaining 18 μL volume of the sample and that of the 2′–5′ linked standard dilutions were used.

The Nuclease P1 (New England Biolabs Inc., M0660S) is supplied as a solution of 100,000 Units/mL along with a vial of 10X Nuclease P1 Reaction Buffer (B0660S). A 2X enzyme mix was prepared such as to contain 2X buffer and 0.56 units of enzyme/μL. For the digestion, 18 μL of the samples (and 2′–5′ product standards) were mixed with 18 μL of the 2X enzyme mix. Thus, the final digestion mixes contained 10 Units of enzyme, 1X reaction buffer (50 mL Sodium Acetate pH 5.5) and 360 pmol of each oligonucleotide. These digestion mixes were then incubated at 37°C for 30 min, following which the enzyme was heat inactivated at 75°C for 10 min.

3.5.7 HPLC-UV absorbance

Ion-pairing reverse phase HPLC experiments were carried out on an Agilent 1260 Infinity II LC System coupled to Agilent 1260 Infinity II DAD WR detector and an Agilent 6230B ESI TOF Mass spectrometer. A C18 capillary column (AdvanceBio Oligonucleotide 4.6x150 mm with particle size 2.7 μm, Agilent) was used to perform reverse phase liquid chromatography. The temperature of the autosampler was set to 4°C. The mobile phases consisted of two eluents. Eluent A was HPLC water (Sigma-Aldrich), 200 mM HFIP (Carl Roth GmbH), 8 mM TEA (Carl Roth GmbH). Eluent B was a 50:50 (v/v) mixture of water and methanol (HPLC grade, Sigma Aldrich, Germany), 200 mM HFIP, 8 mM TEA. The injection volume was 20 μL.

The samples were eluted with a gradient of 20% B to 55.5% B over the course of 18.5 min with a flowrate of 1.5 mL/min at 60°C. Afterwards the column was flushed at 100% Eluent B for 3.31 min and then re-equilibrated at 20% Eluent B for 3.67 min. Retention times were analyzed via a UV Diode Array Detector (Agilent 1260 Infinity II Diode Array Detector WR G7115A) at 260 nm with a bandwidth of 4 nm. A standard solution was prepared at known concentration of strands and, using the auto-sampler, different volumes of the standard solution were injected. This was used to make the calibration curve. The peaks in the UV chromatogram were assigned based on the mass spectrum obtained in negative mode. The ionization parameters are – drying gas temp: 325°C, drying gas flow: 8 L/min, nebulizer pressure: 35 psi, sheath gas temp: 350°C, sheath gas flow: 11 L/min, VCap: 3500 V, and nozzle voltage 1000 V.

3.5.8 Quantification with HPLC

The quantification of strands (for Figures 3.7, 3.8, 3.11, 3.13, 3.14, 3.15 and 3.17) was done with HPLC. The UV 260 nm chromatogram was integrated for each strand peak using MassHunter Qualitative Analysis Navigator and the integrated values were used for the quantification. Using a calibration curve of picomoles vs integrated values of the UV chromatogram the concentrations of the strands were quantified. The concentrations of the strands were normalized with respect to the template strand in the corresponding sample to correct for slight pipetting variations. The concentration of *BA* can be observed in Figure 3.11, where the template concentration varies slightly across different samples. For pH 11 at 25°C rapid hydrolysis of the template is observed, whereas for pH 11 at 10°C, this is only significant for later time points. The remaining samples do not reveal hydrolysis. Since most of the quantifications done were for lower pH and/or temperature, it was assumed this did not introduce artifacts.

3.5.9 Illumina sequencing

All the samples were sequenced by the Gene Center Munich (LMU) using the NGS Illumina NextSeq 1000 machine (flow cell type P2, 2 x 100 bp with 138 cycles for 50bp single-end reads). 20 million reads were ordered for each sample. The raw sequencing data obtained, in FastQ format, was processed in this order by demultiplexing, quality score trimming, and regular expression filtering. Before sequencing, the samples were prepped using the SMARTer smRNA-Seq Kit (Takara) for library preparation. Demultiplexing was performed with software from Galaxy servers [2], provided by the Gene Center Munich. During sequencing, each read base was assigned a Phred quality score $Q = -10 \log_{10} P$, where P is the probability of an incorrectly read base [39]. Using Trimmomatic [13] the low quality segments were trimmed by

running a sliding window of 4 nt in the 3′ to 5′ end direction over the sequence that allowed a minimum average Phred quality of 20, otherwise trimming at the leftmost base of the window, corresponding to an average accuracy of at least 99%. To find the reads that contained a ligation product *ab* the following RegEx was used in series:

(?<=AAAGCATC)(.*)(?=CTCATAGGAAA)
(^[ATCG]3,3)

The first to obtain the intermediate region between the two primer sequences (as this should correspond to the random region) and the second to filter only the 3-mer regions from those. To obtain the random segment in the primer *a* in the t_0 sample the following ReGex was used:

(?<=AAAGCATC).*
^.0,3

The first expression extracts the sequence right after the primer segment and the second trims the first three nucleotides, which should correspond to the random segment of the primer.

Appendix

3.A Identification of the 2′–5′ linkage by nuclease P1 digestion

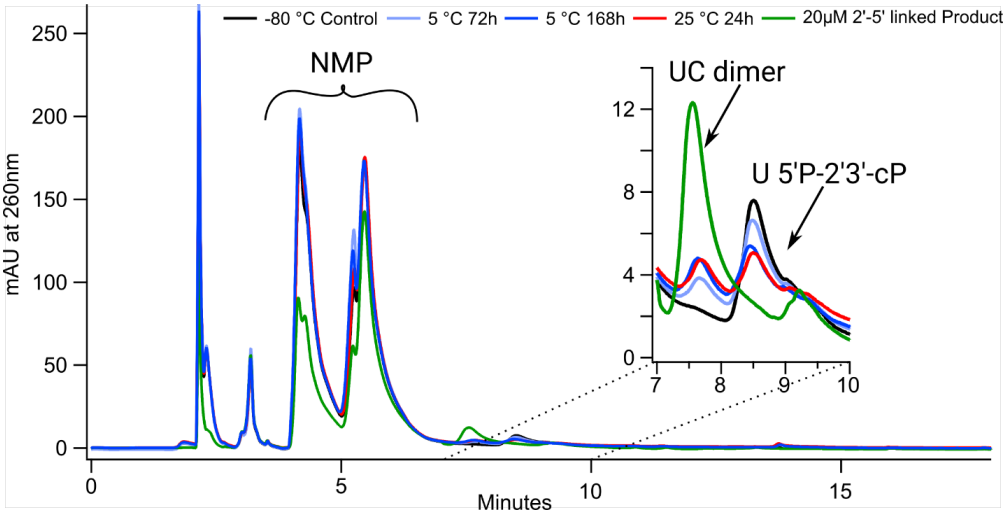


Figure 3.21: UV chromatogram of the nuclease P1 digests of ligation reactions. After the ligation reaction, the ethanol precipitated samples were digested with Nuclease P1 (protocol described in section 3.5.6). Following the digestion, the digests were analyzed by HPLC coupled to an ESI TOF mass spectrometer. The UV chromatograms were obtained at 260 nm. Nucleosides monophosphate (likely 5′-) elute between 4-7 min following which two major peaks can be seen between 7 and 10 min. Inset shows a zoomed-in view of the 7 to 10-min region, with arrows highlighting the peak identities. A minor peak can also be seen at 14 min. The digests traces of different samples are colored as following- black: -80 °C control; light blue: ligation at 5°C for 72h; deep blue: ligation at 5°C for 168h; red: ligation at 25°C for 24 h; green: 20 µM 2′-5′ linked product standard.

The analysis of the digested sample was done with the same HPLC system described in Section 3.5.7. However, an updated method to resolve the mono- and di-nucleotide (up to 5-mer can be resolved) was used: the column was kept at 30 °C and the flow rate was 0.6 mL/min. The gradient started with 1% Eluent B for 4 min after which it was increased first to 4% in 3 min, then to 8% in 3.2 min, and 35% in 5 min. Following which the column was flushed with 100% Eluent B and then re-equilibrated at 1% Eluent B for 7 min each. For the peak identification by ESI TOF [6]mass spectrometer in negative mode following ionization parameters were used - drying gas temp: 325°C, drying gas flow: 8 L/min, nebulizer pressure: 45 psi, sheath gas temp: 400°C, sheath gas flow: 11 L/min, VCap: 3500 V, and nozzle voltage 2000 V.

[6]short for Electrospray Ionization Time-of-Flight

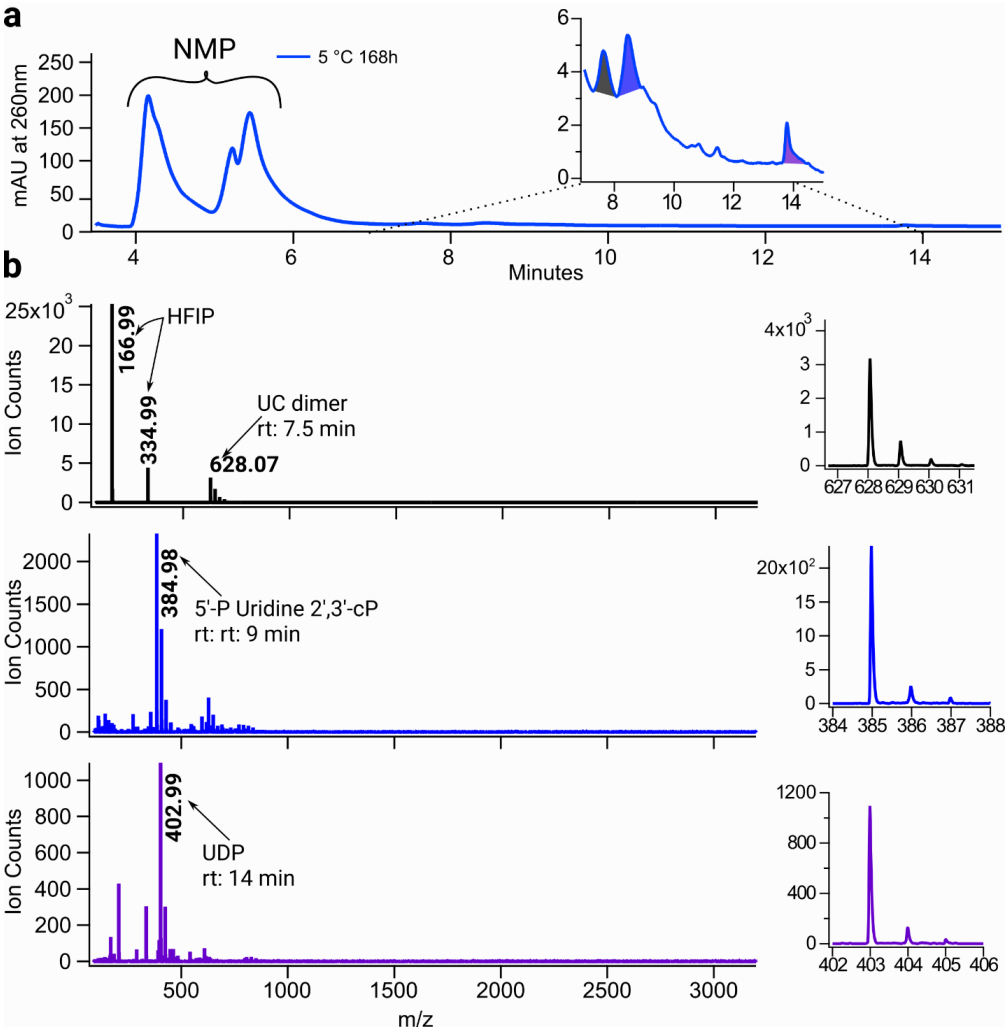


Figure 3.22: Peak identification by mass spectrometry. **a**, UV chromatogram of the nuclease P1 digests for the sample ligated at 5 °C for 168 hours. The inset shows a zoomed-in view from 7-15 min showing three distinct peaks highlighted in black (retention time: 7.5 min), blue (retention time: 8.5 min) and purple (retention time: 14 min). Nucleoside monophosphate peaks are between retention time 4 to 7-minutes and are marked with the braces. **b**, Background subtracted mass spectra extracted from the highlighted region shown in the UV chromatogram in **a**. Three major distributions can be seen in the black spectrum- m/z of 166.99 and 334.99 depict HFIP and its adducts which is present in the HPLC eluents as buffer along with TEA. The next prominent distribution is for the m/z of 628.07 which corresponds to the m/z of the **UC** dimer. Similarly, major peak in the blue spectrum at m/z of 384.99 corresponds to 2′,3′-cyclic, 5′- Uridine bisphosphate and the major peak in purple spectrum at m/z 402.99 corresponds to UDP [7] (or Uridine (2′)3′, 5′ bisphosphate). The adjoining spectra to each of the full spectrum shows a zoomed-in view of the corresponding m/z.

The UV chromatogram of different digests and that of 2′-5′ linked product standard is presented in Figure 3.21. With the HPLC method (Section 3.5.7), the non-interacting species

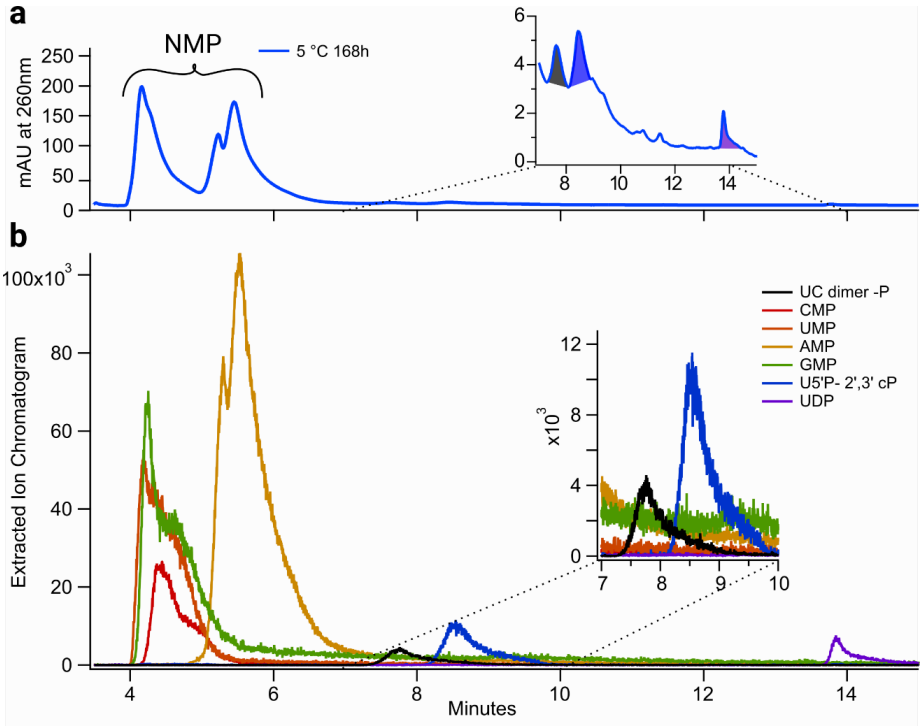


Figure 3.23: Confirmation of the UV peaks by extracted ion chromatogram. **a**, (same as Figure 3.22 **a**) UV chromatogram of the nuclease P1 digests for the sample ligated at 5 °C for 168 hours. The inset shows a zoomed-in view from 7-15 min showing three distinct peaks highlighted in black (retention time: 7.5 min), blue (retention time: 8.5 min) and purple (retention time: 14 min). Nucleoside monophosphate peaks are between retention time 4 to 7-min and are marked with the braces. **b**, Using the theoretical m/z of the molecules identified with their retention times and mass spectra, the ion chromatograms were extracted. This confirms that the molecules identified are indeed present in a single peak and do not pertain to background in the mass spectra or have multiple retention times.

(no negative charge or low hydrophobicity to interact with the ion-pairing C18 column) elute in the column void fraction (0-3 min), followed by the nucleosides. Nucleosides monophosphate elute first (4-7 min) as marked in the Figure 3.21, after which, different species with multiple phosphates elute. Three distinct peaks are observed in the chromatogram after the first 7 min for the digested sample and can be better visualized in the zoomed-in view in Figure 3.22 **a** and 3.23 **a** for the digestion of the sample ligated at 5°C for 168 h. Figure 3.22 and 3.23 detail on how the identity of the peaks in the UV chromatograms were determined using ESI TOF data.

For the peak at approximately 7.5 min, the green trace in Figure 3.21 depicts the digestion of the 2′-5′ linked standard product. The lack of this peak in the -80 °C control sample (black UV chromatogram) suggests the identity of this peak to be that of the 2′-5′ linked **UC** dimer (with a 5′ phosphate). The black mass spectrum in Figure 3.22 and the extracted ion chromatogram for the m/z of the **UC** dimer (628.0699) confirm this identification of the peak. For the peak at 8.5 min, it is interesting to note that it is only present in the digests of the ligation reactions and the -80 °C control but not in that of the digested 2′–5′ linked product. This corresponds to the digestion of the phosphodiester linkage between the penultimate and the last nucleotide

(**U**) with the 2′,3′ cyclic phosphate of the primer *a,* thus corresponding to 2′,3′ cyclic-, 5′-uridine bisphosphate.

Another major peak is observed at 14 min (slight bump can be seen in Figure 3.21) which can be better appreciated in the UV chromatogram in Figure 3.22 **a** and 3.23 **a**. The mass spectrum of that region (purple spectra) shows an m/z of 402.99 peak, corresponding to the mass of UDP. While this could be any possible isomer of UDP, its presence in the digested ligation mix suggests it to be either the cleavage of the linkage between the penultimate and the last nucleotide (**U**) with the inactive primer *a* or the hydrolysis product of the 2′,3′ cyclic-, 5′- Uridine bisphosphate, both of which lead to (2′)3′, 5′ uridine bisphosphate.

Quantification of the of the 2′–5′ linked product was done in the following steps:

- Firstly, as described above and in Section 3.5.7, an aliquot of the 20 µL sample, was used to quantify the amount of total product *ab* in the ligated samples using the HPLC UV chromatogram. Similar quantification was done for the serial dilution of the strand *ab* 2′-5′.
- Nuclease P1 digestion of the serial dilution of *ab* 2′–5′, generates a concentration series of the 2′–5′ linked **UC** dimer. Due to the complete digestion, the concentration of the *ab* 2′–5′ dilutions (computed in step 1) gives us the concentrations of the 2′–5′ linked **UC** dimer. This was used to make a calibration curve (pmol vs integrated counts) for the 2′–5′ linked **UC** dimer.
- This calibration curve was used to quantify the amount of 2′–5′ linked **UC** dimer in the ligation samples after the Nuclease P1 digestion. This amount of the 2′–5′ linked **UC** dimer corresponds to the amount of the ligated product with the 2′–5′ linkage.
- Thus the concentrations of both total ligated product and the 2′–5′ linked product are obtained which are then used to calculate the composition of the phosphodiester linkages formed at the ligation site shown in Section 3.3.5.

Discussion and Outlook

Nucleic acids play an important role in contemporary biology in both information storage and functionality. It is theorized that an early replication network consisting solely of RNA predates the sophisticated protein machinery necessary for modern biological functions. This concept finds support in the existence of functional RNA machinery, such as the ribosome, which may be a remnant from that primordial era. While diverse pools of nucleic acids have been shown to undergo molecular evolution, selecting for functional sequences becomes more challenging the longer the sequences of the initial pool are. Due to sequence space limitations, not all the sequences can be represented in a pool with dozens or hundreds of base-pairs. This would impede retrieval of functional sequences that have rare secondary structures. Selection and elongation of the sequences in the pool must have been concomitant phenomena.

Environmental factors exert selective pressures on nucleic acid strands, termed extrinsic selective pressure. Cyclic flux changes, like tidal waves, may periodically remove and replenish prebiotic pools of oligonucleotides, leading to the enrichment of sequences capable of forming clusters or compartments, as demonstrated in Chapter 1 regarding sequences undergoing LLPS. Sequences phase-separated through the formation of base-pair networks which grew long enough to form DNA condensates. Only sequences with a four-letter alphabet were able to form these networks and phase separate. This necessary compositional diversity allows for the stabilization of the structure, as binary alphabets have more binding possibilities owing to their smaller sequence space which could lead to self-folding or other non-specific secondary structure.

The chemical mechanisms governing pool elongation and replication also impose selective pressures, termed intrinsic selective pressure. This was explored with enzymatic model system in Chapter 2. The replication of short random binary oligonucleotide pools, with an initial bias to specific nucleotides homogenized the overall pool composition to about 50% of each nucleotide while at the same time preserving positional biases that depend on the initial pool. Structured sequences with periodicity facilitated replication through self- and pool-templation, although excessively stable secondary structures such as fully bound duplexes got stalled.

In both of these model systems, sequences that base-pair through the formation of cooperative networks of several sequences, while avoiding fully bound duplexes, have an advantage. In the case of Chapter 1, this is an advantage of survival against dilution, in comparison to other sequences that remains in solution. In the case of Chapter 2, this is a replicative advantage. Periodic sequences have more binding possibilities, which increases the chances of being recruited for replication. In a prebiotic pool, with shorter sequences, cooperative inter-sequence binding also allows to bundle and associate more information and functional groups together, which could later evolve to a different segments of a single genome.

In order to explore the intrinsic selective pressures of prebiotic replication on nucleic acid pools, non-enzymatic replication chemistries need to be described and understood. In Chapter 3, the kinetics and fidelity of the templated ligation of short RNA fragments with

cyclic 2′, 3′-phosphates. The timescales necessary for this mechanism (about 3 d to observe 40% yield) do not yet allow for molecular evolution experiments, hence the enzymatic model system used in Chapter 2. An SP that binds to form a network was shown, with this activation chemistry, to ligate consecutively and yield an 100-mer. On the one hand, this result paves the way to prebiotically create pools of longer oligomers, of lengths closer to that of ribozymes. On the other hand, sequences capable of forming such networks could have an advantage against dilution, through the formation of condensates (Chapter 1) and a replicative advantage over other sequences that do not base pair in this way (Chapter 2). The capacity of sequences to form cooperative networks grants them a survival advantage. This is seen from different perspectives, either physical sheltering by the formation of condensates that phase separate, or by the replicative advantage to have secondary structures that are not stuck in a fully bound duplex allowing for efficient elongation.

In the future, SPs that phase-separate known from Chapter 1 could be investigated in RNA to bridge the gap with the non-enzymatic chemistry described in Chapter 3. This clustering of sequences could bring the interacting moieties closer together and yield even longer polymers than the 100-mer observed. These condensates may also have a protecting effect against hydrolysis, both of the phosphodiester bonds and of the >P moiety which should be investigated. The diversity of sequences in the initial pool for phase separation studies should also be increased, possibly with the addition of random stretches. The evolution of these stretches could be followed with NGS as the sequencing analysis pipeline is developed for Chapter 2. This way the joint selective pressure of phase separation and ligation could be assessed.

Bibliography

[1] Mounir G AbouHaidar and Ivan G Ivanov. Non-enzymatic rna hydrolysis promoted by the combined catalytic activity of buffers and magnesium ions. *Zeitschrift für Naturforschung C*, 54(7-8):542–548, 1999.

[2] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*, 46:W537–W544, 2018.

[3] Siddharth Agarwal, Dino Osmanovic, Melissa A Klocke, and Elisa Franco. The growth rate of DNA condensate droplets increases with the size of participating subunits. *ACS nano*, 16(8):11842–11851, 2022.

[4] Eva Agustriana, Isa Nuryana, Fina Amreta Laksmi, Kartika Sari Dewi, Hans Wijaya, Nanik Rahmani, Danu Risqi Yudiargo, Astadewi Ismadara, Helbert, Moch Irfan Hadi, et al. Optimized expression of large fragment DNA polymerase I from Geobacillus stearothermophilus in Escherichia coli expression system. *Preparative Biochemistry & Biotechnology*, pages 1–10, 2022.

[5] William M Aumiller Jr, Fatma Pir Cakmak, Bradley W Davis, and Christine D Keating. RNA-based coacervates as a model for membraneless organelles: formation, properties, and interfacial liposome assembly. *Langmuir*, 32(39):10042–10053, 2016.

[6] Nenad Ban, Poul Nissen, Jeffrey Hansen, Peter B Moore, and Thomas A Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289(5481):905–920, 2000.

[7] LM Barge, E Branscomb, JR Brucato, SSS Cardoso, JHE Cartwright, SO Danielache, D Galante, TP Kee, Y Miguel, S Mojzsis, et al. Thermodynamics, disequilibrium, evolution: Far-from-equilibrium geological and chemical considerations for origin-of-life research. *Origins of Life and Evolution of Biospheres*, 47:39–56, 2017.

[8] Giacomo Bartolucci, Adriana Calaça Serrão, Philipp Schwintek, Alexandra Kühnlein, Yash Rana, Philipp Janto, Dorothea Hofer, Christof B Mast, Dieter Braun, and Christoph A Weber. Sequence self-selection by cyclic phase separation. *Proceedings of the National Academy of Sciences*, 120(43):e2218876120, 2023.

[9] Robert T Batey, Robert P Rambo, and Jennifer A Doudna. Tertiary motifs in rna structure and folding. *Angewandte Chemie International Edition*, 38(16):2326–2343, 1999.

[10] Sidney Becker, Christina Schneider, Antony Crisp, and Thomas Carell. Non-canonical nucleosides and chemistry of the emergence of life. *Nature communications*, 9(1):5174, 2018.

[11] Stephen J Benkovic, Ann M Valentine, and Frank Salinas. Replisome-mediated DNA replication. *Annual review of biochemistry*, 70(1):181–208, 2001.

[12] Silvia Biffi, Roberto Cerbino, Francesca Bomboi, Elvezia Maria Paraboschi, Rosanna Asselta, Francesco Sciortino, and Tommaso Bellini. Phase behavior and critical activated dynamics of limited-valence DNA nanostars. *Proceedings of the National Academy of Sciences*, 110(39):15633–15637, 2013.

[13] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[14] Ronald. Breslow and Marc. Labelle. Sequential general base-acid catalysis in the hydrolysis of rna by imidazole. *Journal of the American Chemical Society*, 108(10):2655–2659, 5 1986.

[15] Patrick Denis Browne, Tue Kjærgaard Nielsen, Witold Kot, Anni Aggerholm, M Thomas P Gilbert, Lara Puetz, Morten Rasmussen, Athanasios Zervas, and Lars Hestbjerg Hansen. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience*, 9(2), 2020.

[16] Samuel E Butcher and Anna Marie Pyle. The molecular interactions that stabilize rna tertiary structure: Rna motifs, patterns, and networks. *Accounts of chemical research*, 44(12):1302–1311, 2011.

[17] Jamal M. Buzayan, Wayne L. Gerlach, and George Bruening. Non-enzymatic cleavage and ligation of rnas complementary to a plant virus satellite rna. *Nature*, 323(6086):349–353, 9 1986.

[18] Adriana Calaça Serrão, Sreekar Wunnava, Avinash V. Dass, Lennard Ufer, Philipp Schwintek, Christof B. Mast, and Dieter Braun. High-fidelity rna copying via 2,3-cyclic phosphate ligation. *Journal of the American Chemical Society*, 0(0):null, 0.

[19] Adriana Calaça Serrão, Felix T Dänekamp, Zsófia Meggyesi, and Dieter Braun. Replication elongates short DNA, reduces sequence bias and develops trimer structure. *Nucleic Acids Research*, page gkad1190, 12 2023.

[20] Michael J Cavaluzzi and Philip N Borer. Revised UV extinction coefficients for nucleoside-5'-monophosphates and unpaired DNA and RNA. *Nucleic acids research*, 32(1):e13–e13, 2004.

[21] Thomas V Christian and William H Konigsberg. Single-molecule FRET reveals proofreading complexes in the large fragment of Bacillus stearothermophilus DNA polymerase I. *AIMS biophysics*, 5(2):144, 2018.

[22] H. James Cleaves, Kevin E. Nelson, and Stanley L. Miller. The prebiotic synthesis of pyrimidines in frozen solution. *Naturwissenschaften*, 93(5):228–231, 2006.

[23] John B Corliss, JA Baross, and SE Hoffman. An hypothesis concerning the relationships between submarine hot springs and the origin of life on earth. *Oceanologica Acta, Special issue*, 1981.

[24] Carlos E Crespo-Hernández and Rafael Arce. Formamidopyrimidines as major products in the low-and high-intensity uv irradiation of guanine derivatives. *Journal of Photochemistry and Photobiology B: Biology*, 73(3):167–175, 2004.

[25] Edward A Curtis and David P Bartel. The hammerhead cleavage reaction in monovalent cations. *RNA*, 7(4):546–552, 2001.

[26] Bruce Damer and David Deamer. The hot spring hypothesis for an origin of life. *Astrobiology*, 20(4):429–452, 2020.

[27] Avinash Vicholous Dass, Sreekar Wunnava, Juliette Langlais, Beatriz von der Esch, Maik Krusche, Lennard Ufer, Nico Chrisam, Romeo CA Dubini, Florian Gartner, Severin Angerpointner, et al. RNA Oligomerisation without Added Catalyst from 2, 3-Cyclic Nucleotides by Drying at Air-Water Interfaces. *ChemSystemsChem*, 5(1):e202200026, 2023.

[28] Christian De Duve. The onset of selection. *Nature*, 433(7026):581–582, 2005.

[29] Luís H de Oliveira, Pollyana Trigueiro, Baptiste Rigaud, Edson C da Silva-Filho, Josy A Osajima, Maria G Fonseca, Jean-François Lambert, Thomas Georgelin, and Maguy Jaber. When RNA meets montmorillonite: Influence of the pH and divalent cations. *Applied Clay Science*, 214:106234, 2021.

[30] Julien Derr, Michael L Manapat, Sudha Rajamani, Kevin Leu, Ramon Xulvi-Brunet, Isaac Joseph, Martin A Nowak, and Irene A Chen. Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences. *Nucleic acids research*, 40(10):4711–4722, 2012.

[31] Dian Ding, Stephanie J Zhang, and Jack W Szostak. Enhanced nonenzymatic RNA copying with in-situ activation of short oligonucleotides. *bioRxiv*, pages 2023–04, 2023.

[32] Dian Ding, Lijun Zhou, Constantin Giurgiu, and Jack W Szostak. Kinetic explanations for the sequence biases observed in the nonenzymatic copying of RNA templates. *Nucleic Acids Research*, 50(1):35–45, 2022.

[33] Christina F Dirscherl, Alan Ianeselli, Damla Tetiker, Thomas Matreux, Robbin M Queener, Christof B Mast, and Dieter Braun. A heated rock crack captures and polymerizes primordial DNA and RNA. *Physical Chemistry Chemical Physics*, 2023.

[34] Elizabeth A Doherty and Jennifer A Doudna. Ribozyme structures and mechanisms. *Annual review of biophysics and biomolecular structure*, 30(1):457–475, 2001.

[35] TY Dora Tang, C Rohaida Che Hak, Alexander J Thompson, Marina K Kuimova, DS Williams, Adam W Perriman, and Stephen Mann. Fatty acid membrane assembly on coacervate microdroplets as a step towards a hybrid protocell model. *Nature chemistry*, 6(6):527–533, 2014.

[36] Björn Drobot, Juan M Iglesias-Artola, Kristian Le Vay, Viktoria Mayr, Mrityunjoy Kar, Moritz Kreysing, Hannes Mutschler, and TY Dora Tang.

Compartmentalised RNA catalysis in membrane-free coacervate protocells. *Nature communications*, 9(1):3643, 2018.

[37] Evgeniia V. Edeleva, Annalena Salditt, Julian Stamp, Philipp Schwintek, Job Boekhoven, and Dieter Braun. Continuous nonenzymatic cross-replication of dna strands with in situ activated dna oligonucleotides. *Chemical Science*, 10(22):5807–5814, 2019.

[38] Aaron E Engelhart, Matthew W Powner, and Jack W Szostak. Functional rnas exhibit tolerance for non-heritable 2–5 versus 3–5 backbone heterogeneity. *Nature chemistry*, 5(5):390–394, 2013.

[39] Brent Ewing, LaDeana Hillier, Michael C. Wendl, and Phil Green. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*, 8(3):175–185, 1998.

[40] Martha J Fedor and James R Williamson. The catalytic diversity of rnas. *Nature reviews Molecular cell biology*, 6(5):399–412, 2005.

[41] Harold Fellermann, Shinpei Tanaka, and Steen Rasmussen. Sequence selection by dynamical symmetry breaking in an autocatalytic binary polymer model. *Physical Review E*, 96(6):1–14, 2017.

[42] James P Ferris and Gözen Ertem. Oligomerization reactions of deoxyribonucleotides on montmorillonite clay: the effect of mononucleotide structure, phosphate activation and montmorillonite composition on phosphodiester bond formation. *Origins of life and evolution of the biosphere*, 20:279–291, 1990.

[43] James P Ferris, Aubrey R Hill Jr, Rihe Liu, and Leslie E Orgel. Synthesis of long prebiotic oligomers on mineral surfaces. *Nature*, 381(6577):59–61, 1996.

[44] Hartmut Follmann and Carol Brownson. Darwin's warm little pond revisited: from molecules to the origin of life. *Naturwissenschaften*, 96(11):1265–1292, 2009.

[45] Mark E Fornace, Nicholas J Porubsky, and Niles A Pierce. A unified dynamic programming framework for the analysis of interacting nucleic acid strands: enhanced models, scalability, and speed. *ACS Synthetic Biology*, 9(10):2665–2678, 2020.

[46] Paul A. Giannaris and Masad J. Damha. Oligoribonucleotides containing 2,5-phosphodiester linkages exhibit binding selectivity for 3,5-rna over 3,5-ssdna. *Nucleic Acids Research*, 21(20):4742–4749, 1993.

[47] Clémentine Gibard, Subhendu Bhowmik, Megha Karki, Eun-Kyong Kim, and Ramanarayanan Krishnamurthy. Phosphorylation, oligomerization and self-assembly in water under potential prebiotic conditions. *Nature Chemistry*, 10(2):212–217, 2018.

[48] Walter Gilbert. Origin of life: The RNA world. *Nature*, 319(6055):618, 1986.

[49] Nigel Goldenfeld and Carl Woese. Life is physics: evolution as a collective phenomenon far from equilibrium. *Annu. Rev. Condens. Matter Phys.*, 2(1):375–399, 2011.

[50] Tobias Göppel, Joachim H Rosenberger, Bernhard Altaner, and Ulrich Gerland. Thermodynamic and Kinetic Sequence Selection in Enzyme-Free Polymer Self-Assembly inside a Non-equilibrium RNA Reactor. *Life*, 12(4):567, 2022.

[51] Rachel Green and Jack W Szostak. Selection of a ribozyme that functions as a superior template in a self-copying reaction. *Science*, 258(5090):1910–1915, 1992.

[52] Haukur Gudnason, Martin Dufva, Dang Duong Bang, and Anders Wolff. Comparison of multiple DNA dyes for real-time PCR: effects of dye concentration and sequence composition on DNA amplification and melting temperature. *Nucleic acids research*, 35(19):e127, 2007.

[53] Xutiange Han, Erchi Wang, Yixiao Cui, Yikai Lin, Hui Chen, Ran An, Xingguo Liang, and Makoto Komiyama. The staining efficiency of cyanine dyes for single-stranded dna is enormously dependent on nucleotide composition. *Electrophoresis*, 40(12-13):1708–1714, 2019.

[54] David P. Horning and Gerald F. Joyce. Amplification of RNA by an RNA polymerase ribozyme. *Proceedings of the National Academy of Sciences*, 113(35):9786–9791, 8 2016.

[55] Alan Ianeselli, Miguel Atienza, Patrick W. Kudella, Ulrich Gerland, Christof B. Mast, and Dieter Braun. Water cycles in a Hadean CO2 atmosphere drive the evolution of long DNA. *Nature Physics*, 18(5):579–585, 2022.

[56] Alan Ianeselli, Christof B Mast, and Dieter Braun. Periodic melting of oligonucleotides by oscillating salt concentrations triggered by microscale water cycles inside heated rock pores. *Angewandte Chemie*, 131(37):13289–13294, 2019.

[57] Alan Ianeselli, Annalena Salditt, Christof Mast, Barbara Ercolano, Corinna L Kufner, Bettina Scheu, and Dieter Braun. Physical non-equilibria for prebiotic nucleic acid chemistry. *Nature Reviews Physics*, 5(3):185–195, 2023.

[58] Luc Jaeger, Martin C. Wright, and Gerald F. Joyce. A complex ligase ribozyme evolved in vitro from a group i ribozyme domain. *Proceedings of the National Academy of Sciences of the United States of America*, 96(26):14712–14717, 12 1999.

[59] Pia Jarvinen, Mikko Oivanen, and Harri Lonnberg. Interconversion and phosphoester hydrolysis of 2',5'- and 3',5'-dinucleoside monophosphates: kinetics and mechanisms. *The Journal of Organic Chemistry*, 56(18):5396–5401, 8 1991.

[60] Byoung-jin Jeon, Dan T Nguyen, Gabrielle R Abraham, Nathaniel Conrad, Deborah K Fygenson, and Omar A Saleh. Salt-dependent properties of a coacervate-like, self-assembled DNA liquid. *Soft Matter*, 14(34):7009–7015, 2018.

[61] G. F. Joyce. Non-enzymatic template-directed synthesis of rna copolymers. *Origins of Life and Evolution of the Biosphere*, 14(1-4):613–620, 1984.

[62] Shintaro Kadoya, Joshua Krissansen-Totton, and David C Catling. Probable cold and alkaline surface environment of the Hadean Earth caused by impact ejecta weathering. *Geochemistry, Geophysics, Geosystems*, 21(1):e2019GC008734, 2020.

[63] Ryan Kennedy, Manuel E. Lladser, Zhiyuan Wu, Chen Zhang, Michael Yarus, Hans De Sterck, and Rob Knight. Natural and artificial RNAs occupy the same restricted region of sequence space. *RNA*, 16(2):280–289, 2010.

[64] James R Kiefer, Chen Mao, Jeffrey C Braman, and Lorena S Beese. Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. *Nature*, 391(6664):304–307, 1998.

[65] Ryszard Kierzek, Liyan He, and Douglas H Turner. Association of 2'-5'oligoribonucleotides. *Nucleic acids research*, 20(7):1685–1690, 1992.

[66] Pauline J Kolbeck, Willem Vanderlinden, Gerd Gemmecker, Christian Gebhardt, Martin Lehmann, Aidin Lak, Thomas Nicolaus, Thorben Cordes, and Jan Lipfert. Molecular structure, DNA binding mode, photophysical properties and recommendations for use of SYBR Gold. *Nucleic acids research*, 49(9):5143–5158, 2021.

[67] Patrick W Kudella, Alexei V Tkachenko, Annalena Salditt, Sergei Maslov, and Dieter Braun. Structured sequences emerge from random pool when replicated by templated ligation. *Proceedings of the National Academy of Sciences*, 118(8), 2021.

[68] Corinna L Kufner, Stefan Krebs, Marlis Fischaleck, Julia Philippou-Massier, Helmut Blum, Dominik B Bucher, Dieter Braun, Wolfgang Zinth, and Christof B Mast. Sequence dependent uv damage of complete pools of oligonucleotides. *Scientific Reports*, 13(1):2638, 2023.

[69] Daniel A. Lafontaine, Danièle Beaudry, Patrick Marquis, and Jean-Pierre Perreault. Intra- and intermolecular nonenzymatic ligations occur within transcripts derived from the peach latent mosaic viroid. *Virology*, 212(2):705–709, 10 1995. DOI: 10.1006/viro.1995.1528 MAG ID: 1974932928 PMID: 7571440 S2ID: 9d7ab4ed652084a2678c00473a5effb723ba52f8.

[70] Matthew Levy and Stanley L Miller. The stability of the RNA bases: implications for the origin of life. *Proceedings of the National Academy of Sciences*, 95(14):7933–7938, 1998.

[71] Matthew Levy, Stanley L Miller, and John Oró. Production of guanine from NH 4 CN polymerizations. *Journal of molecular evolution*, 49:165–168, 1999.

[72] Li Li, Noam Prywes, Chun Pong Tam, Derek K O'flaherty, Victor S Lelyveld, Enver Cagri Izgu, Ayan Pal, and Jack W Szostak. Enhanced nonenzymatic RNA copying with 2-aminoimidazole activated nucleotides. *Journal of the American Chemical Society*, 139(5):1810–1813, 2017.

[73] Yingfu Li and Ronald R. Breaker. Kinetics of rna degradation by specific base catalysis of transesterification involving the 2'-hydroxyl group. *Journal of the American Chemical Society*, 121(23):5364–5372, 1999.

[74] Qiang Liu, Erik C Thorland, and Steve S Sommer. Inhibition of PCR amplification by a point mutation downstream of a primer. *Biotechniques*, 22(2):292–300, 1997.

[75] R. Lohrmann and L. E. Orgel. Prebiotic synthesis: Phosphorylation in aqueous solution. *Science*, 161(3836):64–66, 7 1968. publisher: American Association for the Advancement of Science.

[76] A. V. Lutay, E. L. Chernolovskaya, M. A. Zenkova, and V. V. Vlassov. The nonenzymatic template-directed ligation of oligonucleotides. *Biogeosciences Discussions*, 3(3):243–249, 6 2006.

[77] A. V. Lutay, Elena L. Chernolovskaya, Marina A. Zenkova, and V. V. Vlasov. Nonenzymatic template-dependent ligation of 2',3'-cyclic phosphate-containing oligonucleotides catalyzed by metal ions. *Doklady Biochemistry and Biophysics*, 401(1):163–166, 3 2005.

[78] A. V. Lutay, Marina A. Zenkova, and Valentin V. Vlassov. Nonenzymatic recombination of rna: possible mechanism for the formation of novel sequences. *Chemistry Biodiversity*, 4(4):762–767, 4 2007.

[79] Mikhail Makarov, Alma C Sanchez Rocha, Robin Krystufek, Ivan Cherepashuk, Volha Dzmitruk, Tatsiana Charnavets, Anneliese M Faustino, Michal Lebl, Kosuke Fujishima, Stephen D Fried, et al. Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. *Journal of the American Chemical Society*, 145(9):5320–5329, 2023.

[80] Angelica Mariani, Claudia Bonfio, Christopher M Johnson, and John D Sutherland. pH-Driven RNA strand separation under prebiotically plausible conditions. *Biochemistry*, 57(45):6382–6386, 2018.

[81] T Matreux, Kristian Le Vay, A Schmid, P Aikkila, L Belohlavek, AZ Çalışkanoğlu, E Salibi, A Kühnlein, C Springsklee, B Scheu, et al. Heat flows in rock cracks naturally optimize salt compositions for ribozymes. *Nature Chemistry*, 13(11):1038–1045, 2021.

[82] Jean-Louis Mergny and Laurent Lacroix. Analysis of thermal melting curves. *Oligonucleotides*, 13(6):515–537, 2003.

[83] Shin Miyakawa and James P. Ferris. Sequence- and regioselectivity in the montmorillonite-catalyzed synthesis of RNA. *Journal of the American Chemical Society*, 125(27):8202–8208, 2003.

[84] Shin Miyakawa and James P Ferris. Sequence- and regioselectivity in the montmorillonite-catalyzed synthesis of RNA. *Journal of the American Chemical Society*, 125(27):8202–8208, 2003.

[85] Ryo Mizuuchi, Alex Blokhuis, Lena Vincent, Philippe Nghe, Niles Lehman, and David Baum. Mineral surfaces select for longer RNA molecules. *Chemical communications*, 55(14):2090–2093, 2019.

[86] Matthias Morasch, Dieter Braun, and Christof B Mast. Heat-flow-driven oligonucleotide gelation separates single-base differences. *Angewandte Chemie*, 128(23):6788–6791, 2016.

[87] Matthias Morasch, Jonathan Liu, Christina F Dirscherl, Alan Ianeselli, Alexandra Kühnlein, Kristian Le Vay, Philipp Schwintek, Saidul Islam, Mérina K Corpinot, Bettina Scheu, et al. Heated gas bubbles enrich, crystallize, dry, phosphorylate and encapsulate prebiotic molecules. *Nature Chemistry*, 11(9):779–788, 2019.

[88] Matthias Morasch, Christof B Mast, Johannes K Langer, Pierre Schilcher, and Dieter Braun. Dry polymerization of 3, 5-cyclic GMP to long strands of RNA. *ChemBioChem*, 15(6):879–883, 2014.

[89] Keiji Murayama, Hikari Okita, Takumi Kuriki, and Hiroyuki Asanuma. Nonenzymatic polymerase-like template-directed synthesis of acyclic l-threoninol nucleic acid. *Nature Communications*, 12(1):804, 2021.

[90] Hannes Mutschler and Philipp Holliger. Non-canonical 3-5 extension of rna with prebiotically plausible ribonucleoside 2,3-cyclic phosphates. *Journal of the American Chemical Society*, 136(14):5193–5196, 4 2014.

[91] Hannes Mutschler, Alexander I. Taylor, Benjamin T. Porebski, Alice Lightowlers, Gillian Houlihan, Mikhail Abramov, Piet Herdewijn, and Philipp Holliger. Random-sequence genetic oligomer pools display an innate potential for ligation and recombination. *eLife*, 7, 11 2018.

[92] Hannes Mutschler, Aniela Wochner, and Philipp Holliger. Freeze-thaw cycles as drivers of complex ribozyme assembly. *Nature chemistry*, 7(6):502–508, 2015.

[93] Michi Nakata, Giuliano Zanchetta, Brandon D Chapman, Christopher D Jones, Julie O Cross, Ronald Pindak, Tommaso Bellini, and Noel A Clark. End-to-end stacking and liquid crystal condensation of 6–to 20–base pair DNA duplexes. *Science*, 318(5854):1276–1279, 2007.

[94] Dan T Nguyen and Omar A Saleh. Tuning phase and aging of DNA hydrogels through molecular design. *Soft Matter*, 13(32):5421–5427, 2017.

[95] Harry F Noller. The driving force for molecular evolution of translation. *RNA*, 10(12):1833–1837, 2004.

[96] Harry F Noller. Evolution of protein synthesis from an rna world. *Cold Spring Harbor perspectives in biology*, 4(4):a003681, 2012.

[97] Yasuhiro Oba, Yoshinori Takano, Hiroshi Naraoka, Naoki Watanabe, and Akira Kouchi. Nucleobase synthesis in interstellar ices. *Nature Communications*, 10(1):8–15, 2019.

[98] Norio Ogata and Hirofumi Morino. Elongation of repetitive DNA by DNA polymerase from a hyperthermophilic bacterium Thermus thermophilus. *Nucleic Acids Research*, 28(20):3999–4004, 10 2000.

[99] Carlos G Oliver, Vladimir Reinharz, and Jérôme Waldispühl. On the emergence of structural complexity in RNA replicators. *Rna*, 25(12):1579–1591, 2019.

[100] Jessica L O'Rear, Shenglong Wang, Andrew L Feig, Leonid Beigelman, Olke C Uhlenbeck, and Daniel Herschlag. Comparison of the hammerhead cleavage reactions stimulated by monovalent and divalent cations. *RNA*, 7(4):537–545, 2001.

[101] Andrey R Pavlov, Nadejda V Pavlova, Sergei A Kozyavkin, and Alexei I Slesarev. Recent developments in the optimization of thermostable DNA polymerases for efficient applications. *Trends in biotechnology*, 22(5):253–260, 2004.

[102] Seng-Meng Phang, Chai-Yaw Teo, Evelyn Lo, and Victor Wong Thi Wong. Cloning and complete sequence of the DNA polymerase-encoding gene (BstpolI) and characterisation of the Klenow-like fragment from Bacillus stearothermophilus. *Gene*, 163(1):65–68, 1995.

[103] Matthew W. Powner, Béatrice Gerland, and John D. Sutherland. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature*, 459(7244):239–242, 5 2009. publisher: Nature Publishing Group.

[104] Noam Prywes, J Craig Blain, Francesca Del Frate, and Jack W Szostak. Nonenzymatic copying of RNA templates containing all four letters is catalyzed by activated oligonucleotides. *eLife*, 5:e17756, 6 2016.

[105] Anna Pyle. Metal ions in the structure and function of rna. *JBIC Journal of Biological Inorganic Chemistry*, 7:679–690, 2002.

[106] Jifeng Qian, Tanya M Ferguson, Deepali N Shinde, Alissa J Ramírez-Borrero, Arend Hintze, Christoph Adami, and Angelika Niemz. Sequence dependence of isothermal DNA amplification via EXPAR. *Nucleic Acids Research*, 40(11):e87–e87, 2012.

[107] M. Renz, R. Lohrmann, and L.E. Orgel. Catalysts for the polymerization of adenosine cyclic 2,3-phosphate on a poly (u) template. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis*, 240(4):463–471, 7 1971.

[108] Lluís Ribas de Pouplana, Patrick Forterre, Jonathan Filée, and Hannu Myllykallio. Origin and evolution of DNA and DNA replication machineries. *The genetic code and the origin of life*, pages 145–168, 2004.

[109] Michael P Robertson and Gerald F Joyce. The origins of the rna world. *Cold Spring Harbor perspectives in biology*, 4(5):a003608, 2012.

[110] Rajat Rohatgi, David P Bartel, and Jack W Szostak. Nonenzymatic, template-directed ligation of oligoribonucleotides is highly regioselective for the formation of 3 '- 5 'phosphodiester bonds. *Journal of the American Chemical Society*, 118(14):3340–3344, 1996.

[111] Joachim H Rosenberger, Tobias Göppel, Patrick W Kudella, Dieter Braun, Ulrich Gerland, and Bernhard Altaner. Self-assembly of informational polymers by templated ligation. *Physical Review X*, 11(3):031055, 2021.

[112] Minik T. Rosing, Dennis K. Bird, Norman H. Sleep, and Christian J. Bjerrum. No climate paradox under the faint early sun. *Nature*, 464(7289):744–747, 4 2010.

[113] Michael J Russell, Roy M Daniel, and Allan J Hall. On the emergence of life via catalytic iron-sulphide membranes. *Terra Nova*, 5(4):343–347, 1993.

[114] Annalena Salditt, Leonie Karr, Elia Salibi, Kristian Le Vay, Dieter Braun, and Hannes Mutschler. Ribozyme-mediated RNA synthesis and replication in a model Hadean microenvironment. *Nature Communications*, 14(1):1495, 2023.

[115] Annalena Salditt, Lorenz M. R. Keil, David P. Horning, Christof B. Mast, Gerald F. Joyce, and Dieter Braun. Thermal Habitat for RNA Amplification and Accumulation. *Phys. Rev. Lett.*, 125:048104, Jul 2020.

[116] Annalena Salditt, Lorenz MR Keil, David P Horning, Christof B Mast, Gerald F Joyce, and Dieter Braun. Thermal habitat for rna amplification and accumulation. *Physical review letters*, 125(4):048104, 2020.

[117] Omar A Saleh, Byoung-jin Jeon, and Tim Liedl. Enzymatic degradation of liquid droplets of DNA is modulated near the phase boundary.

[118] John SantaLucia Jr. A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, 95(4):1460–1465, 1998.

[119] Erwin Schrodinger. What is life?: the physical aspect of the living cell. 1946.

[120] David M Shechner, Robert A. Grant, Sarah C. Bagby, Yelena Koldobskaya, Joseph A. Piccirilli, and David P. Bartel. Crystal structure of the catalytic core of an RNA-polymerase ribozyme. *Science*, 326(5957):1271–1275, 11 2009.

[121] J. Sheng, L. Li, A. E. Engelhart, J. H. Gan, J. W. Wang, and J. W. Szostak. Structural insights into the effects of 2 '-5 ' linkages on the RNA duplex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(8):3050–3055, 2 2014.

[122] Friedrich C Simmel, Bernard Yurke, and Hari R Singh. Principles and applications of nucleic acid strand displacement reactions. *Chemical reviews*, 119(10):6326–6369, 2019.

[123] Emilie Yeonwha Song, Eddy I. Jiménez, Huacan Lin, Kristian Le Vay, Ramanarayanan Krishnamurthy, and Hannes Mutschler. Prebiotically Plausible RNA Activation Compatible with Ribozyme-Catalyzed Ligation. *Angewandte Chemie*, 60(6):2952–2957, 2020.

[124] Marilyne Sosson, Daniel Pfeffer, and Clemens Richert. Enzyme-free ligation of dimers and trimers to rna primers. *Nucleic Acids Research*, 47(8):3836–3845, 5 2019.

[125] Ralph Stadhouders, Suzan D Pas, Jeer Anber, Jolanda Voermans, Ted HM Mes, and Martin Schutten. The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5 nuclease assay. *The Journal of Molecular Diagnostics*, 12(1):109–117, 2010.

[126] Hannah S Steinert, Jörg Rinnenthal, and Harald Schwalbe. Individual basepair stability of DNA and RNA studied by NMR-detected solvent exchange. *Biophysical journal*, 102(11):2564–2574, 2012.

[127] Michael Stich, Carlos Briones, and Susanna C. Manrubia. On the structural repertoire of pools of short, random RNA sequences. *Journal of Theoretical Biology*, 252(4):750–763, 2008.

[128] Daekyu Sun and Laurence H Hurley. The importance of negative superhelicity in inducing the formation of g-quadruplex and i-motif structures in the c-myc promoter: implications for drug targeting and control of gene expression. *Journal of medicinal chemistry*, 52(9):2863–2874, 2009.

[129] N Kyle Tanner. Ribozymes: the characteristics and properties of catalytic RNAs. *FEMS microbiology reviews*, 23(3):257–275, 1999.

[130] Pallavi Thaplyal and Philip C Bevilacqua. Experimental approaches for measuring pKa's in RNA and DNA. In *Methods in enzymology*, volume 549, pages 189–219. Elsevier, 2014.

[131] Alexei V. Tkachenko and Sergei Maslov. Onset of natural selection in populations of autocatalytic heteropolymers. *Journal of Chemical Physics*, 149(13), 2018.

[132] Zoe R Todd, Albert C Fahrenbach, Sukrit Ranjan, Christopher J Magnani, Jack W Szostak, and Dimitar D Sasselov. Ultraviolet-driven deamination of cytidine ribonucleotides under planetary conditions. *Astrobiology*, 20(7):878–888, 2020.

[133] J. D. Toner and D. C. Catling. Alkaline lake settings for concentrated prebiotic cyanide and the origin of life. *Geochimica Et Cosmochimica Acta*, 260:124–132, 9 2019.

[134] Jonathan D. Toner and David C. Catling. A carbonate-rich lake solution to the phosphate problem of the origin of life. *Proceedings of the National Academy of Sciences of the United States of America*, 117(2):883–888, 1 2020.

[135] Andrew S Tupper and Paul G Higgs. Rolling-circle and strand-displacement mechanisms for non-enzymatic RNA replication at the time of the origin of life. *Journal of Theoretical Biology*, 527:110822, 2021.

[136] D. A. Usher and D. Yee. Geometry of the dry-state oligomerization of 2,3-cyclic phosphates. *Journal of Molecular Evolution*, 13(4):287–293, 11 1979.

[137] Peter TS Van der Gulik and Dave Speijer. How amino acids and peptides shaped the RNA world. *Life*, 5(1):230–246, 2015.

[138] M. S. Verlander and L. E. Orgel. Analysis of high molecular weight material from the polymerization of adenosine cyclic 2, 3-phosphate. *Journal of Molecular Evolution*, 3(2):115–120, 1974.

[139] MS Verlander, R Lohrmann, and LE Orgel. Catalysts for the self-polymerization of adenosine cyclic 2, 3-phosphate. *Journal of Molecular Evolution*, 2:303–316, 1973.

[140] Enrique Viguera, Danielle Canceill, and S Dusko Ehrlich. In vitro replication slippage by DNA polymerases from thermophilic organisms. *Journal of molecular biology*, 312(2):323–333, 2001.

[141] Alexander V. Vlassov, Brian H. Johnston, Laura F. Landweber, and Sergei A. Kazakov. Ligation activity of fragmented ribozymes in frozen solution: implications for the rna world. *Nucleic Acids Research*, 32(9):2966–2974, 5 2004.

[142] Stephanie R Vogel, Christopher Deck, and Clemens Richert. Accelerating chemical replication steps of RNA involving activated ribonucleotides and downstream-binding elements. *Chemical communications*, (39):4922–4924, 2005.

[143] Falk Wachowius and Philipp Holliger. Non-enzymatic assembly of a minimized rna polymerase ribozyme. *ChemSystemsChem*, 1:1–4, 2019.

[144] Marita Wasner, Dominique Arion, Gadi Borkow, Anne Noronha, Andre H. Uddin, Michael A. Parniak, and Masad J. Damha. Physicochemical and biochemical properties of 2',5'-linked rna and 2',5'-rna:3',5'-rna "hybrid" duplexes ,. *Biochemistry*, 37(20):7478–7486, 5 1998.

[145] N Watanabe, WMC Sameera, H Hidaka, A Miyazaki, and A Kouchi. Nucleobase synthesis in interstellar ices. *Chemical physics letters*, 737:136820, 2019.

[146] Brian T Wimberly, Ditlev E Brodersen, William M Clemons Jr, Robert J Morgan-Warren, Andrew P Carter, Clemens Vonrhein, Thomas Hartsch, and V Ramakrishnan. Structure of the 30s ribosomal subunit. *Nature*, 407(6802):327–339, 2000.

[147] Long-Fei Wu, Ziwei Liu, Samuel J Roberts, Meng Su, Jack W Szostak, and John D Sutherland. Template-free assembly of functional rnas by loop-closing ligation. *Journal of the American Chemical Society*, 144(30):13920–13927, 2022.

[148] Yongzheng Xing, Enjun Cheng, Yang Yang, Ping Chen, Tao Zhang, Yawei Sun, Zhongqiang Yang, and Dongsheng Liu. Self-assembled DNA hydrogels with designable thermal and enzymatic responsiveness. *Advanced Materials*, 23(9):1117–1121, 2011.

[149] Zhongyang Xing, Alessio Caciagli, Tianyang Cao, Iliya Stoev, Mykolas Zupkauskas, Thomas O'Neill, Tobias Wenzel, Robin Lamboll, Dongsheng Liu, and Erika Eiser. Microrheology of DNA hydrogels. *Proceedings of the National Academy of Sciences*, 115(32):8137–8142, 2018.

[150] Xiaojun Xu, Tao Yu, and Shi-Jie Chen. Understanding the kinetic mechanism of RNA single base pair formation. *Proceedings of the National Academy of Sciences*, 113(1):116–121, 2016.

[151] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*, 34(2):564–574, 2006.

[152] Hanyang Yu, Su Zhang, and John C Chaput. Darwinian evolution of an alternative genetic system provides support for tna as an rna progenitor. *Nature chemistry*, 4(3):183–187, 2012.

[153] Marat M Yusupov, Gulnara Zh Yusupova, Albion Baucom, Kate Lieberman, Thomas N Earnest, JHD Cate, and Harry F Noller. Crystal structure of the ribosome at 5.5 Å resolution. *science*, 292(5518):883–896, 2001.

[154] Joseph N Zadeh, Conrad D Steenberg, Justin S Bois, Brian R Wolfe, Marshall B Pierce, Asif R Khan, Robert M Dirks, and Niles A Pierce. Nupack: Analysis and design of nucleic acid systems. *Journal of computational chemistry*, 32(1):170–173, 2011.

[155] G Zanchetta, M Nakata, M Buscaglia, NA Clark, and T Bellini. Liquid crystal ordering of DNA and RNA oligomers with partially overlapping sequences. *Journal of Physics: Condensed Matter*, 20(49):494214, 2008.

[156] Ke Zhang, Joseph Hodge, Anamika Chatterjee, Tae Seok Moon, and Kimberly M. Parker. Duplex structure of double-stranded RNA provides stability against hydrolysis relative to single-stranded RNA. *Environmental Science and Technology*, 55(12):8045–8053, 2021.

[157] Stephanie J. Zhang, Daniel Duzdevich, Dian Ding, and Jack W. Szostak. Freeze-thaw cycles enable a prebiotically plausible and continuous pathway from nucleotide activation to nonenzymatic rna copying. *Proceedings of the National Academy of Sciences*, 119(17):e2116429119, 4 2022. publisher: Proceedings of the National Academy of Sciences.

[158] Lijun Zhou, Seohyun Chris Kim, Katherine H. Ho, Derek K. O'Flaherty, Constantin Giurgiu, Tom H. Wright, and Jack W. Szostak. Non-enzymatic primer extension with strand displacement. *eLife*, 8:1–14, 2019.

[159] Lijun Zhou, Derek K O'Flaherty, and Jack W Szostak. Template-directed copying of RNA by non-enzymatic ligation. *Angewandte Chemie*, 132(36):15812–15817, 2020.

[160] Lijun Zhou, Derek K O'Flaherty, and Jack W Szostak. Assembly of a ribozyme ligase from short oligomers by nonenzymatic ligation. *Journal of the American Chemical Society*, 142(37):15961–15965, 2020.

[161] Hubert Zipper, Herwig Brunner, Jürgen Bernhagen, and Frank Vitzthum. Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. *Nucleic acids research*, 32(12):e103–e103, 2004.

List of Publications

List of publications published during the course of the doctoral studies and associated (shared-) first author publications:

Giacomo Bartolucci*, **Adriana Calaça Serrão***, Philipp Schwintek*, Alexandra Kühnlein, Yash Rana, Philipp Janto, Dorothea Hofer, Cristof B. Mast and Dieter Braun; Sequence self-selection by cyclic phase separation, *Proceedings of the National Academy of Sciences* **(2023)**, doi:10.1073/pnas.2218876120

Adriana Calaça Serrão*, Felix Dänekamp*, Zsófia Meggyesi and Dieter Braun; Replication elongates short DNA, reduces sequence bias and develops trimer structure, *Nucleic Acids Research* **(2024)**, doi: 10.1093/nar/gkad1190

Adriana Calaça Serrão*, Sreekar Wunnava*, Avinash V. Dass, Lennard Ufer, Philipp Schwintek, Christof B. Mast and Dieter Braun; High-Fidelity RNA Copying via 2′, 3′-Cyclic Phosphate Ligation, *Journal of the American Chemical Society* **(2024)**, doi: 10.1021/jacs.3c10813 (accepted)

* equal contribution

PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Sequence self-selection by cyclic phase separation

Giacomo Bartolucci[a,b,1], Adriana Calaça Serrão[c,1], Philipp Schwintek[c,1], Alexandra Kühnlein[c], Yash Rana[a,b], Philipp Janto[c], Dorothea Hofer[c], Christof B. Mast[c], Dieter Braun[c,2], and Christoph A. Weber[d,2]

Edited by Dimitar D. Sasselov, Harvard-Smithsonian Center for Astrophysics, Cambridge, MA; received November 8, 2022; accepted September 6, 2023 by Editorial Board Member Herbert Levine

The emergence of functional oligonucleotides on early Earth required a molecular selection mechanism to screen for specific sequences with prebiotic functions. Cyclic processes such as daily temperature oscillations were ubiquitous in this environment and could trigger oligonucleotide phase separation. Here, we propose sequence selection based on phase separation cycles realized through sedimentation in a system subjected to the feeding of oligonucleotides. Using theory and experiments with DNA, we show sequence-specific enrichment in the sedimented dense phase, in particular of short 22-mer DNA sequences. The underlying mechanism selects for complementarity, as it enriches sequences that tightly interact in the dense phase through base-pairing. Our mechanism also enables initially weakly biased pools to enhance their sequence bias or to replace the previously most abundant sequences as the cycles progress. Our findings provide an example of a selection mechanism that may have eased screening for auto-catalytic self-replicating oligonucleotides.

molecular selection | phase separation | prebiotic oligonucleotides | molecular origin of life | DNA

Oligonucleotides can catalyse biochemical reactions and store genetic information (1–3). The mechanisms through which functional oligonucleotides became sufficiently abundant are crucial to understanding the molecular origin of life (4). In addition to sequence motifs, sufficient strand length is also a requirement for functional folds (5). Therefore, the assembly of long-chained prebiotic oligonucleotides has been the focus of many recent studies (6–9).

However, the probability of randomly assembling a specific sequence of length L with m different nucleotides is proportional to m^{-L}. Since sequence space grows exponentially with sequence length, functional sequences are either not present or too dilute to interact and undergo chemical reactions. It is thus one of the central mysteries of the molecular origin of life how long enough sequences that enable self-replication could be selected from a large random pool of short non-functional oligonucleotides.

Due to the lack of complex biological machinery at the molecular origin of life, various physicochemical selective mechanisms have been considered (10). Examples are biased replication (11, 12), accumulation due to gradients of temperature or salt (13), the accumulation at liquid-vapor interfaces (14), as well as the length selective accumulation at mineral surfaces (15, 16). Multiple of the aforementioned mechanisms may also act synergistically. Another possible mechanism is related to the coexistence of two liquid-like phases. In particular, recent studies have shown that oligonucleotides can phase separate, forming coacervates (17, 18), liquid crystals (19, 20), or hydrogels (21–23), which can lead to a local enrichment of specific oligonucleotides.

An especially elegant case emerges when phase separation is caused directly by the base pairing of sequence segments among oligonucleotides. The strong interactions among complementary oligonucleotide strands (approximately $5\,k_bT$ per base pair) lead to low saturation concentration above which phase separation occurs (24) and an oligonucleotide-dense phase that is composed of strongly correlated sequences. Thus, oligonucleotide phase separation via base pairing provides a mechanism to strongly accumulate a specific set of oligonucleotide sequences.

In a realistic prebiotic environment, such as an under-water rocky pore (14, 25, 26), the phase separation of oligonucleotides can be expected to be subject to periodic, typically daily, changes in the environment. In addition, such systems can exchange oligomers with the environment continuously, for example, by fluid flows (Fig. 1*A*). Without phase separation, the oligomer composition is set by the composition of the environment. However, when the oligomers can phase separate, the oligomer-dense phase can grow by continuously recruiting sequences from the pool. Initially, we focus on periodic flows composed of short spikes separated by long waiting times in which the flow vanishes.

Significance

A central mystery of the molecular origin of life is the emergence of oligonucleotides, such as RNA, that can self-replicate. In our work, we theoretically study and experimentally verify a simple though minimal mechanism capable of screening for such specific oligonucleotide sequences. This mechanism relies on two physical ingredients ubiquitous on early Earth: cycles of phase separation into oligonucleotide-dense and dilute phases and cyclic oligonucleotide exchange with a surrounding pool. We show that specific sequences can enrich oligonucleotide composition, evolving away from an initial pool. This non-equilibrium selection mechanism may provide the missing link in how specific short-chained peptides, RNA, and DNA sequences were recruited from prebiotic pools steering the assembly of self-replicating oligonucleotides at the molecular origin of life.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.D.S. is a guest editor invited by the Editorial Board.

[1] G.B, A.C.S, and P.S. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: dieter.braun@lmu.de or christoph.weber@physik.uni-augsburg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2218876120/-/DCSupplemental.

Published October 17, 2023.

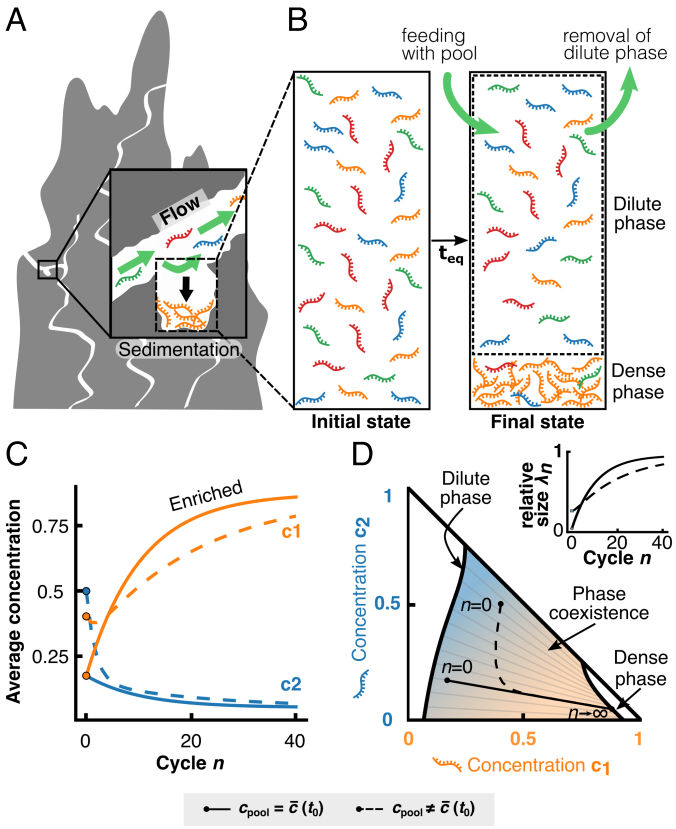


Fig. 1. Sequence selection via phase separation under varying feeding conditions. (*A*) Illustration of a time-dependent oligonucleotide flow through porous rocks on early earth. By sedimentation, phase-separated sequences can be enriched in pores, while others are flushed away by the flow. In the case of intermittent flows that are non-zero for periods much shorter than t_{eq} (time to reach phase equilibrium), separated by waiting times much longer than t_{eq}, this scenario of sequence exchange can be mimicked via cycles of phase separation. (*B*) At each cycle, the system phase separates into a dense and a dilute phase, then a fraction of the dilute phase is removed and replaced with samples coming from a fixed sequence pool. (*C*) Theoretical result for the selection kinetics of multiple cycles depicted in *B*, discussed in systems composed of solvent and two sequences 1 and 2 with concentrations c_1 (orange) and c_2 (blue), respectively, see Eq. **1**. Solid and dashed lines correspond to α=0.75 and α=0.25, where α is the fraction of the dilute phase that gets replaced with the pool at each cycle. The two cases displayed differ also in initial average concentration $\bar{\boldsymbol{c}}(t_0)$ that can be equal to the pool ($\boldsymbol{c}_{pool} = \bar{\boldsymbol{c}}(t_0)$, solid lines) or deviate from it ($\boldsymbol{c}_{pool} \neq \bar{\boldsymbol{c}}(t_0)$, dashed lines). (*D*) The selection kinetics can be represented as a trajectory (solid, dashed black line) in the corresponding phase diagram, where outer white regions correspond to homogeneous mixtures while phase separation occurs in the coloured area. Gray tie lines connect the concentrations of the coexisting phases. During the kinetics, the dense phase grows as indicated by an increase of its relative size λ_n (*Inset*).

The duration of the spikes is short compared to the time to reach phase equilibrium t_{eq} while the waiting time between spikes is large compared to t_{eq}. In this case, we can mimic the continuous exchange with the pool by a simplified cyclic protocol composed of two different steps: i) A feeding step that corresponds to the replacement of a part of the dilute phase by the pool, followed by ii) a relaxation period toward phase separation equilibrium; see Fig. 1*B*. Later, we examine the opposite regime, i.e., where the flow is constant in time, see *DNA Phase Separation in a Continuous Feeding Flow*.

In this study, we ask whether this recruitment can significantly alter the oligonucleotide composition in the pore and thereby provide a physical mechanism of selection of specific sequences. The key question is how much the oligonucleotide composition of the system can evolve away from the pool, which serves as a reference for the selection process. We investigate this question through theory and experiments using DNA as a model oligomer. We decided to use DNA instead of RNA since DNA is more stable and our study focuses on a selection mechanism that relies on base pairing, which is very similar for both (27). We show that phase separation combined with feeding cycles by a pool indeed provides a strong selection mechanism giving rise to distinct routes for molecular selection of specific oligonucleotide sequences.

Results and Discussion

Theory of Cyclic Phase Separation with Feeding. Here, we first discuss the theory governing the kinetics of an oligonucleotide mixture of volume V which is composed of M different sequences. This system undergoes cycles alternating between i) a period where the material is exchanged with a pool of fixed composition $\boldsymbol{c}_{pool}$, and ii) a period of phase separation (Fig. 1*B*). Specifically, within (ii), the mixture phase separates into a dense and a dilute phase with sufficient time to relax to thermodynamic phase equilibrium, while during the feeding step (i), a fraction α of the dilute phase is replaced by the pool. After n cycles, the average composition of the system is described by the M-dimensional vector, $\bar{\boldsymbol{c}}(t_n) = \lambda(t_n)\boldsymbol{c}^{\mathrm{I}}(t_n) + (1-\lambda(t_n))\,\boldsymbol{c}^{\mathrm{II}}(t_n)$, where the vector components are the average concentrations of sequences. Moreover, $\boldsymbol{c}^{\mathrm{I}}(t_n)$ and $\boldsymbol{c}^{\mathrm{II}}(t_n)$ denote the concentrations of the dense and dilute phase, respectively, and $\lambda(t_n) = V^{\mathrm{I}}(t_n)/V$ is the fraction of the system occupied by the dense phase, where $V^{\mathrm{I}}(t_n)$ denotes the volume of the dense phase. The average composition of the system changes with cycle time t_n due to the feeding step (i) and is given by (see *SI Appendix*, section 1 for more information):

$$\bar{\boldsymbol{c}}(t_{n+1}) = \Big[\big(1-\lambda(t_n)\big)\Big(\alpha\boldsymbol{c}_{\mathrm{pool}} + (1-\alpha)\boldsymbol{c}^{\mathrm{II}}(t_n)\Big) + \lambda(t_n)\boldsymbol{c}^{\mathrm{I}}(t_n)\Big], \qquad [1]$$

where $\boldsymbol{c}_{pool}$ is the concentration vector characterizing the composition of the pool that remains constant in time. The fraction of the dense phase $\lambda(t_n)$, and the concentrations of the dense and dilute phase $\boldsymbol{c}^{\mathrm{I}}(t_n)$ and $\boldsymbol{c}^{\mathrm{II}}(t_n)$ at cycle time t_n can be determined by a Maxwell's construction in a M-dimensional state space for $\bar{\boldsymbol{c}}(t_n)$ obtained from solving the iteration Eq. **1**. The construction amounts to solving a set of non-linear equations that describe the balance of the chemical potentials and the osmotic pressure between the phases. Their solution requires estimating the sequence-specific interactions among the different oligonucleotides. Details on the numerical method and the determination of interaction matrices are discussed in *SI Appendix*, section 2.

To study the selection of oligonucleotide by cyclic phase separation, we considered the exchange of the oligonucleotide-poor phase by a pool of constant composition $\boldsymbol{c}_{pool}$, where the pool acted as a reference for the selection kinetics. We studied two cases where the pool had the same composition as the initial average concentration at $t = t_0$, $\boldsymbol{c}_{pool} = \bar{\boldsymbol{c}}(t_0)$, or they differed in terms of composition, $\boldsymbol{c}_{pool} \neq \bar{\boldsymbol{c}}(t_0)$. Representative time traces for both cases are shown in Fig. 1*C* for a mixture composed of solvent and two oligonucleotide sequences. We find that for both cases one sequence is enriched while the other sequence decreases in concentration as cycles proceed (orange and blue dashed lines, respectively). These concentration traces can be represented as trajectories of the average concentrations $\bar{\boldsymbol{c}}(t_n)$ in the ternary

phase diagram; Fig 1*D*. For average concentrations in the white region of the phase diagram the system remains homogeneous, while in the coloured region, phase separation occurs. In light gray, we show the tie lines connecting the coexisting dense and dilute phase concentrations. Each average concentration on the trajectory within this coloured region leads to a unique fraction of the dense phase, $\lambda(t_n)$, and a pair of concentrations corresponding to the dense and dilute phase, $\boldsymbol{c}^{\mathrm{I}}(t_n)$ and $\boldsymbol{c}^{\mathrm{II}}(t_n)$, respectively. As cycles proceed, the fraction of the dense phase $\lambda(t_n)$ grows (Fig. 1 *D*, *Inset*). During this growth, sequence 1 is selected over sequence 2. As volume growth saturates at $\lambda(t_\infty) = 1$, the selection process stops and the system settles in a stationary state.

The sequence composition of this stationary state is set by the tie line defined by the pool composition $\boldsymbol{c}_{\mathrm{pool}}$ (straight solid black line in Fig 1*D*). Most importantly, the slope and length of this pool tie line determine if and how well sequences are selected. Only if the pool tie line deviates from the diagonal tie line in the phase diagram, there is sequence selection during the growth of the dense phase. Selection is more pronounced if the pool tie line is longer in the phase diagram since more volume growth can occur. This case can be realized for example by strong interactions among sequences leading to the dilute and dense binodal branches being far apart in the phase diagram. Strikingly, both conditions, non-diagonal pool tie lines and strong sequence interactions, are particularly fulfilled in mixtures of oligonucleotides that can interact via base pairing. From our theoretical study, we conclude that phase separation subject to cyclic feeding can provide a selection mechanism particularly relevant in oligonucleotides mixtures.

Observation of a Dense Phase of DNA. To experimentally scrutinize the prerequisite of our proposed selection mechanism, we investigate phase-separation in oligonucleotide mixtures. To this end, we have designed several experimental systems. The constructs were motivated by the theoretical model which suggested that a group of sequences were selected if they interacted strongly among themselves but weakly with other sequences. The three sequence pairs (1 to 3) are composed of sequences with three binding regions each (a, b, and c or a', b', and c') which are separated by dimeric spacer sequences. These segments are individually reverse-complementary, i.e., a' is the reverse complement of a. However, the sequences i and ii are not the reverse complement of each other, because the order of the individual segments is not reversed (since i = 5' a,b,c 3' and ii = 5' a',b',c' 3'). This choice avoids a fully complementary double-stranded structure and allows each sequence to bind to three other sequences, forming a branched structure (Fig. 2 *A* and *B*). See *SI Appendix*, section 4 for a simulation of the secondary structure of each of the sequence pairs with NUPACK. The formation of branched DNA aggregates with short sequences leads to a dense phase. In fact, it was previously shown that mixtures of oligonucleotides are able to form dense phases through mutual base-pairing between long strands (14, 18, 21, 22, 28–30). However, these studies investigated rather long strands and solely ref. (31) studied phase separation of short DNA strands in the length regime of 20 to 25 bp which are more likely in prebiotic soups.

To characterize the phase separation propensity of our designed systems, we used time-lapse fluorescence microscopy (*SI Appendix*, section 13). In particular, we imaged each system over time in thin temperature-controlled microfluidic chambers. For each sequence pair, both strands were at 25 μM. Salt

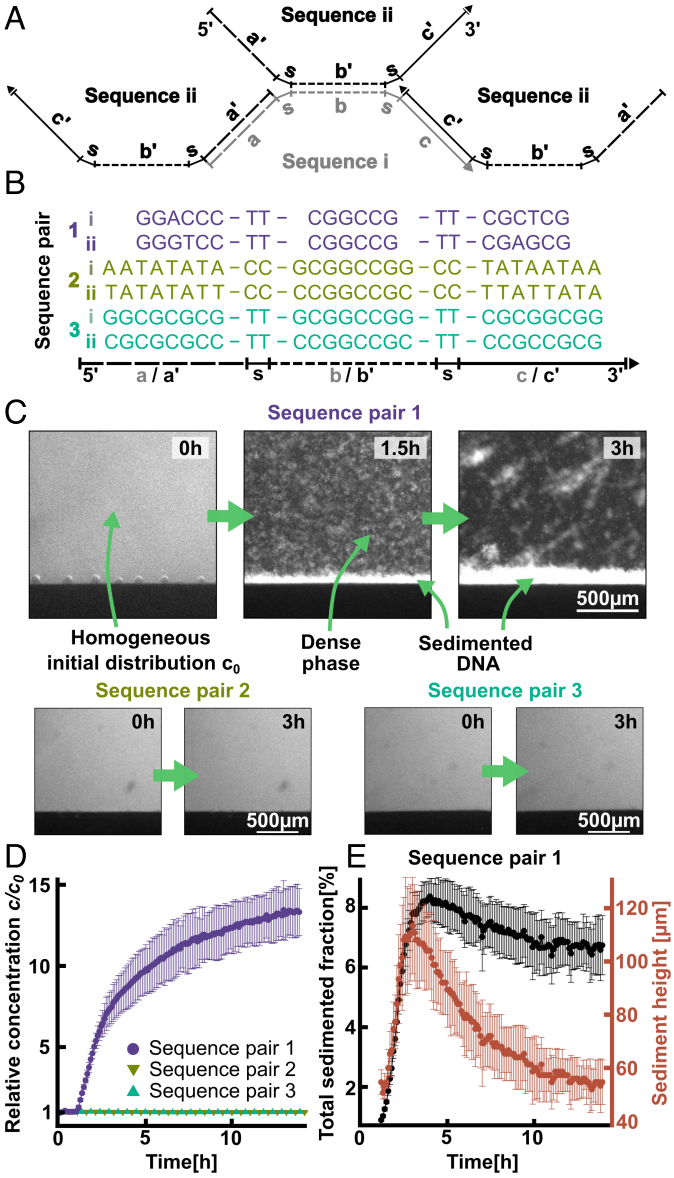


Fig. 2. Phase separation and sedimentation behavior of three sequence pairs. (*A* and *B*) Sequence i is composed of three segments a, b, and c with spacer s. Its pair ii consists of reverse complements a', b' and c'. The inverted arrangement of a' and b' creates a network from three binding sites and prevents the formation of a linear double-stranded duplex. (*C*) Fluorescence time laps images in a vertical, 500 μm thin microfluidic chamber to prevent convection flow. Concentrations of strands were 25 μM in a buffer of 10 mM Tris-HCl pH 7, 10 mM $MgCl_2$ and 125 mM NaCl. Fluorescence labelling was provided by 5X SYBR Green I. After cooling from 65 °C to 15 °C, sequence pair 1 phase separated and sedimented to the bottom of the chamber. Sequence pairs 2 and 3 did not form a dense phase, and thus showed a homogeneous fluorescence signal. (*D*) Sequence pair 1 showed an up to 13-fold enhanced relative concentration while sequence pairs 2 and 3 showed no phase separation. (*E*) The sedimentation behaviour of sequence pair 1 is studied by measuring its fluorescence. The total amount of sedimented DNA plateaued at 6 to 8% after 5 h while the sedimented DNA contracted about twofold. The sticking of dense phase DNA to the chamber walls could not be fully prevented. Error bars are SDs of three independent experiments.

concentrations were fixed at 125 mM NaCl and 10 mM $MgCl_2$, pH was controlled using 10 mM TRIS pH 7 buffer, and 5X SYBR Green I was used for fluorescent labelling. Other buffer conditions were also screened (*SI Appendix*, section 6). Prior to each experiment, the solutions were heated inside the microfluidic chamber to 65 °C to ensure homogeneous initial conditions. The samples were then slowly cooled at a rate of 6 K/min to 15 °C and

incubated at that temperature for at least 3 h. This temperature is lower than the melting temperature for all of the sequence pairs, corresponding to a scenario where most of the sequences are bound (*SI Appendix*, section 8). Choosing a higher incubation temperature resulted in smaller aggregates in bulk (*SI Appendix*, section 9 as well as Videos 4 and 5, video description).

Microscope images for all three systems are shown in Fig. 2*C*, where "0 h" corresponds to the moment when the cooling step has reduced the temperature to 15 °C. Within about 10 min, the first dense phase DNA nucleated for sequence pair 1 (Video Description) grew within 1.5 h to a size of a few micrometers and sedimented at speeds between 0.1 to 2.5 μm/s (*SI Appendix*, section 9). As a result, a phase of sedimented DNA accumulated at the bottom of the chamber. Assuming the particles to behave like sinking spheres subjected to Stokes drag, we estimated their densities to be around 3% higher than water. We could thus determine the characteristic sedimentation length to be on the scale of tens of micrometers. For sedimentation to occur, the chamber (pore) height has to exceed the sedimentation length. Inverting this logic, the minimal particle radius for our setup to observe sedimentation would be 0.3 μm. The DNA concentration in the dense phase increased up to 13-fold (Fig. 2*D*).

The total amount of molecules that sedimented saturates at about 8% of the initial material at about 3 h, decreasing then only slightly over time (Fig. 2*E*, black data points). The height of sedimented DNA reached a maximum of about 100 μm at 3 h but then compacted to about half the height (Fig. 2*E*, red data points). Similar behavior has been observed in literature for systems composed of longer DNA strands (21).

No dense phase was observed for sequence pairs 2 or 3, despite their longer segments that suggest stronger binding affinities. This could be caused by non-specific hybridization tendencies of binary (especially G- and C-rich) sequences, which can lead to alternative secondary structures, such as hairpin-rich configurations, internal loops or G-quadruplexes (32). Consequently, the formation of such structures may prevent network formation. We also examined five additional 22-nt sequence pairs similar to pair 1, systematically varying the base composition of the GC-only binding segments. Notably, the inclusion of a single A/T nucleotide in the outer segments a/a' or c/c', thus utilizing the full 4-letter alphabet, proved crucial for phase separation (see *SI Appendix*, section 5 for more details).

Cycles of Phase Separation and Feeding. Based on our observation that sequence pair 1 can form a dense DNA-rich phase, we experimentally scrutinize the theoretically proposed selection mechanism shown in Fig. 1 *A–D* that relied on a cyclic material inflow. We subjected the phase-separating DNA to cyclic feeding steps by replacing the *Top* fraction of the dilute phase in the vial with the pool (Fig. 3*A*, steps 1 to 3). The theory suggests that exchanging the complete dilute phase with the pool reaches the final stationary state with minimal amount of cycles (*SI Appendix*, section 1). However, a complete removal of the dilute phase by pipetting turns out to be experimentally difficult since this also risks removing sedimented dense DNA. To avoid kinetically trapped states of phase-separating oligonucleotides, we additionally include annealing and melting steps in the cycle (Fig. 3*A*, steps 3 to 4 and back to step 1). This procedure enabled a fast relaxation to thermodynamic equilibrium after each feeding step.

We investigated a system composed of equal fractions of the sequence pairs 1 and 2 (Fig. 2*B*), where solely the sequence pair 1 showed phase separation before. As control we considered a system composed of sequence pairs 2 and 3 where we could not observe the formation of a dense DNA phase (Fig. 2*B*). We determined the strand concentrations of each system in the *Top* and *Bottom* fractions of the vial using HPLC. We monitored the kinetics over six feeding cycles for the system composed of sequence pairs 1 and 2 (Fig. 3*B*) and compare it to the non-phase-separating control (*Inset* in *B*). Both systems were initialized with equimolar concentrations of the two respective sequence pairs.

We found that the concentrations for the control hardly increased per cycle with slopes about or less than 2 μM/cycle.

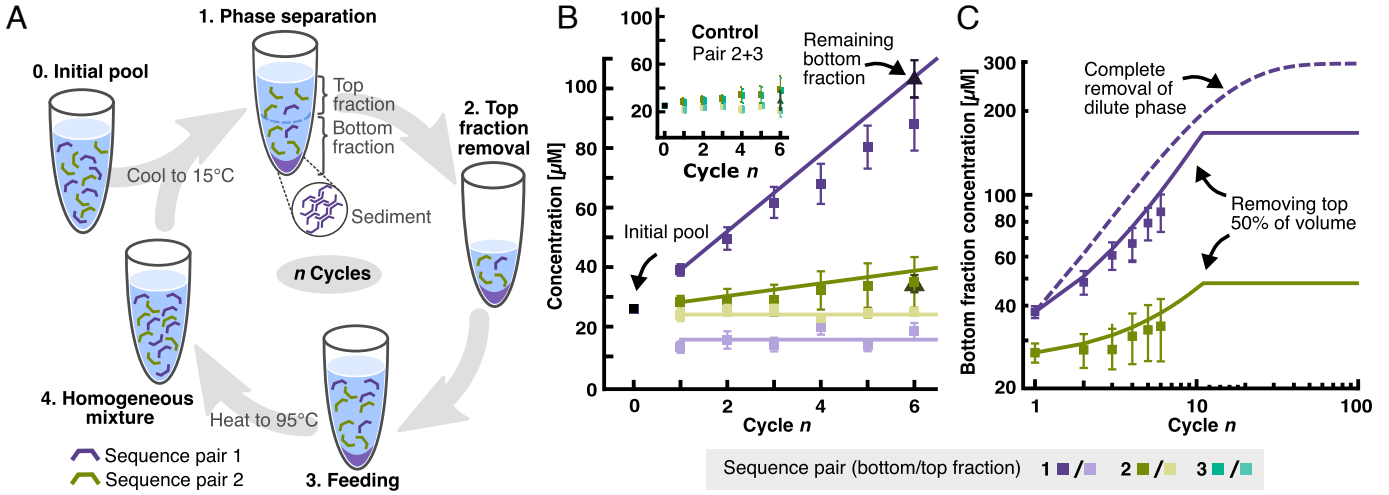


Fig. 3. Cycles of phase separation and feeding steps select specific oligonucleotide sequences from the initial pool. (*A*) Cyclic experimental protocol based on Fig. 1*B*. (*B*) The initial pool contained a 25 μM concentration of sequences pairs 1 and 2. After sedimentation, the top half of the volume (Top fraction) was removed ($\alpha = 0.5$) and fed after each cycle with the same volume of the initial pool. Using quantification with HPLC, we found that sequence pair 1 (purple) was enriched while the concentration of sequence pair 2 (green) remained approximately constant. The same flat dynamics was found for a control system using non-phase-separating sequence pairs 2 and 3 (inset, due to significant data overlap markers are halved). In addition, the concentration of all supernatants and the final sediment were measured by absorbance at 260 nm (triangular markers; *SI Appendix*, section 10). (*C*) Solid lines show theoretical predictions. The *Bottom* fraction concentration saturates once the sedimented DNA has filled the *Bottom* fraction of the chamber. If the whole supernatant could have been removed at each step of the cycle, only slightly amplified selection would be predicted (dashed line).

This non-zero increase is probably due to the adhesion of strands to the vial surface. For the phase-separating system with sequence pair 1 we observed that the concentration strongly increased, approximately linear with a slope of about (10.2 ± 0.4) μM/cycle (purple), while the sequence pair 2 in the mixture got only weakly enriched by the cycling. This observation confirmed that specific sequences could get selected by phase separation from the dilute phase.

Additionally, we also tested whether SYBR Green I changes the general behavior of the sequence pair mixtures over feeding cycles. In its absence, the partitioning of sequence pair 1 to the *Bottom* fraction also occurred, while sequence pair 2 remained constant on both fractions. Interestingly, the linear increase in concentration on the *Bottom* fraction is about 1.6 times higher than in the presence of SYBR Green I but shows very similar results compared to the experiments with SYBR (*SI Appendix*, section 11). However, in order to match theory parameters based on fluorescence measurements, we kept the SYBR concentration constant throughout all experiments.

In the experiments, the selection occurred concomitantly with the growth of the dense phase, which is consistent with our theoretical results. As cycles proceeded, more phase-separating DNA was recruited and led to an increase in the concentration of the *Bottom* fraction (Fig. 3*C*). In contrast, the concentration of the *Top* fraction remained constant at about 14.7 μM (Fig. 3*C*, light purple). A constant *Top* fraction concentration during cycles implies that the system remained on the same tie line while the volume of the sedimented DNA was growing. This corresponds to the simple theoretical scenario where the system is initialized at the pool tie line, as outlined in Fig. 1*D* ($c(t_0) = c_{\text{pool}}$).

We quantitatively compared the experimental results for the *Bottom* fraction concentration with the theoretical model. Since the experimental selection kinetics occurred on a single tie line, the dense and dilute phase concentrations, $\boldsymbol{c}^{\text{I}}$ and $\boldsymbol{c}^{\text{II}}$ remained constant over time, while the sediment size increases. For the dilute phase concentration $\boldsymbol{c}^{\text{II}}$, we use the experimental concentration value of the *Top* fraction. The sediment concentration $\boldsymbol{c}^{\text{I}}$ could be estimated for the theory using the experimental value for the initial average sequence concentration $\bar{c}(t_0)$ and the initial sedimented DNA size $\lambda(t_0)$. Using these values, we find that the theoretical results (solid lines in Fig. 3*B*) agree well with the experimental data points.

Based on the agreement between experiment and theory, we could use the theory to extrapolate the selection kinetics for a larger amount of cycles (Fig. 3*C*, solid lines). For the experimental partial removal of the dilute phase, we found that selection approximately doubled after 20 cycles. The selection kinetics saturate because the sediment has grown to the volume corresponding to the *Bottom* fraction.

Finally, we used the theory to consider the ideal case of the complete removal of the dilute phase. We find that, for this ideal case, the sequence pair can enrich by twofold better than the for *Top* fraction removal and more than 10-fold compared to the initial pool (Fig. 3*C*, dashed lines).

In summary, we have shown experimentally that discrete cycles of feeding, i.e., replacing the dilute phase with a pool, lead to the enrichment of specific sequence pairs through the formation of dense phases, confirming the theoretically proposed selection mechanism.

DNA Phase Separation in a Continuous Feeding Flow. The conditions on the early Earth scenario in general deviated from the simplified scenario of periodic feeding cycles separated by waiting times to relax to phase equilibrium. Here, we demonstrate a continuous implementation of the selection mechanism shown in Fig. 3. Using a microfluidic setup, we mimic a continuous exchange of solute and simulate a feeding flow of nucleic acids and salts flowing through a rock pore (Fig. 1*A*). Previous studies have shown that gases and liquids can effectively pass through cracks in hydrothermal structures like shattered glassy basalt, providing one possible implementation of the scenario mentioned above (33, 34). This flow generates a selection pressure, requiring sequence pairs to rapidly phase separate and settle at the pore's bottom to avoid being flushed out, as this would threaten the dilution by a larger reservoir such as the ocean, potentially resulting in hydrolysis-induced death or loss of information.

To test the concept, we designed a microfluidic system with a continuous flow of a pool and compared it to fluid flow theory using COMSOL Multiphysics. In this setup (*SI Appendix*, section 13), a pore (3 mm × 6 mm × 500 μm) at 15 °C is connected to a feeding pool of sequence pair 1 through a channel (See *Left* green arrow in Fig. 4*A*). The results of this experiment are shown in the fluorescence micrographs in Fig. 4*A*. Using an inflow speed of 2 μm/s, we observe upconcentration of sequence

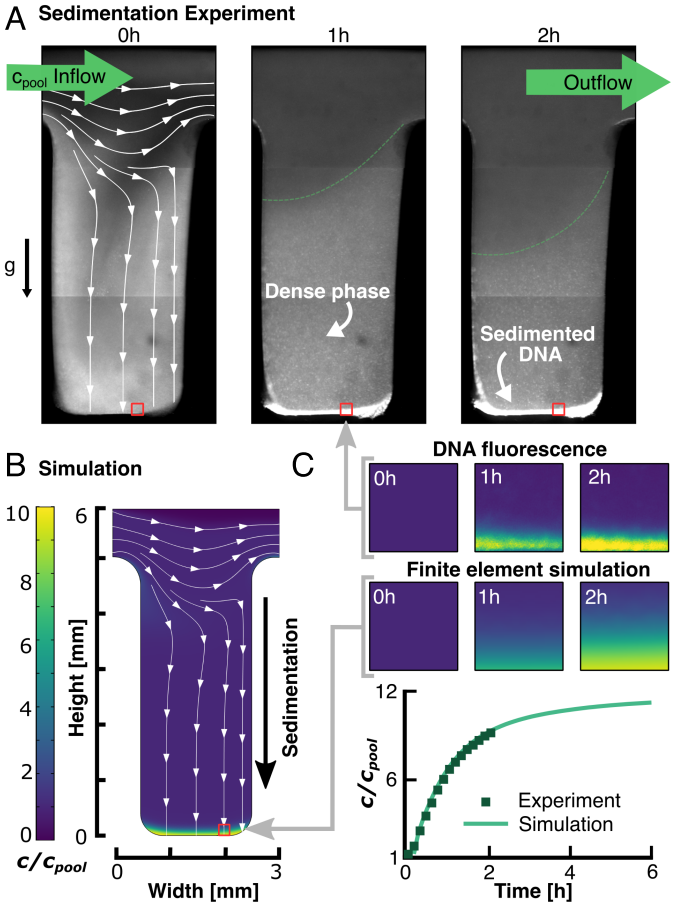


Fig. 4. DNA Phase separation with continuous feeding flow. (*A*) Fluorescence time-lapse images of a microfluidic chamber with a 2 μm/s inflow of sequence pair 1 with concentration c_{pool}. DNA that phase separates can sediment, thus is not advected out of the chamber. This implements cycles of phase separation in a system with continuous flow. (*B*) Finite element simulation of fluid flow with sedimentation and diffusion confirms the experimental findings in detail when assuming a downward sedimentation speed of 0.1 μm/s and a diffusion coefficient of 5 μ m^2/s. (*C*) The relative concentration c/c_{pool} at the bottom matches well between experiment and simulation, seen for a 200 μm sized squared cut-out or when plotted over time.

pair 1 by phase separation and sedimentation inside the pore despite the continuous outflow.

In our setup, we found that after 2 h, the concentration of sequence pair 1, $c(t)$, is enriched 8-fold relative to the inflow concentration c_{pool}. The inflow and the choice of pore geometry have to be tuned not to perturb the sedimentation speed of the dense phase DNA (Fig. 4*B*). By numerically solving the hydrodynamic flow equations in addition to sedimentation and diffusion of dense phase DNA, the simulated increase in concentration agrees well with the experiments in both space and time (Fig. 4*C*). In summary, our findings demonstrate that for a sufficiently slow flow (in the 10 μm/s regime), sequences can be effectively selected from the continuously fed pool and accumulate within a pore. This emphasizes the viability of the selection mechanism in a prebiotic context.

Selection in Pools with Many Sequences. Up to now, we have investigated specific pools composed of only a few designed sequence pairs for the proof of principle. It remains unclear how robust our proposed selection mechanism is for realistic pools that are formed via polymerization and contain many different sequences. To tackle this question, we use our theory of cyclic oligonucleotide phase separation with discrete feeding cycles and consider pools that could emerge from the polymerization of different units. For simplicity, we focus on sequences of fixed length L composed of two different units, 0 and 1. Both units can be either thought of as two different nucleotides, or as two nucleotide segments of fixed compositions, like the building blocks of the sequences introduced in Fig. 2*A*.

Following ref. 35, sequence ensembles can be characterized by two parameters, the relative composition of the two units r and the blockiness b_l. The latter determines the chain correlations of the two units: For $b_l = 1$, the model favors homopolymers (..11.. or ..00..), while for $b_l = -1$, sequences are anti-correlated heteropolymers (..1010..); see *SI Appendix*, section 3*A* for more details on the model. Phase separation of an ensemble of different sequences occurs once the system crosses the cloud point in the phase diagram; for details, see *SI Appendix*, section 3*B*. Subjecting such a phase-separated sequence pool to removal and feeding cycles, we find qualitatively different selection scenarios depending on the parameters r and b_l.

For initial pools of low blockiness ($b_l < 0$), we find that the sequence bias of the pool is strongly amplified for a large number of cycles n (Fig. 5*A*). This behavior results from the strong interaction propensity among sequences of the same type. Dominantly, such sequences are recruited from the dilute phase after pool replacement, while other sequences partition into the dilute phase and get subsequently removed by the replacement step. In contrast, for initial pools of high blockiness ($b_l > 0$), the initial sequence bias is completely altered while cycling (Fig. 5*B*).

In particular, homopolymeric sequences are disfavored, while heteropolymeric sequences get more favored as cycles proceed. The reason is that a homopolymeric sequence cannot interact with copies of itself, while more heteropolymeric sequences can. These trends are summarized for largely different sequence ensembles for different values of unit compositions r and blockiness b_l by determining the most abundant sequence before cycling and for a large amount of cycles n_{f} (Fig. 5*C*).

We identify a domain at low blockiness ($b_l < 0$) where the most abundant sequences are amplified. In this domain, the number of sequences in solution decreases significantly. Thus, a description in terms of a few strongly interacting sequences becomes more and more accurate as cycles proceed.

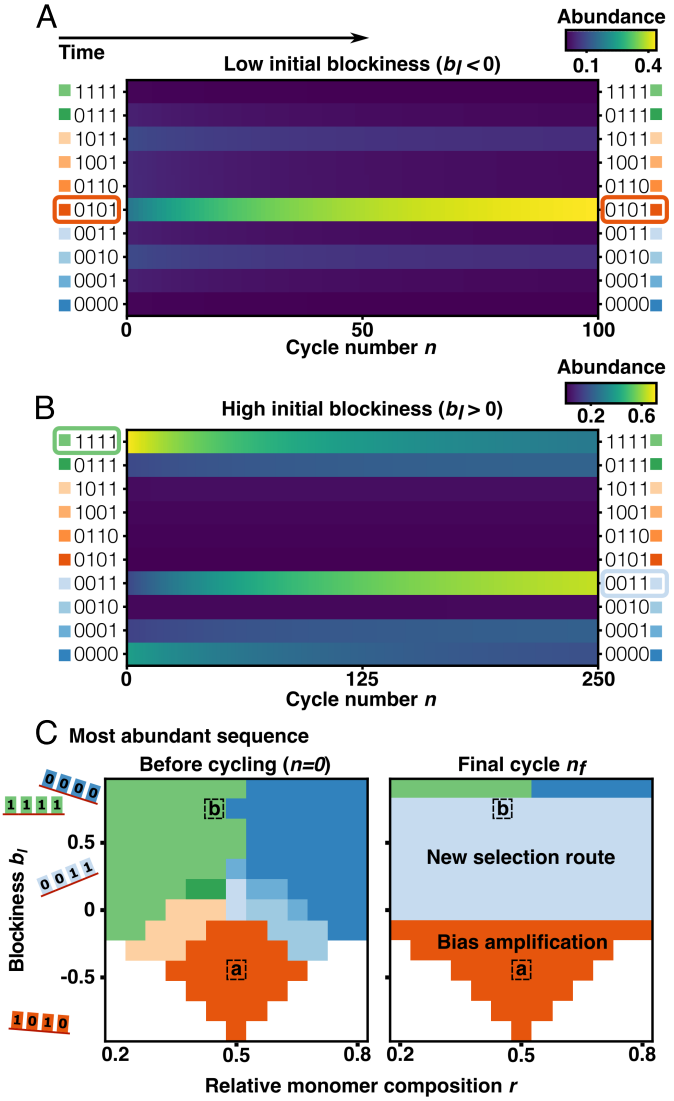


Fig. 5. Cycles of phase separation can amplify pool bias or offer an alternative selection route. We consider pools composed of sequence ensembles characterized by two parameters: blockiness b_l and relative composition r of the units 0 and 1. (*A* and *B*) For low initial blockiness ($b_l < 0$), the initially most abundant sequence (here: 0101) gets further enriched. In contrast, for high initial blockiness ($b_l > 0$), homopolymeric sequences (e.g. 1111 and 0000) are depleted with cycling while a more heteropolymeric sequence (0011) is strongly amplified. The color code indicates relative sequence abundances. (*C*) The most abundant sequence is shown for different values of b_l and r before cycling (*Left*) after cycling n_{f}-times (*Right*). Further amplification of the initially most abundant sequence is found for low blockiness ($b_l < 0$), while new selection routes can emerge at high blockiness ($b_l > 0$).

For a quantification of the variation in abundance due to phase separation cycles, see *SI Appendix*, Fig. S4*H*. New selection routes can emerge for sequence ensembles of intermediate or large blockiness ($b_l > 0$). The later regime leads to a switch of the most abundant sequence when subjecting the system to a large number of feeding cycles.

Conclusion

Here, we showed that the ability of oligonucleotides to phase separate can give rise to an evolutionary selection mechanism if subject to feeding cycles. In particular, replacing the dilute phase with a constant pool composed of different oligonucleotide sequences leads to the growth of a dense phase of specific

sequences while others are depleted relative to the pool. We have quantitatively confirmed our theoretical predictions for discrete cycles by experiments using designed DNA sequences. In addition, we showed that the same mechanism also caused the selection of specific sequences in a prebiotic-relevant scenario where the system is subjected to a continuous flow of the pool.

The key property of the proposed selection mechanism is that it is highly sequence-specific, also in the presence of other interacting sequences. Specifically, sequences that interact strongly with other sequences are enriched in the dense phase while weakly interacting sequences are expelled and thus leave the system through the removal step. A key observation of our work is that the selection mechanism also works for very short oligonucleotides. In our experiments, sequences of 22 nucleotides with base pairing regions of 6 nt form cooperative base-pairing networks at room temperature and phase separate, while others with the length of 28 nucleotides and base pairing regions of 8 nt do not. In contrast to the strong length selectivity of mineral surfaces, for example, this mechanism counter-intuitively prioritizes sequence over length in this case. In the future, it would be interesting to investigate the combined effect of DNA phase separation with other selection mechanisms, particularly how it affects the sequence and length distribution of freshly polymerized oligonucleotide pools (36).

Using theory, we studied realistic, multi-sequence pools that result from polymerization. We found robust and pronounced selection kinetics already for sequences composed of only four segments of nucleotide sequences. We distinguish two qualitative scenarios of sequence selection, where either the initial sequences bias is strongly amplified, or the initial bias is swapped and other sequences are selected.

The robustness of our selection mechanism, particularly for short oligonucleotides, suggests its relevance at the molecular origin of life, where specific short-chained peptides, RNA, and DNA sequences were recruited during their assembly from prebiotic pools. The cyclic removal of weakly interacting sequences can guide the selection of longer sequences which face dilution by the exponentially growing size of sequence space. Moreover, the dense phase could have provided enhanced stability against degrading chemical reactions such as catalytic cleavage (28) or hydrolysis due to the duplex formation (37). In fact, there is a correlation between catalytic sequences and phase separation in functional ribozyme polymerases (38). Ultimately, we expect an enhanced selection propensity when combining self-replication with our selection mechanism relying on base-pairing interactions.

Materials and Methods

Strand Design. DNA oligonucleotide systems were designed using the NUPACK software package 3.2.2 (39). The strands were constrained to contain three binding sites separated by spacers either composed of TT or CC. Systems that formed the intended secondary structure (each strand base-pairing with three other strands) were chosen (*SI Appendix*, section 4). The oligomers were ordered from biomers.net GmbH, in a dry state, with high-performance liquid chromatography purification. The sequences were as follows (5′-3′) - Sequence pair 1, sequence i: GGA CCC TTC GGC CGT TCG CTCG; sequence ii: GGG TCC TTC GGC CGT TCG AGCG; Sequence pair 2, sequence i: AAT ATA TAC CGC GGC CGG CCT ATA ATA A; sequence 2: TAT ATA TTC CCC GGC CGC CCT TAT TATA; Sequence pair 3, sequence i: GGC GCG CGT TGC GGC CGG TTC GCG GCGG; sequence ii: CGC GCG CCT TCC GGC CGC TTC CGC CGCG. All the strands were stored at −20 °C, diluted in nuclease-free water at 200 μM. Before every experiment, the strands were denatured at 95 °C for 2 min.

Reaction Mixtures. Initial pools of 15 μL were prepared with 25 μM of each respective DNA strand, 10 mM Tris Buffer-HCl pH 7, 5X SYBR Green I (intercalating dye; excitation 450 to 490 nm, emission 510 to 530 nm), 125 mM NaCl, and 10 mM $MgCl_2$. The mixtures were heated to 95 °C for 2 min to ensure full de-hybridization of the strands. The temperature protocol that allows hybridization and consequent phase separation was *i*. 95 °C for 2 min, *ii*. 65 °C for 10 s, *iii*. cooling to 15 °C (ramp rate: 6 K per minute), iv. 15 °C for at least 3 h. Temperature protocols were performed in a standard thermocycler (Bio-Rad CFX96 Real-Time System). Melting curves were measured in triplicates using the same reaction mixture and temperature profile as for the sedimentation experiments (*SI Appendix*, section 8). Baseline correction using a reference measurement with only SYBR Green I. In the case of feeding cycle experiments, after sedimentation, 7.5 μL of the dilute phase, corresponding to 50% of the initial volume, was removed by carefully pipetting only at the center of the meniscus to avoid removing material from the sediment. Afterward, 7.5 μL of the initial pool stock was added to the remaining *Bottom* fraction. The aforementioned temperature protocol was then repeated, completing one feeding cycle.

Sedimentation Imaging. The imaging experiments were performed in a microfluidic chamber containing multiple wells, cut out of 500 μm Teflon foil and sandwiched between two sapphire plates (*SI Appendix*, Fig. S16). The sample volume (about 15 μL per well) was loaded by using microloader pipette tips. The temperature of the chamber was controlled using three Peltier elements. To remove the waste heat from the Peltier elements, a Julabo 300F water bath (JULABO GmbH) was used to cool the back of the chamber. The entire chamber is held in place by screwing a steel frame on top using a homogenous torque of 0.2 Nm. After loading, the wells were sealed with Parafilm to avoid evaporation. Monitoring of the sedimentation was performed using a self-built fluorescence microscope composed of a 490-nm LED (M490L4, Thorlabs), a 2.5x Fluar objective (Zeiss), and the FITC/Cy5 H Dualband Filterset (AHF). Multiple wells could be imaged by moving the chamber perpendicularly to the light axis with two NEMA23 Stepper Motors and a C-Beam Linear Actuator (Ooznest Limited). Images were taken using a Stingray-F145B CCD camera (ALLIED Vision Technologies) connected via FireWire to a computer running a self-written Labview code operating camera, motors, LED's and Peltier elements (*SI Appendix*, Fig. S15). Flow-through experiments were conducted using a similar chamber without Peltier Elements, only using the water bath at homogeneous 15 °C. In this case, the sapphires have holes on the backside, where an outlet and two inlet tubings were attached. Inlet tubing 1 contained 20 uM of each strand of system 3 and 10X SYBR Green I, while inlet tubing 2 contained 20 mM TRIS pH 7, 250 mM NaCl, and 20 mM $MgCl_2$. Flowspeed was adjusted using the Nemesys Controler NEM-B002-02 D (Cetoni GmbH) with two 100 μl syringes. Hardware was controlled using a self-written labview (National Instruments) software (*SI Appendix*, Fig. 4).

High-Performance Liquid Chromatography (HPLC). Ion-pairing reverse-phase HPLC experiments were carried out on a column liquid chromatography system equipped with an auto-sampler and a bio-inert quaternary pump (Agilent 1260 Infinity II Bio-Inert Pump G5654A, Agilent Technologies). A C18 capillary column (AdvanceBio Oligonucleotide 4.6x150 mm with particle size 2.7 μm, Agilent) was used to perform reverse-phase liquid chromatography. The temperature of the autosampler was set to 4 °C. The mobile phases consisted of two eluents. Eluent A was HPLC water (Sigma-Aldrich), 200 mM 1,1,1,3,3,3, -Hexafluoro-2-propanol (HFIP) (Carl Roth GmbH), 8 mM Triethylamine (TEA) (Carl Roth GmbH). Eluent B was a 50:50 (v/v) mixture of water and methanol (HPLC grade, Sigma Aldrich, Germany), 200 mM HFIP, and 8 mM TEA. The injection volume for each measurement was 100 μL. The samples were eluted with a gradient of 1% B to 58.6% B over the course of 45 min with a flow rate of 1 mL/min. Prior to the gradient, the column was flushed with 1% B for 5 min. Retention times were analyzed via a UV Diode Array Detector (Agilent 1260 Infinity II Diode Array Detector WR G7115A) at 260 nm with a bandwidth of 4 nm. Samples were diluted for HPLC loading in the following manner: 7.5 μL of sample, 105 μL nuclease-free water, and 75 μL of a 5M urea solution. They were heated to 95 °C for 2 min afterward to ensure de-hybridization of the strands and dissolution of any sediment. Then, 105 μL of the diluted samples were transferred into N9 glass vials (Macherey-Nagel GmbH) and stored at 4 °C in

the auto-sampler of the HPLC-MS system (1260 Infinity II, Agilent Technologies) until injection.

Finite-Element Simulations. Simulations were performed in 2D using COMSOL Multiphysics 5.4. The simulation file with all the detailed parameters is given in the supplement in binary format. Additionally, the simulation is given as an auto-generated report in a hierarchical html compressed into a Zip-File. For more detailed information, see *SI Appendix*, section 8.

Data, Materials, and Software Availability. The data and the codes that support the findings of this study are available at the following online repository (40).

ACKNOWLEDGMENTS. D. Braun acknowledges support from the European Research Council (ERC) Evotrap, Grant No. 787356, the Simons Foundation (Grant No. 327125), the CRC 235 Emergence of Life (Project-ID 364653263), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2094 – 390783311, and the Center for NanoScience (CeNS). C. Weber acknowledges the ERC under the European Union's Horizon 2020 research and innovation programme (Fuelled Life, Grant No. 949021) for financial support.

Author affiliations: [a]Division Biological Physics, Max Planck Institute for the Physics of Complex Systems, Dresden 01187, Germany; [b]Center for Systems Biology Dresden, Dresden 01307, Germany; [c]Ludwigs-Maximilian-Universität München and Center for NanoScience, Munich 80799, Germany; and [d]Faculty of Mathematics, Natural Sciences, and Materials Engineering: Institute of Physics, University of Augsburg, Augsburg 86159, Germany

Author contributions: G.B., A.C.S., P.S., C.B.M., D.B., and C.A.W. designed research; G.B., A.C.S., P.S., A.K., Y.R., P.J., D.H., C.B.M., D.B., and C.A.W. performed research; A.K. and C.B.M. contributed new reagents/analytic tools; G.B., A.C.S., P.S., C.B.M., D.B., and C.A.W. analyzed data; and G.B., A.C.S., P.S., D.B., and C.A.W. wrote the paper.

1. M. P. Robertson, G. F. Joyce, The Origins of the RNA World. *Cold Spring Harb. Persp. Biol.* **4**, a003608 (2012).
2. K. Kruger *et al*., Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell* **31**, 147–157 (1982).
3. M. J. Fedor, J. R. Williamson, The catalytic diversity of RNAs. *Nat. Rev. Mol. Cell Biol.* **6**, 399–412 (2005).
4. G. Walter, The RNA world Superlattices point ahead. *Nature* **319**, 618 (1986).
5. K. R. Birikh, P. A. Heaton, F. Eckstein, The structure, function and application of the hammerhead ribozyme. *Euro. J. Biochem.* **245**, 1–16 (1997).
6. C. Deck, M. Jauker, C. Richert, Efficient enzyme-free copying of all four nucleobases templated by immobilized RNA. *Nat. Chem.* **3**, 603–608 (2011).
7. A. Mariani, D. A. Russell, T. Javelle, J. D. Sutherland, A light-releasable potentially prebiotic nucleotide activating agent. *J. Am. Chem. Soc.* **140**, 8657–8661 (2018).
8. T. Walton, W. Zhang, L. Li, C. P. Tam, J. W. Szostak, The mechanism of nonenzymatic template copying with imidazole-activated nucleotides. *Angew. Chem.- Int. Ed.* **58**, 10812–10819 (2019).
9. S. Wunnava *et al*., Acid-catalyzed RNA-oligomerization from 3',5'-cGMP. *Chem. Euro J.* **27**, 17581–17585 (2021).
10. I. Budin, J. W. Szostak, Expanding roles for diverse physical phenomena during the origin of life. *Annu. Rev. Biophys.* **39**, 245–263 (2010).
11. A. V. Tkachenko, S. Maslov, Spontaneous emergence of autocatalytic information-coding polymers. *J. Chem. Phys.* **143** (2015). https://pubs.aip.org/aip/jcp/article/143/4/045102/71438/Spontaneous-emergence-of-autocatalytic-information.
12. A. V. Tkachenko, S. Maslov, Onset of natural selection in populations of autocatalytic heteropolymers. *J. Chem. Phys.* **149** (2018). https://pubs.aip.org/aip/jcp/article/143/4/045102/71438/Spontaneous-emergence-of-autocatalytic-information.
13. C. B. Mast, S. Schink, U. Gerland, D. Braun, Escalation of polymerization in a thermal gradient. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 8030–8035 (2013).
14. M. Morasch *et al*., Heated gas bubbles enrich, crystallize, dry, phosphorylate and encapsulate prebiotic molecules. *Nat. Chem.* **11**, 779–788 (2019).
15. R. Mizuuchi *et al*., Mineral surfaces select for longer RNA molecules. *Chem. Commun.* **55**, 2090–2093 (2019).
16. L. H. de Oliveira *et al*., When RNA meets montmorillonite: Influence of the pH and divalent cations. *Appl. Clay Sci.* **214**, 106234 (2021).
17. W. M. Aumiller, F. Pir Cakmak, B. W. Davis, C. D. Keating, RNA-based coacervates as a model for membraneless organelles: Formation, properties, and interfacial liposome assembly. *Langmuir* **32**, 10042–10053 (2016).
18. Bj. Jeon *et al*., Salt-dependent properties of a coacervate-like, self-assembled DNA liquid. *Soft Matter* **14**, 7009–7015 (2018).
19. M. Nakata *et al*., End-to-end stacking and liquid crystal condensation of 6-to 20-base pair DNA duplexes. *Science* **318**, 1276–1279 (2007).
20. G. Zanchetta, M. Nakata, M. Buscaglia, N. A. Clark, T. Bellini, Liquid crystal ordering of DNA and RNA oligomers with partially overlapping sequences. *J. Phys. Condens. Matter* **20**, 494214 (2008).
21. D. T. Nguyen, O. A. Saleh, Tuning phase and aging of DNA hydrogels through molecular design. *Soft Matter* **13**, 5421–5427 (2017).
22. Z. Xing *et al*., Microrheology of DNA hydrogels. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 8137–8142 (2018).
23. Y. Sato, T. Sakamoto, M. Takinoue, Sequence-based engineering of dynamic functions of micrometer-sized DNA droplets. *Sci. Adv.* **6**, eaba3471 (2020).
24. D. M. Mitrea, R. W. Kriwacki, Phase separation in biology; Functional organization of a higher order Short linear motifs - The unexplored frontier of the eukaryotic proteome. *Cell Commun. Sig.* **14**, 1–20 (2016).
25. L. M. Barge *et al*., Thermodynamics, disequilibrium, evolution: Far-from-equilibrium geological and chemical considerations for origin-of-life research. *Orig. Life Evol. Biosph.* **47**, 39–56 (2017).
26. F. Westall *et al*., A hydrothermal-sedimentary context for the origin of life. *Astrobiology* **18**, 259–293 (2018).
27. J. SantaLucia Jr, A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1460–1465 (1998).
28. O. A. Saleh, B. J. Jeon, T. Liedl, Enzymatic degradation of liquid droplets of DNA is modulated near the phase boundary. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 16160–16166 (2020).
29. Y. Xing *et al*., Self-assembled DNA hydrogels with designable thermal and enzymatic responsiveness. *Adv. Mater.* **23**, 1117–1121 (2011).
30. S. Biffi *et al*., Phase behavior and critical activated dynamics of limited-valence DNA nanostars. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15633–15637 (2013).
31. S. Agarwal, D. Osmanovic, M. A. Klocke, E. Franco, The growth rate of DNA condensate droplets increases with the size of participating subunits. *ACS Nano* **16**, 11842–11851 (2022).
32. D. Sun, L. H. Hurley, The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: Implications for drug targeting and control of gene expression. *J. Med. Chem.* **52**, 2863–2874 (2009).
33. M. Colombier *et al*., Degassing and gas percolation in basaltic magmas. *Earth Planet. Sci. Lett.* **573**, 117134 (2021).
34. F. Westall *et al*., A hydrothermal-sedimentary context for the origin of life. *Astrobiology* **18**, 259–293 (2018).
35. G. H. Fredrickson, S. T. Milner, L. Leibler, Multicritical phenomena and microphase ordering in random block copolymers melts. *Macromolecules* **25**, 6341–6354 (1992).
36. A. V. Dass *et al*., Rna oligomerisation without added catalyst from 2', 3'-cyclic nucleotides by drying at air-water interfaces. *Chem. Syst. Chem.* **5**, e202200026 (2023).
37. K. Zhang, J. Hodge, A. Chatterjee, T. S. Moon, K. M. Parker, Duplex structure of double-stranded RNA provides stability against hydrolysis relative to single-stranded RNA. *Environ. Sci. Technol.* **55**, 8045–8053 (2021).
38. A. Salditt *et al*., Thermal habitat for RNA amplification and accumulation. *Phys. Rev. Lett.* **125**, 048104 (2020).
39. J. N. Zadeh *et al*., Nupack: Analysis and design of nucleic acid systems. *J. Comput. Chem.* **32**, 170–173 (2011).
40. G. Bartolucci, A. Calaça Serrão, P. Schwintek, Supplementary codes. Selection_via_PS. GitHub. https://github.com/GiacoBarto/selection_via_PS. Deposited 4 August 2023.

Nucleic Acids Research, 2024, **52**, 1290–1297
https://doi.org/10.1093/nar/gkad1190
Advance access publication date: 14 December 2023
Molecular Biology

Replication elongates short DNA, reduces sequence bias and develops trimer structure

Adriana Calaça Serrão †, Felix T. Dänekamp †, Zsófia Meggyesi and Dieter Braun *

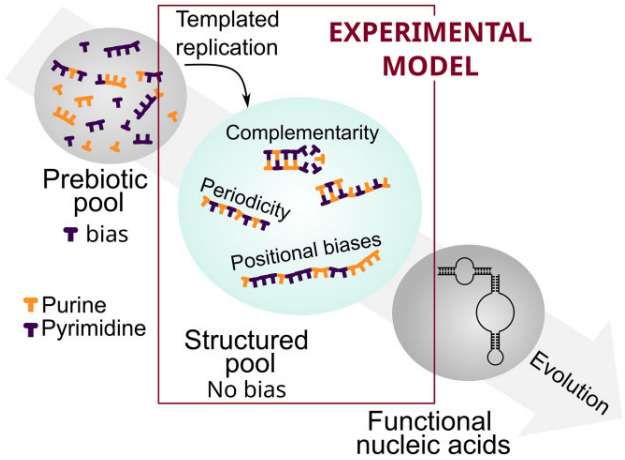
Systems Biophysics, Physics Department, Center for NanoScience, Ludwig-Maximilians-Universität München, Amalienstraße 54, 80799 Munich, Germany

*To whom correspondence should be addressed. Tel: +49 89 2180 1484; Email: dieter.braun@lmu.de
†The first two authors should be regarded as Joint First Authors.

Abstract

The origin of molecular evolution required the replication of short oligonucleotides to form longer polymers. Prebiotically plausible oligonucleotide pools tend to contain more of some nucleobases than others. It has been unclear whether this initial bias persists and how it affects replication. To investigate this, we examined the evolution of 12-mer biased short DNA pools using an enzymatic model system. This allowed us to study the long timescales involved in evolution, since it is not yet possible with currently investigated prebiotic replication chemistries. Our analysis using next-generation sequencing from different time points revealed that the initial nucleotide bias of the pool disappeared in the elongated pool after isothermal replication. In contrast, the nucleotide composition at each position in the elongated sequences remained biased and varied with both position and initial bias. Furthermore, we observed the emergence of highly periodic dimer and trimer motifs in the rapidly elongated sequences. This shift in nucleotide composition and the emergence of structure through templated replication could help explain how biased prebiotic pools could undergo molecular evolution and lead to complex functional nucleic acids.

Graphical abstract



Introduction

The replication of short oligonucleotides to create longer polymers is a central step in the origin of more functional nucleic acids. It has been addressed through enzymatic (1,2) and non-enzymatic replication (3–5), mostly from specific sequences or naive pools of short oligomers. However, condensation of mononucleotides in a primordial context often leads to short oligomer pools with a sequence bias, namely with one nucleobase incorporated more into the product strands (6–9). This bias, on the one hand, may be due to an imbalanced abundance in the environment caused by different rates of nucleotide formation and degradation in different conditions (10–14). On the other hand, even when the environment has equimolar concentrations of all reacting nucleotides, the rate of the condensation reactions themselves may also vary for different nucleotides (6,9,15).

Functional nucleic acid strands are usually long, with several tens or hundreds of base pairs (16), and have specific secondary structure (17,18). Even though such catalytic nucleic acids occupy only a subsection of the possible sequence space (19,20), they are still more compositionally diverse than the biased pools obtained from nucleotide condensation studies (10). The mechanism through which such functional strands evolve from a pool of short biased oligomers, both elongating and driving the evolution of sequence information, is not fully understood (21,22).

Received: July 13, 2023. **Revised:** November 15, 2023. **Editorial Decision:** November 25, 2023. **Accepted:** November 30, 2023

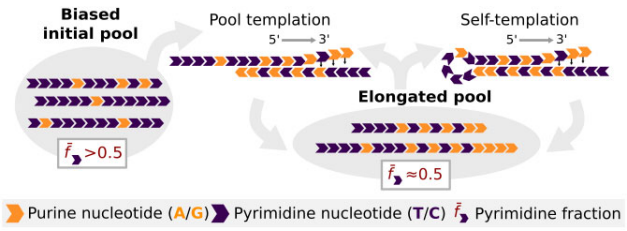


Figure 1. Polymerization starts from a binary initial pool (AT or GC) with a bias $\bar{f}$ of either purine (orange) or pyrimidine (purple) nucleotides. Distinct sequences from the pool base-pair to form short duplexes and are enzymatically extended (5′–3′) complementarily to the template ('pool templation'). Longer sequences may also self-template through hairpin-like secondary structures ('self-templation'). The biased pyrimidine fraction $\bar{f}_T$ or $\bar{f}_C$ in the initial pool is countered by complementary elongation.

Templated replication is a potential mechanism through which both the compositional diversity and sequence length can increase to facilitate the exploration of sequence space while replicating sequence information (10). Due to the complementarity of Watson–Crick base pairing, necessary for templation, a strong bias to one nucleobase leads to the complementary base being correspondingly more incorporated in the nascent strand. This in turn homogenizes the average pool nucleotide fraction (Figure 1). While the overall nucleotide composition is expected to diversify, several studies have shown that templated replication can act as a selection mechanism in itself, enriching specific sequence motifs (1,23–25). More experimental investigations are needed to grasp the influence of the initial biases of the pool on the sequence level. The goal of our study was to specifically understand which motifs are enriched starting from such biased initial pools and whether the replicated pool holds memory of the initial bias.

In contemporary biology, strand separation and elongation occur in tandem (26,27). The displacement of any pre-hybridized strands is performed enzymatically. However, strand displacement can also be triggered by the hybridization of other sequences in the pool (28). This non-enzymatic strand displacement has recently been described for a prebiotic RNA replication system (29). When compared to other prebiotic mechanisms proposed for strand separation, such as pH (22,30), heat and salt fluctuations (31), strand displacement has the advantage that it can also occur isothermally and with a constant chemical environment (28). It thereby erases the need for a specific set of cycle conditions that are potentially more difficult to satisfy and isolates the impact of replication on sequence structure from other environment variables.

We investigated how the sequence landscape of short biased DNA pools evolves upon templated polymerization with *Bacillus stearothermophilus* strand displacing polymerase (*Bst*). *Bst* binds to double-stranded regions and elongates the strand in the 5′ to 3′ direction with high fidelity (32–34), displacing downstream bound strands (Supplementary Data, Section II). A single strand can therefore go through several replication rounds, even in isothermal conditions—first through pool templation and later, when a certain length threshold is crossed, through self-templation (Figure 1). This is therefore a robust model system for prebiotic primer extension starting from a diverse pool. With the faster enzymatic kinetics, the influence of the replication cycles on the pool composition and diversity can be assessed.

The initial pools studied consisted of short 12-mer DNA strands, with a binary composition of either AT or GC, and of all the four possible biases (A-rich, G-rich, etc.). After following the sequence space over the course of incubation with *Bst*, we found that the initial nucleotide bias of the pool disappears, so that the resulting pool has a nucleotide fraction of 0.5 (i.e. 50% A and 50% T). While this new pool is now homogenized in terms of overall nucleotide composition, individual segments of the elongated strands still retain traces of the initial bias, due to the directionality of the polymerization. This shows that even though the overall pool nucleotide fraction changes through replication, the structure within sequences depends on the initial state. Furthermore, we have also observed that highly periodic motifs are present in sequences that elongate fast.

Materials and methods

Polymerization with *Bst*

The polymerization reactions were performed with *Bst* 2.0 DNA Polymerase (New England Biolabs, #M0537S). The conditions were according to the protocol provided by the manufacturer: 1× Isothermal Amplification Buffer, 8 mM $MgSO_4$ (for a total of 10 mM with 2 mM $MgSO_4$ from the 1× buffer), 320 U/ml *Bst* (all supplied when ordering the enzyme), 1.4 mM of each nucleotide triphosphate and 10 μM DNA. AT samples were supplied with 1.4 mM dATP and dTTP and GC experiments with 1.4 mM dGTP and dCTP (all from Sigma–Aldrich), and the ATGC experiments with all four nucleotides (1.4 mM of each). All experiments were conducted with initial DNA samples containing only random 12-mers provided by biomers.net, with binary base alphabets (AT, GC) in varying base content and for the ATGC experiment a full base alphabet (Supplementary Data, Section I). The ordered base content differs from the effective base content detected with next-generation sequencing (NGS) (Figures 3 and 4, and Supplementary Data, Section VI). The polymerization reactions were incubated in a thermocycler with the following protocol: (i) constant temperature (35°C for AT, 65°C for GC, 45°C for ATGC) for the reported time and (ii) 90°C for 20 min to deactivate *Bst*. The incubation temperature was lower for AT than for GC due to differences in melting temperature, and based on a temperature screening performed with *Bst* (Supplementary Data, Section VII).

PAGE and gel imaging

The samples were run in a denaturing 15% polyacrylamide in 50% urea, with a 19:1 ratio of acrylamide to bis-acrylamide and polymerized with tetramethylethylenediamine and ammonium persulfate. The gels were pre-heated in the electrophoretic chamber at 300 V for 27 min. The samples were then loaded, in a mixture with a ratio of 2:7 of sample to loading dye. Loading dye is prepared in-house [for 10 ml: 9.5 ml formamide, 0.5 ml glycerol, 1 μl ethylenediaminetetraacetic acid (EDTA, 0.5 M) and 100 μl Orange G dye (New England Biolabs, #B7022S)]. The running protocol for the gels in the electrophoretic chamber was 50 V for 5 min followed by 300 V for 25 min. After the run, the gels were stained with a 2× SYBR Gold (Thermo Fischer Scientific, #S11494) dilution in Tris–borate–EDTA (TBE) buffer 1×. They were then rinsed

with 1× TBE buffer twice and imaged using a Bio-Rad Chemi-Doc™ MP imaging system. The 20–100 bp ladder (DNA oligo length standard 20/100 Ladder, IDT, #51-05-15-02) was supplied in a final concentration of 2.04 ng/μl (for each rung) and the 100–1517 bp ladder (100 bp DNA Ladder, New England Biolabs, #N3231S) in a final concentration of 71.4 ng/μl (for all rungs; concentrations vary by *n*-mer as described by the manufacturer). Finally, the obtained micrographs were loaded into and analyzed with a self-written LabVIEW program (Supplementary Data, Section VIII).

Sequencing

Samples were sequenced by the Gene Center Munich (LMU) using the NGS Illumina NextSeq 1000 machine (flow cell type P2, 2 × 50 bp with 138 cycles for 100 bp single-end reads; at most 120 bp with two indexes were read, with declining quality toward the end). Fifty million reads were ordered for each sample. The raw sequencing data obtained, in FastQ format, were processed in this order by demultiplexing, quality score trimming and regular expression filtering. Demultiplexing was performed with software from Galaxy servers (35), provided by the Gene Center Munich. During sequencing, each read base was assigned a Phred quality score $Q = -10 \log_{10} P$, where P is the probability of an incorrectly read base (36). Using Trimmomatic (37), we trimmed low-quality segments by running a sliding window of 4 nt in the 3′ to 5′ direction over the sequence that allowed a minimum average Phred quality of 20, otherwise trimming at the leftmost base of the window, corresponding to an average accuracy of at least 99%. As the experimentally obtained sequences were appended on the 3′ terminus with a CT tail followed by an AGAT during sequencing preparation, those needed to be found and cut, for which we employed the following regular expressions:

(^[AT]{12,})(?=([CT]{4,}AGAT)) for AT
(^[CG]{12,})(?=([CT]{4,}AGAT)) for GC
(^[ATGC]{12,})(?=([CT]{4,}AGAT)) for ATGC

This also ensured that only binary sequences were included in the analysis of binary pools. For the ATGC experiment, a further adapter filtering step was employed to recover the sequencing signal from the adapter contaminated reads (Supplementary Data, Section XII).

Results and discussion

Length distribution of binary pools over time

Our starting pools with 10 μM total DNA were composed of random 12 nt long single-stranded binary sequences (AT or GC only) with a bias in the nucleotide fraction. The four binary pools studied were labeled according to the more abundant nucleotide and were revealed to initially contain 60% A (A_0), 75% T (T_0), 70% G (G_0) and 69% C (C_0) by sequencing (see 'Materials and methods' section and Supplementary Data, Section IV). The sequence space was $2^{12} = 4096$, but sequences were not represented equally due to the bias. From these initial pools, sequences were isothermally amplified with the strand displacing enzyme *Bst*. The incubation temperatures were 35°C for AT pools and 65°C for GC pools. In a temperature screening, these led to the most extensive elongation (Supplementary Data, Section VII).

The evolution of sequence lengths over time was analyzed through PAGE (Figure 2A and B). Different time points were analyzed for AT and GC pools to account for the different kinetics of nucleotide incorporation (Figure 2C and D). The polymerization was stopped at 32 h for AT and 8 h for GC when the length distribution reached a state with an abundance of sequences well beyond 100 nt in the PAGE gel. Both the A-biased (A_0) and the G-biased (G_0) pools displayed replication to sequences longer than 100-mer within the first 2 h. In case of the A_0 pool, most of the short initial sequences (<20 nt) were depleted after 2 h, whereas for G_0 these remained detectable even for later time points. The remaining pools, T-biased (T_0) and C-biased (C_0), exhibit similar length distribution kinetics (A_0 to T_0 and G_0 to C_0, respectively) (Supplementary Data, Section V).

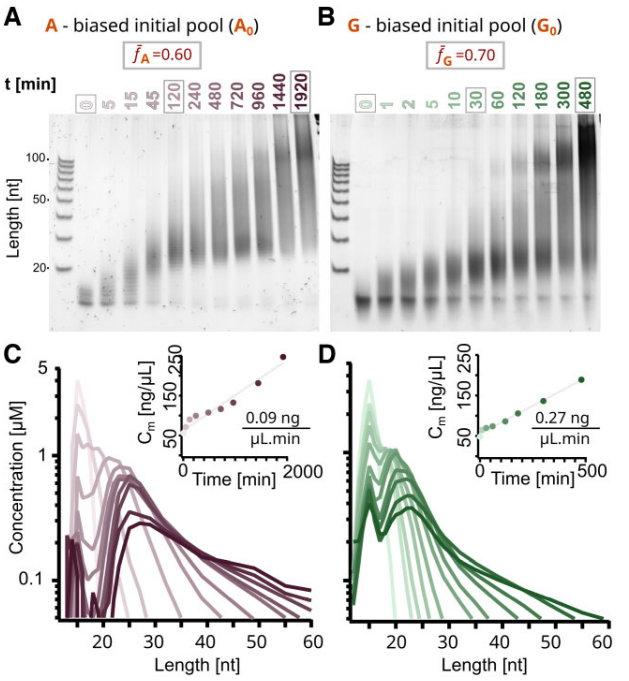


Figure 2. Templated polymerization of random DNA 12-mers leads to products longer than 100-mer. Polyacrylamide gel electrophoresis (PAGE) analysis shows the length distribution of (**A**) A-biased (A_0) and (**B**) G-biased (G_0) pools over time. The molar concentration of sequences was quantified and plotted over sequence length for each time point corresponding to individual lanes. A_0 corresponds to pink (**C**) and G_0 to green (**D**), with hue increasing over time. The total DNA mass concentration grows linearly with time (insets) and was fitted in gray. The concentrations were obtained from the gels by PAGE smear quantification (Supplementary Data, Sections VIII and IX).

The concentration profiles over strand length were obtained via ladder-calibrated SYBR Gold fluorescence intensity in PAGE gels and depicted for all time points in Figure 2C and D (Supplementary Data, Sections VIII and IX). The contribution of nucleotide composition to SYBR Gold intensity was ruled out by performing a screen with sequences of different compositions at known concentrations (Supplementary Data, Section IX.A). For both A_0 and G_0 pools, the molar concentration at later time points forms a double-peaked length distribution with a long tail that continues to lengths longer than 300 nt. The first peak, around 12 nt, could be explained by the sequences of the initial pool that were not recruited for replication. The second peak, between 20 and 30 nt, could be due to fully hybridized duplexes that have a melting temperature above the incubation temperature (38).

While the total number of sequences is constant because single nucleotides get added to already existing sequences, the total DNA mass increases linearly with time as more nucleotides are incorporated (Figure 2C and D, insets). The difference in kinetics observed (about three times slower for AT experiments) can be explained by both the temperature-dependent efficiency of *Bst* and nucleotide-dependent differences in the rate of incorporation (39,40).

Disappearance of nucleotide bias in the AT pool

To assess the sequence content of our product strands, we used NGS. For each of the four initial pools, three time point samples were sequenced (indicated in Figure 2A and B by the gray outlines). These represent the initial pool, an early time point pool, from which we learned about 'fast replicators', and a late time point pool to understand the sequence distribution in the 'left-behind' pool, respectively. The maximum sequence length captured is 112 nt, corresponding to the maximum read length of 120 nt minus the CT tail of at least 4 nt and the AGAT adapter, isolating the left-behind sequences from the fast replicators in the sequenced late time point. In the case of A_0 and T_0, we sequenced the samples at 0, 2 and 32 h, whereas in the case of G_0 and C_0, at 0, 0.5 and 8 h. The resulting data sets allowed us to characterize how the bias in the initial pools affects the pool evolution on short and long timescales.

The analysis of the AT data sets (A_0 and T_0) is depicted in Figure 3. As polymerization leads to all possible integer lengths from the initial 12-mer sequences (to the maximum range), we plotted the fraction $f_T(i)$ of nucleotide T at each position i for sequences of the same length. We then stacked the graphs so that the positions align across lengths (Figure 3A and B). The position is plotted in the 5′ to 3′ direction, the same direction as *Bst* elongates the sequences. This way, for every sequence length, the probability of finding the nucleotide T at each position can be read.

The initial pools (Figure 3A and B, top) consisted of 12-mer sequences with an overall T fraction ($\bar{f}_T$) of 0.40 for A_0 and 0.75 for T_0. The heavier bias in the T_0 pool can be explained due to DNA synthesis variability. Across positions, the distribution of the nucleotide fraction is close to homogeneous, with no apparent patterns in the initial pool that could propagate with replication.

The initial bias is countered by polymerization, and the overall pool average approaches equal nucleotide fraction for both A_0 and T_0 ($\bar{f}_T = 0.5$). This can be seen for the 2 and 32 h time points (Supplementary Data, Section IV). As most of the sequences in the pool are biased toward one nucleotide, sequences are likely to find a similarly biased template. Template-directed polymerization incorporates complementary nucleotides to the templates, inverting the bias in the newly forming strand segments. Note that all the sequences in the initial pool are 12-mer and that primer and template is a notation that solely depends on the direction of elongation; i.e. *Bst* adds nucleotides to the 3′ end of the primer (Figure 1).

Periodicity of fast AT replicators

While the pool-averaged bias was homogenized, in-strand positional biases were amplified. Due to the 5′–3′ direction of the polymerase, any bias at priming the first 12 nt at the 5′ terminus will be preserved over the complete reaction period. Additionally, since the nucleotides added are mostly complementary, the nascent segment will be inversely biased. These

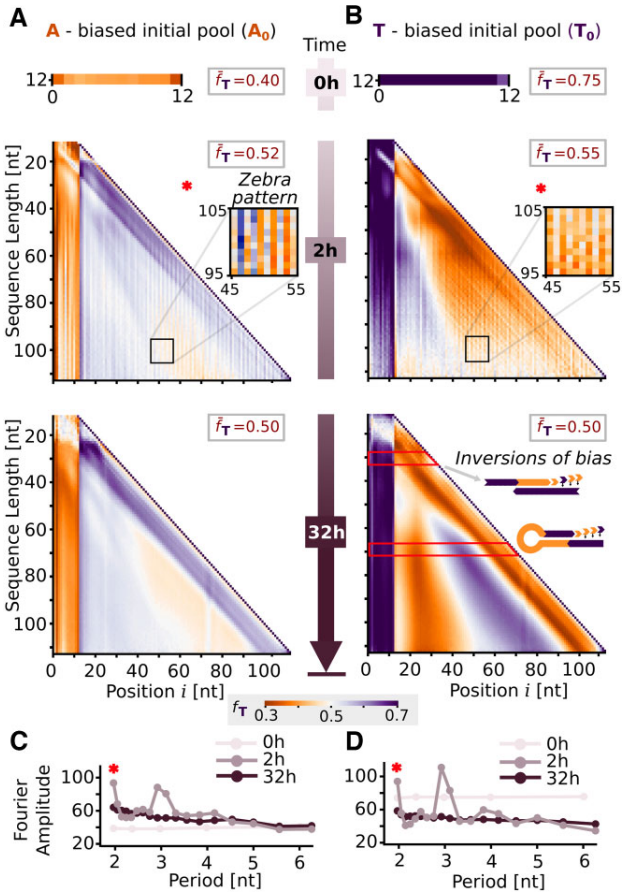


Figure 3. Effects of initial bias and elongation on sequence composition for AT (A_0 and T_0 experiments). (**A**, **B**) Evolution of nucleotide fraction f_T across sequence lengths and positions in sequences for the initial pool, an early time point and a late time point (0, 2 and 32 h). A-rich regions ($f_T \ll 0.5$) are represented in orange and T-rich regions ($f_T \gg 0.5$) in purple. The initially biased average pool nucleotide fraction $\bar{f}_T$ is countered as the pool undergoes polymerization, homogenizing to 0.5 at later time points. The first 12 nucleotides at the 5′ end retain the initial sequence bias for all graphs, due to the directionality of the polymerization mechanism (5′–3′). In addition, an inverse bias at 3′ is explained by pool templation from the biased pool. For the 2 h time point, horizontally alternating 'zebra' patterns of f_T are visible, illustrated by the insets with increased contrast. At 32 h, gradients of alternating nucleotide fraction suggest self-complementarity, possibly a consequence of self-templation. (**C**, **D**) Periodicity is plotted as the amplitude of the Fourier modes of a discrete Fourier transform performed on the position-dependent conditional probability of A for the 50-mer long sequences (Supplementary Data, Section XI.C). The fast replicator sequences from the 2 h early time point display patterns with period 2 nt, matching the zebra patterns of the nucleotide fraction graphs, as well as period 3 nt.

bias inversions are observed as starting 12-mer columns for the 2 and 32 h time points, for both of the analyzed pools.

Fast replicators, corresponding to sequences observed at early time points, feature patterned structure. They display a zebra pattern, visible through the vertical stripes indicating alternating average nucleotide fractions (Figure 3A and B, insets). To understand the interdependence between in-strand sequence motifs, we calculated a matrix that correlates the nucleotide fraction at each position to all positions of each respective sequence for sequences of length 50 (Supplementary Data, Section XI). The $\bar{f}_T$ plots do not allow to do so as they

average over all sequences of the same length. The correlation matrices for 2 h time points, for both A_0 and T_0, revealed a diagonal correlation indicative of periodicity. To obtain the dominant period of the patterns, a discrete Fourier transform (Supplementary Data, Section XI.C) was applied to every row of the correlation matrices and averaged across all rows and sequences (Figure 3C and D). The graphs spike at periods 2 and 3 nt above the baseline Fourier amplitude of 50, which random sequences would display (the baseline equals the average pool nucleotide fraction in percent). Fast replicators display a period of length 2 nt, matching the zebra patterns of the nucleotide fraction graphs. Additionally, a periodicity of length 3 nt is revealed.

After 32 h of polymerization, the zebra patterns in the fraction of T have been replaced by smooth gradients. A reason for this may be that the fast replicators have elongated even more and are no longer captured by sequencing analysis. The gradients are antisymmetric around the center, corresponding to alternating inversions of bias. This indicates self-complementarity, suggesting self-templation through the formation of hairpins as a mechanism of elongation. Self-templation is favored over pool templation when possible since it is kinetically more likely to find a complementary region within the proximity of the same molecule than within another molecule of the pool (Figure 1). Furthermore, the emergence of self-complementarity at the late time point suggests its possible adverse effect on replication, causing certain sequences to be left behind, as these sequences form stable, fully bound duplexes.

Similar patterns in GC pools, but lack of 2 nt periodicity for fast replicators

Similarly to the AT experiments, the G_0 and C_0 samples were analyzed with NGS (Figure 4). The three time points chosen in this case were adapted to the faster GC elongation kinetics. The initial pools had symmetric biases, with $\bar{f}_{C,G_0} = 0.30$ and $\bar{f}_{C,C_0} = 0.69$ for C_0. In the case of the polymerized pools, the sequences obtained were overall shorter than in the case of the AT data sets, even for the later time points. This may be due to a combination of the different polymerization dynamics and a lower sequencing efficiency for GC samples (Supplementary Data, Section III), which yields fewer and lower quality reads for a similar initial concentration. For this reason, the GC graphs (Figure 4A and B) are noisier and have shorter maximum length.

For the earlier time point, at 0.5 h incubation time, the alternating vertical stripes that indicated zebra patterns in the AT graphs are not present. The positional dependences within sequences of a specific length were analyzed by conditional probability graphs, which revealed periodicity in GC samples as they did for AT (Supplementary Data, Section XI). The Fourier transform graphs, unlike the AT ones, lack 2 nt periodicity while still displaying an increased periodicity of 3 nt (Figure 4C and D). This indicates that 3 nt periodicity is a feature of fast replicators independent of the initial pool type.

The inversion of bias both on the 5′ to 3′ end and in the intermediate region is evident for both the 0.5 and 8 h time points as in AT. These can be explained with the pool and self-templation mechanisms in addition to the directionality of polymerization. For both AT and GC, 4-mer motifs that are reverse complementary display similar abundances after polymerization has occurred, which can be seen by the symmetry

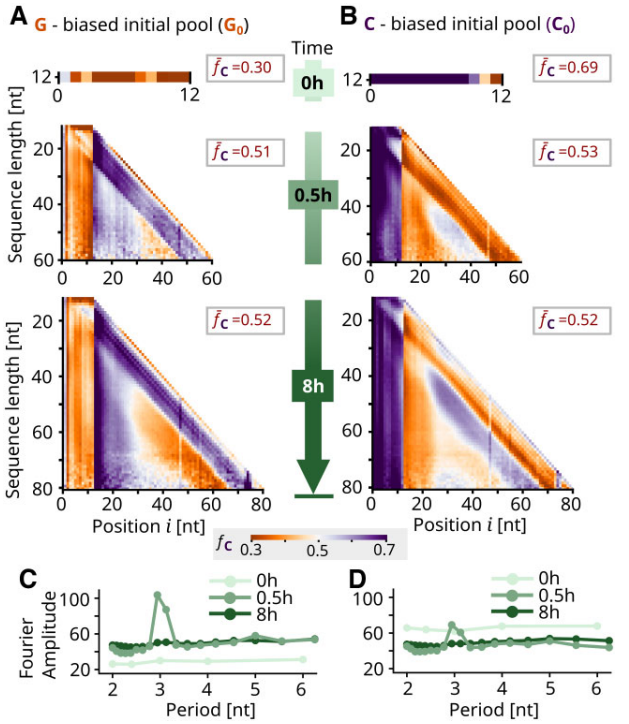


Figure 4. Results for GC (G_0 and C_0 experiments). (**A**, **B**) The nucleotide fraction f_C for the 0, 0.5 and 8 h time points was decomposed by length and position, which leads to graphs similar to those for AT. Again, the initially biased average pool nucleotide fraction $\bar{f}_C$ is homogenized with time and the first 12 nt retain the initial bias, while the following segment is inversely biased due to pool templation. However, no zebra patterns are visible in the early 0.5 h time point. (**C**, **D**) The Fourier modes (for G, 50-mer) confirm this absence of 2 nt periodicity but do indicate 3 nt periodicity.

in the graphs (Supplementary Data, Section X). The increase in overall pool complementarity leads to the convergence of the pool average nucleotide fraction to $\bar{f}_C = 0.52$ after 8 h for both experiments.

Mechanistic insights

For elongation to occur, two sequences need to form overlap duplexes or a sequence needs to self-template. However, if the resulting duplex is excessively stable after replication, it hinders strand separation and further replication, effectively leading to the sequences being left behind. To understand sequence evolution, we analyzed both early and late time points, aiming to distinguish characteristics of fast replicators from left-behind sequences. Late time points reveal antisymmetric bias inversion regions indicative of fully self-bound sequences, which are too stable to replicate and therefore remain in the left-behind pool.

To gain insights into how self-complementarity evolves during replication, we analyzed the longest potentially self-complementary region in each sequence (Figure 5A). This was achieved by comparing a strand's sequence from the 3′ end to the 5′ end and identifying the longest complementary overlap. The results were then averaged among sequences of the same length for both AT and GC pools. To establish a reference point, a random pool was generated with a nucleotide fraction of $\bar{f}_{T/C,pool} = 0.50$. This reference provides a baseline for the maximum length of self-complementary regions in the

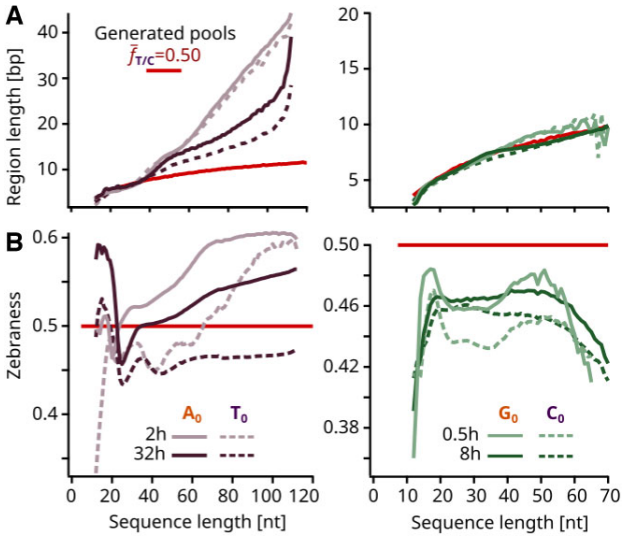


Figure 5. Analysis of self-complementarity and zebraness revealed inverse effects between AT and GC experiments. (**A**) Longest self-complementary regions that were found for each sequence, plotted averaged per sequence length. AT pools (left panel) displayed higher self-complementarity compared to a randomly generated homogeneous pool, particularly for sequences longer than 40 nt. No significant deviations from a randomly generated pool were present in the GC pools (right panel). (**B**) Zebraness by sequence length, defined as the fraction of alternating 2-mer motifs (XY, YX). In the case of AT experiments (left panel), the 2 h early time point sequences possessed a higher average zebraness than their 32 h late time point counterparts. Additionally, the zebraness was higher for longer sequences, suggesting that 2 nt periodicity is present in fast AT replicators. In contrast, GC samples (right panel) had a generally lower zebraness, consistently below 0.5. Furthermore, the zebraness decreased for longer strands, indicating that the bulky non-alternating 2-mer motifs (XX, YY) are favored for the fast GC replicators.

absence of pool- or sequence-level patterns. AT sequences exhibit significantly longer self-complementary regions, particularly among fast replicators. Conversely, GC regions align perfectly in length with the generated reference sample.

Notably, the longer AT self-complementary regions of fast replicators coincide with an increased 2 nt periodicity. To understand this characteristic, we introduced the concept of zebraness, defined as the fraction of alternating (zebra) 2-mer motifs (XY or YX) (41) (Figure 5B). Correspondingly, 2-mer bulky motifs are defined as homodimers (XX or YY). The findings reveal that in AT sequences, zebraness of fast replicators is higher than 0.5 and consistently exceeds that of left-behind sequences. In contrast, for GC sequences, zebraness consistently falls below 0.5. Thus, zebraness appears to confer a replicative advantage to AT sequences, while not benefiting GC sequences. The difference between the AT and GC fast replicators can be explained by the intrinsic differences in the stacking energies ΔG of zebra—averaged from the motifs XY and YX—and bulky XX/YY motifs that have been determined in (42) (literature values ΔG^{SH}, all in kcal/mol):

$$\Delta G^{\mathrm{zebra}}_{\mathrm{AT}} = (\Delta G^{\mathrm{SH}}_{\mathrm{AT/TA}} + \Delta G^{\mathrm{SH}}_{\mathrm{TA/AT}})/2 = -0.73,$$

$$\Delta G^{\mathrm{bulky}}_{\mathrm{AT}} = \Delta G^{\mathrm{SH}}_{\mathrm{AA/TT}} = -1.00,$$

$$\Delta G^{\mathrm{zebra}}_{\mathrm{GC}} = (\Delta G^{\mathrm{SH}}_{\mathrm{GC/CG}} + \Delta G^{\mathrm{SH}}_{\mathrm{CG/GC}})/2 = -2.20,$$

$$\Delta G^{\mathrm{bulky}}_{\mathrm{GC}} = \Delta G^{\mathrm{SH}}_{\mathrm{GG/CC}} = -1.84.$$

Thus, for AT, bulky motifs are more stabilizing than zebra motifs, whereas for GC the opposite is true, with the stacking energy difference $\Delta G^{\mathrm{bulky}} - \Delta G^{\mathrm{zebra}}$ equaling 0.27 kcal/mol for AT versus −0.36 kcal/mol for GC. Stacking energy of neighboring nucleotide pairs is the main contributor for duplex stability (43), explaining its strong effect on sequence evolution. The sequences rich in the most destabilizing motif type replicate the fastest into very long strands. This prevents them from being stuck in very stable secondary structures and renders them more accessible for several rounds of priming. Additionally, long zebra regions are fully self-complementary, allowing a single strand to have many possible transient fold-back conformations and undergo several rounds of self-templation, which could be a replication mode of AT fast replicators. Due to the elevated stability of the G:C base pair in comparison to the A:T base pair, in addition to stacking, for GC this mode of replication might lead to overly stable self-folded conformations impeding their status as fast replicators.

The enhanced 3 nt periodicity is a distinct characteristic of fast replicators in both AT and GC experiments (Figures 3 and 4C and D). We propose a mechanism by which 3 nt periodicity balances the formation of duplexes for elongation with the avoidance of overly stable ones, enabling fast replication. Unlike zebra sequences, which are reverse complementary to themselves, 3 nt periodic sequences cannot as easily self-template through hairpin formation unless they are composed of (at least) two regions with repeating reverse complementary 3-mer motifs. However, their periodic regions offer an increased amount of potential binding sites for reverse complementary periodic regions of other sequences, allowing for the formation of duplex regions for elongation to start. Two subpopulations of sequences with reverse complementary periodic 3-mer motifs may form efficient primer–template pairs that rapidly bind, elongate and separate again, effectively cooperating to achieve fast replication. The advantage of 3 nt periodicity over longer 4 or 5 nt periodicities is not only the higher amount of potential binding sites, but more importantly the small sequence space associated with 3-mers. This results in only four '3 nt periodic partner' subpopulations (containing periodic motifs AAT/ATT, ATA/TAT, TAA/TTA and AAA/TTT) instead of the combination of six pool-templating plus four self-templating subpopulations for 4 nt periodicities or a total of sixteen '5 nt periodic partner' subpopulations in the pool.

While this study focused on isolating effects of replication in binary systems, we performed a supplementary experiment to check whether the conclusions drawn in these simpler and more accessible systems also apply to more complex full-alphabet experiments. Analyzing the replication of a 4 nt data set ($ATGC_0$) at two time points (0 and 64 h) recovered patterns found in binary systems (Supplementary Data, Section XIII). For example, the nucleotide fractions also revealed positional biases, due to the combination of templation and directionality. The polymerized sequences exhibited longer self-complementary regions than a generated random pool, as it was the case for AT pools. These regions displayed an increase in the prevalence of AT and TA motifs with increasing length. Similarly to the late time point pools of the binary systems, the Fourier transform of the 50 nt sequences did not reveal any periodicity for ATGC, which indicates that the fast replicators were not captured with this later time point. The full-alphabet analysis thereby demonstrates that mecha-

nisms and effects of replication isolated in binary systems are recoverable in ATGC data.

Conclusion

We demonstrated that in following templated replication, pools display a positional bias and the average pool nucleotide fractions become more homogeneous. Replication from two independently synthesized initial pools with the same bias resulted in reproducible length distributions, average pool nucleotide fractions and sequence structure (Supplementary Data, Section VI).

We experimentally verified that compositional diversity, represented by the average pool nucleotide fraction, arises from biased binary pools via templated replication. This is a necessary characteristic for the exploration of sequence space with the possibility of generating a functional sequence. Similar conclusions have previously been described for binary DNA systems *in silico* (10), particularly for templated ligation.

Simultaneously, the replication of an initially biased pool resulted in regions in the replicated sequence that possess the same or the symmetric bias, which alternate and balance each other on average. This allows for a biased exploration of subsections of sequence space with structured sequences, without restricting the sequence space to a subset of similar sequences. Different nucleotide biases have been shown to correlate with enrichment of different secondary structures (20), implying that the sequences obtained from our templated replication may exhibit a diverse range of secondary structure, which is in turn correlated with functionality.

Symmetry breaking, triggered by the selection for the reverse complement due to templation mechanisms, has been experimentally described for templated ligation. In a previous study (1) where binary AT pools were studied, two different subpopulations of sequences were found to contain a high amount of reverse complement sequences, with different nucleotide biases being enriched for each subpopulation (an A-rich and a T-rich). Indeed, we observed a comparable behavior within single sequences.

We also found that highly periodic sequences are replicated faster, interestingly amplifying a periodic trimer structure in all studied pools. We attribute this to the potential emergence of cooperative sequence networks made up of subpopulations within the pools. These subpopulations would be characterized by reverse-complementary 3-mer periodic motif sequences that would cross-catalyze each other's rapid elongation.

Besides this agreement in 3 nt structure, the 2 nt periodicity differed for the two binary systems investigated. AT pools favored the 2 nt zebra motifs AT and TA, whereas GC pools preferred the bulky motifs GG and CC, likely due to intrinsic differences in stacking energies. Our findings, especially of the high self-complementarity in long AT sequences (Figure 5), support the mechanism of 'hairpin elongation' for repetitive DNA, as previously suggested (44). Repetitive DNA strands possess a high number of potential fold-back sites for hairpin formation. Repeated complete or partial melting, possibly induced by the strand displacing activity of *Bst*, alternating with hairpin formation and self-templation, would quickly elongate highly repetitive sequences.

In this study, we employed an experimental model system to provide insight into the role of replication as a mechanism of selection. Using a protein-based replication system with strand displacement (*Bst*), we identified which sequence patterns emerged as the fittest by analyzing the fast replicators. In addition, we characterized the dependence of the emergent structure on the initial pool. Overall, our findings contribute to elucidate the steps involved in the molecular evolution of short unstructured nucleic acids into long functional sequences.

Data availability

All data and code relevant to the study are available at https://doi.org/10.6084/m9.figshare.23674773, uploaded as supplementary information or, in the case of the raw sequencing FASTQ data, provided in the NCBI repository PRJNA965926 available at http://www.ncbi.nlm.nih.gov/bioproject/965926.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank Christof B. Mast, Sreekar Wunnava and Paula Aikkila for comments on the manuscript, Annalena Salditt for helpful discussions on data analysis, and Stefan Krebs and Marlis Fischalek at the Gene Center Munich for their help with the library preparation and the sequencing of the samples.

Author contributions: Project conception: A.C.S. and D.B. Research design: A.C.S., F.T.D. and D.B. Methodology development: A.C.S. and F.T.D. Experiments: A.C.S., F.T.D. and Z.M. Data analysis: A.C.S. and F.T.D. Programming: F.T.D. Manuscript writing: A.C.S. and F.T.D. Manuscript reviewing: A.C.S., F.T.D. and D.B. Supervision: A.C.S. and D.B. Funding acquisition: D.B.

Funding

European Research Council [787356]; Deutsche Forschungsgemeinschaft [364653263 and 390783311]; Center for NanoScience. Open access funding provided by the Volkswagen Foundation.

Conflict of interest statement

None declared.

References

1. Kudella,P.W., Tkachenko,A.V., Salditt,A., Maslov,S. and Braun,D. (2021) Structured sequences emerge from random pool when replicated by templated ligation. *Proc. Natl Acad. Sci. U.S.A.*, **118**, e2018830118.
2. Salditt,A., Karr,L., Salibi,E., Le Vay,K., Braun,D. and Mutschler,H. (2023) Ribozyme-mediated RNA synthesis and replication in a model Hadean microenvironment. *Nat. Commun.*, **14**, 1495.
3. Zhou,L., O'Flaherty,D.K. and Szostak,J.W. (2020) Assembly of a ribozyme ligase from short oligomers by nonenzymatic ligation. *J. Am. Chem. Soc.*, **142**, 15961–15965.
4. Zhou,L., O'Flaherty,D.K. and Szostak,J.W. (2020) Template-directed copying of RNA by non-enzymatic ligation. *Angew. Chem*, **132**, 15812–15817.

5. Ding,D., Zhou,L., Mittal,S. and Szostak,J.W. (2023) Experimental tests of the virtual circular genome model for nonenzymatic RNA replication. *J. Am. Chem. Soc.*, **145**, 7504–7515.
6. Dass,A.V., Wunnava,S., Langlais,J., von der Esch,B., Krusche,M., Ufer,L., Chrisam,N., Dubini,R.C., Gartner,F., Angerpointner,S., *et al.* (2023) RNA oligomerisation without added catalyst from 2′,3′-cyclic nucleotides by drying at air–water interfaces. *ChemSystemsChem*, **5**, e202200026.
7. Dirscherl,C.F., Ianeselli,A., Tetiker,D., Matreux,T., Queener,R.M., Mast,C.B. and Braun,D. (2023) A heated rock crack captures and polymerizes primordial DNA and RNA. *Phys. Chem. Chem. Phys.*, **25**, 3375–3386.
8. Ferris,J.P. and Ertem,G. (1990) Oligomerization reactions of deoxyribonucleotides on montmorillonite clay: the effect of mononucleotide structure, phosphate activation and montmorillonite composition on phosphodiester bond formation. *Origins Life Evol. Biosph.*, **20**, 279–291.
9. Ding,D., Zhou,L., Giurgiu,C. and Szostak,J.W. (2022) Kinetic explanations for the sequence biases observed in the nonenzymatic copying of RNA templates. *Nucleic Acids Res.*, **50**, 35–45.
10. Derr,J., Manapat,M.L., Rajamani,S., Leu,K., Xulvi-Brunet,R., Joseph,I., Nowak,M.A. and Chen,I.A. (2012) Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences. *Nucleic Acids Res.*, **40**, 4711–4722.
11. Levy,M. and Miller,S.L. (1998) The stability of the RNA bases: implications for the origin of life. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 7933–7938.
12. Levy,M., Miller,S.L. and Oró,J. (1999) Production of guanine from NH_4CN polymerizations. *J. Mol. Evol.*, **49**, 165–168.
13. Oba,Y., Takano,Y., Naraoka,H., Watanabe,N. and Kouchi,A. (2019) Nucleobase synthesis in interstellar ices. *Nat. Commun.*, **10**, 8–15.
14. Cleaves,H.J., Nelson,K.E. and Miller,S.L. (2006) The prebiotic synthesis of pyrimidines in frozen solution. *Naturwissenschaften*, **93**, 228–231.
15. Miyakawa,S. and Ferris,J.P. (2003) Sequence- and regioselectivity in the montmorillonite-catalyzed synthesis of RNA. *J. Am. Chem. Soc.*, **125**, 8202–8208.
16. Tanner,N.K. (1999) Ribozymes: the characteristics and properties of catalytic RNAs. *FEMS Microbiol. Rev.*, **23**, 257–275.
17. Doudna,J.A. and Cech,T.R. (2002) Site-specific RNA self-cleavage: the chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
18. Schultes,E.A. and Bartel,D.P. (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.
19. Kennedy,R., Lladser,M.E., Wu,Z., Zhang,C., Yarus,M., De Sterck,H. and Knight,R. (2010) Natural and artificial RNAs occupy the same restricted region of sequence space. *RNA*, **16**, 280–289.
20. Stich,M., Briones,C. and Manrubia,S.C. (2008) On the structural repertoire of pools of short, random RNA sequences. *J. Theor. Biol.*, **252**, 750–763.
21. De Duve,C. (2005) The onset of selection. *Nature*, **433**, 581–582.
22. Ianeselli,A., Atienza,M., Kudella,P.W., Gerland,U., Mast,C.B. and Braun,D. (2022) Water cycles in a Hadean CO_2 atmosphere drive the evolution of long DNA. *Nat. Phys.*, **18**, 579–585.
23. Tkachenko,A.V. and Maslov,S. (2018) Onset of natural selection in populations of autocatalytic heteropolymers. *J. Chem. Phys.*, **149**, 134901.
24. Göppel,T., Rosenberger,J.H., Altaner,B. and Gerland,U. (2022) Thermodynamic and kinetic sequence selection in enzyme-free polymer self-assembly inside a non-equilibrium RNA reactor. *Life*, **12**, 567.
25. Fellermann,H., Tanaka,S. and Rasmussen,S. (2017) Sequence selection by dynamical symmetry breaking in an autocatalytic binary polymer model. *Phys. Rev. E*, **96**, 062407.
26. Benkovic,S.J., Valentine,A.M. and Salinas,F. (2001) Replisome-mediated DNA replication. *Annu. Rev. Biochem.*, **70**, 181–208.
27. Simmel,F.C., Yurke,B. and Singh,H.R. (2019) Principles and applications of nucleic acid strand displacement reactions. *Chem. Rev.*, **119**, 6326–6369.
28. Tupper,A.S. and Higgs,P.G. (2021) Rolling-circle and strand-displacement mechanisms for non-enzymatic RNA replication at the time of the origin of life. *J. Theor. Biol.*, **527**, 110822.
29. Zhou,L., Kim,S.C., Ho,K.H., O'Flaherty,D.K., Giurgiu,C., Wright,T.H. and Szostak,J.W. (2019) Non-enzymatic primer extension with strand displacement. *eLife*, **8**, e51888.
30. Mariani,A., Bonfio,C., Johnson,C.M. and Sutherland,J.D. (2018) pH-driven RNA strand separation under prebiotically plausible conditions. *Biochemistry*, **57**, 6382–6386.
31. Ianeselli,A., Mast,C.B. and Braun,D. (2019) Periodic melting of oligonucleotides by oscillating salt concentrations triggered by microscale water cycles inside heated rock pores. *Angew. Chem.*, **131**, 13289–13294.
32. Christian,T.V. and Konigsberg,W.H. (2018) Single-molecule FRET reveals proofreading complexes in the large fragment of *Bacillus stearothermophilus* DNA polymerase I. *AIMS Biophys.*, **5**, 144.
33. Phang,S.-M., Teo,C.-Y., Lo,E. and Wong,V. W.T. (1995) Cloning and complete sequence of the DNA polymerase-encoding gene (BstpolI) and characterisation of the Klenow-like fragment from *Bacillus stearothermophilus*. *Gene*, **163**, 65–68.
34. Agustriana,E., Nuryana,I., Laksmi,F.A., Dewi,K.S., Wijaya,H., Rahmani,N., Yudiargo,D.R., Ismadara,A., Helbert, Hadi,M.I., *et al.* (2022) Optimized expression of large fragment DNA polymerase I from *Geobacillus stearothermophilus* in *Escherichia coli* expression system. *Prep. Biochem. Biotechnol.*, **53**, 384–393.
35. Afgan,E., Baker,D., Batut,B., van den Beek,M., Bouvier,D., Čech,M., Chilton,J., Clements,D., Coraor,N., Grüning,B.A., *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.
36. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
37. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
38. Rosenberger,J.H., Göppel,T., Kudella,P.W., Braun,D., Gerland,U. and Altaner,B. (2021) Self-assembly of informational polymers by templated ligation. *Phys. Rev. X*, **11**, 031055.
39. Qian,J., Ferguson,T.M., Shinde,D.N., Ramírez-Borrero,A.J., Hintze,A., Adami,C. and Niemz,A. (2012) Sequence dependence of isothermal DNA amplification via EXPAR. *Nucleic Acids Res.*, **40**, e87.
40. Pavlov,A.R., Pavlova,N.V., Kozyavkin,S.A. and Slesarev,A.I. (2004) Recent developments in the optimization of thermostable DNA polymerases for efficient applications. *Trends Biotechnol.*, **22**, 253–260.
41. Göppel,T., Rosenberger,J.H., Altaner,B. and Gerland,U. (2022) Thermodynamic and kinetic sequence selection in enzyme-free polymer self-assembly inside a non-equilibrium RNA reactor. *Life*, **12**, 567.
42. SantaLucia,J. and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415–440.
43. Yakovchuk,P., Protozanova,E. and Frank-Kamenetskii,M.D. (2006) Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.*, **34**, 564–574.
44. Ogata,N. and Morino,H. (2000) Elongation of repetitive DNA by DNA polymerase from a hyperthermophilic bacterium *Thermus thermophilus*. *Nucleic Acids Res.*, **28**, 3999–4004.

Received: July 13, 2023. **Revised:** November 15, 2023. **Editorial Decision:** November 25, 2023. **Accepted:** November 30, 2023

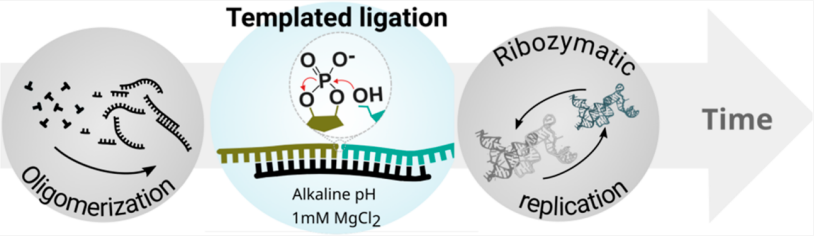
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS
Article

High-Fidelity RNA Copying via 2′,3′-Cyclic Phosphate Ligation

Adriana Calaça Serrão,[§] Sreekar Wunnava,[§] Avinash V. Dass, Lennard Ufer, Philipp Schwintek, Christof B. Mast, and Dieter Braun*

Cite This: https://doi.org/10.1021/jacs.3c10813
Read Online



ABSTRACT: Templated ligation offers an efficient approach to replicate long strands in an RNA world. The 2′,3′-cyclic phosphate (>P) is a prebiotically available activation that also forms during RNA hydrolysis. Using gel electrophoresis and high-performance liquid chromatography, we found that the templated ligation of RNA with >P proceeds in simple low-salt aqueous solutions with 1 mM $MgCl_2$ under alkaline pH ranging from 9 to 11 and temperatures from −20 to 25 °C. No additional catalysts were required. In contrast to previous reports, we found an increase in the number of canonical linkages to 50%. The reaction proceeds in a sequence-specific manner, with an experimentally determined ligation fidelity of 82% at the 3′ end and 91% at the 5′ end of the ligation site. With splinted oligomers, five ligations created a 96-mer strand, demonstrating a pathway for the ribozyme assembly. Due to the low salt requirements, the ligation conditions will be compatible with strand separation. Templated ligation mediated by 2′,3′-cyclic phosphate in alkaline conditions therefore offers a performant replication and elongation reaction for RNA on early Earth.

INTRODUCTION

Nucleic acid replication is essential for the propagation of genetic information and therefore is a central step for the origin of life.[1] An early form of molecular evolution that preceded catalytic polymers (i.e., enzymes and ribozymes) required a nonenzymatic copying mechanism. While primer extension by the addition of mono-, di-, or tri-nucleotides has been demonstrated to copy shorter sequences,[2−5] its processivity is limiting, and the combination with strand separation is a considerable hurdle. A simple way to overcome these issues could be templated ligation, which reduces the number of steps-per-length required to generate an oligonucleotide, and if possible, operating at low magnesium concentration would offer an easier strand separation, particularly when coupled to nonequilibrium microenvironments with salt and pH cycling.[6] Additionally, ligation chain reactions are known to offer exponential replication.[7]

The state-of-the-art method for templated ligation makes use of phosphoramidates, such as phosphorimidazolides.[8,9] The need for a separate (*ex situ*) presynthesis step with condensing agents, coupled with their short half-life, reduces their prebiotic likelihood. Furthermore, *in situ* activation with prebiotically plausible organocatalysts in ligation-compatible scenarios has not been shown. Disregarding the need for multiple synthesis steps required to make the phosphorimidazolides,[10,11] imidazole-activated oligonucleotides are less reactive than their mononucleotide counterparts, lowering the yield of templated ligation compared to that of polymerization.[8] Moreover, studies demonstrating the assembly of a long catalytic RNA by templated ligation of imidazole-activated oligonucleotides required a high concentration of Mg^{2+}, which leads to product inhibition.[8,9,12] This strongly supports the need for an efficient ligation system compatible with strand separation.

The quest for such a system led us to cyclic phosphates since they generate short oligomers in the dry state,[13−15] which retain the active >P ends and could act as raw material for ligation. More importantly, they represent a simple and endogenous activated group, minimizing the need for complex multistep synthesis and *ex situ* activation. 2′,3′-Cyclic phosphate (>P) endings are likely readily available in the prebiotic pool since they are the primary product of prebiotic nucleotide synthesis[16] and phosphorylation reactions.[17,18] Moreover, ribozymatic[19] and alkaline hydrolysis of RNA strands via transesterification[20−22] produce >P ends, which

Received: October 2, 2023
Revised: February 5, 2024
Accepted: February 6, 2024

are substrates for ribozymes catalyzing phosphodiester bond formation.[23,24]

Hydrolysis of >P results in 2′- or 3′-monophosphate, the recycling of which, i.e., the recyclization to activated oligonucleotides with >P, under prebiotic conditions was also demonstrated,[17,18,23] suggesting a way for the *in situ* recycling of hydrolyzed substrates. Biochemical protocols using 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) at low temperatures are also common.[24−26] The widespread presence of these activated phosphate groups in established prebiotic pathways and their relatively longer half-life motivated us to investigate their role in ligating RNA oligomers in a templated setting and their relevance to early RNA replication.

Previous work with >P containing oligonucleotides has demonstrated template copying only through ligation with DNA/RNA chimeras, resulting in low yield and a predominance of 2′−5′ linkages.[26,27] Nontemplated ligation of random sequences with >P has also been shown to proceed in eutectic phase at low rates.[28] However, the potential of ligation of RNA sequences with >P ends for quantitative genetic copying remained largely unexplored.

Our study explored the ligation of a completely RNA-based system in simple conditions devoid of additional organocatalysts. We investigated the impact of reaction conditions on yield, kinetics, and sequence fidelity, achieving a maximum yield at 5 °C and pH 10. This was achieved under reduced salt conditions, enabling compatibility of the settings with strand separation. Moreover, we observed an increased ratio of 50% of the canonical 3′−5′ linkages at the ligation site, an improvement over previously reported aqueous condensations involving >P.[26,29,30] Additionally, we demonstrated that the reaction is highly sequence-selective, even for a single nucleotide at the ligation site. Finally, we could demonstrate that multistep ligations within a splinted RNA system using >P can generate long RNA molecules on the length scale of 100 nucleotides through a cross-templating reaction.

This work serves as a proof of principle for nonenzymatically replicating and generating long RNA, in a sequence-specific manner using a simple and ubiquitous phosphate chemistry, albeit at conditions of elevated pH. Such alkaline conditions are however common in fresh-water volcanic lakes,[31,32] indicating it is a possible scenario on early Earth.

RESULTS AND DISCUSSION

Learning from our previous work[15] and other studies on ligation with >P,[26,27,29,33] we decided to investigate the ligation reaction with >P under alkaline conditions. This was also corroborated by a broader pH screen (Supporting Information S4, Figure S4.1). All the reactions were performed in aqueous alkaline conditions with 50 mM 2-(cyclohexylamino)ethane-1-sulfonic acid (CHES) buffer and an equimolar concentration of the participating strands. A possible influence of the buffer itself was not supported in a buffer screen at pH 9 (Supporting Information S4, Figure S4.2). The two ligating strands are labeled as primer *a* and *b* and the template, *BA*. All the reactions were performed in the presence of 1 mM Mg^{2+} as it was found to be optimal (Supporting Information S4, Figure S4.3). The analysis of the ligation reaction was done either by polyacrylamide gel electrophoresis (PAGE, Figure 1c; see Supporting Information S1.3 and S2.2) or high-performance liquid chromatography (HPLC, Figures 1d, Supporting Information S1.4 and S2.1), or a combination of both.

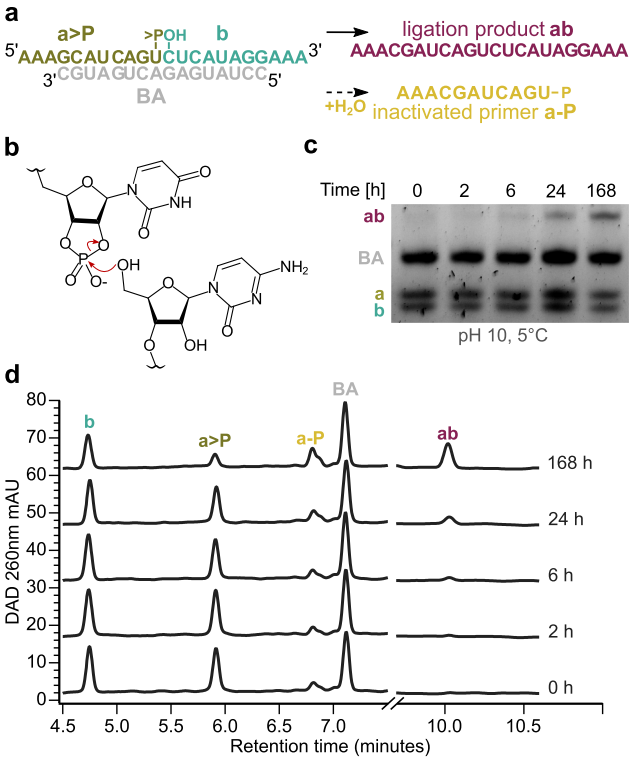


Figure 1. Nonenzymatic template-directed ligation of short RNA strands. (a) Schematics of reaction design. Both primers *a* and*b* bind on the complementary template, *BA*. The primer *a* has a 2′,3′-cyclic phosphate, while *b* contains a 5′-OH group. (b) 5′-OH performs a nucleophilic attack on the cyclic phosphate group and forms a phosphodiester bond between the two primers, leading to the ligation product strand *ab*. As a side reaction, the cyclic phosphate in *a* can also hydrolyze, rendering *a* inactive. (c) Denaturing polyacrylamide gel electrophoresis (PAGE) analysis of ligation reaction over time. Reaction contained 1 μM primers, 1 μM template, 50 mM CHES, pH 10, and 1 mM $MgCl_2$ at 5 °C. (d) Stacked HPLC chromatograms (absorbance at 260 nm) of the same reaction mixtures as in (c). The product peak increases over time as the primers get depleted.

To study the impact of temperature and pH on the ligation reaction, three temperatures of 5, 10, and 25 °C and pH 9, 10, and 11 were tested, as shown in Figure 2. The reaction yielded both the ligation product (*ab*) and the inactivated primer side product (*a*−*P*) while consuming the two primers *a* > *P* and *b*. The template (*BA*) was not consumed by the reaction. Thus, its concentration was used to correct for small pipetting errors through normalization.

The concentration over time of the strands *a* > *P*, *b*, *ab*, *a*−*P*, for different temperature−pH combinations is plotted in Figure 2b (see Figure 2a for color-coded schematics). It is important to note that in all the experiments, the initial concentration of *a*−*P* is on average 26% of total *a*, suggesting that a part of *a* > *P* was already hydrolyzed in the stock solutions. This capped the maximum concentration of *ab* at 0.74 μM, the concentration of initial *a* > *P*. The yield of *ab* depends on the temperature and pH combination of the reaction. The highest concentration of *ab* at 7 days (0.28 μM) was obtained at 5 °C, pH 10 (Figure 2b,c, red rectangle). Calculating for the real initial amount of *a* > *P*, the yield was 38%. The obtained yield was measured under a limiting concentration of the primer with a cyclic phosphate end (*a* > *P*), meaning that the formation of hydrolyzed *a*−*P* contributed

https://doi.org/10.1021/jacs.3c10813
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

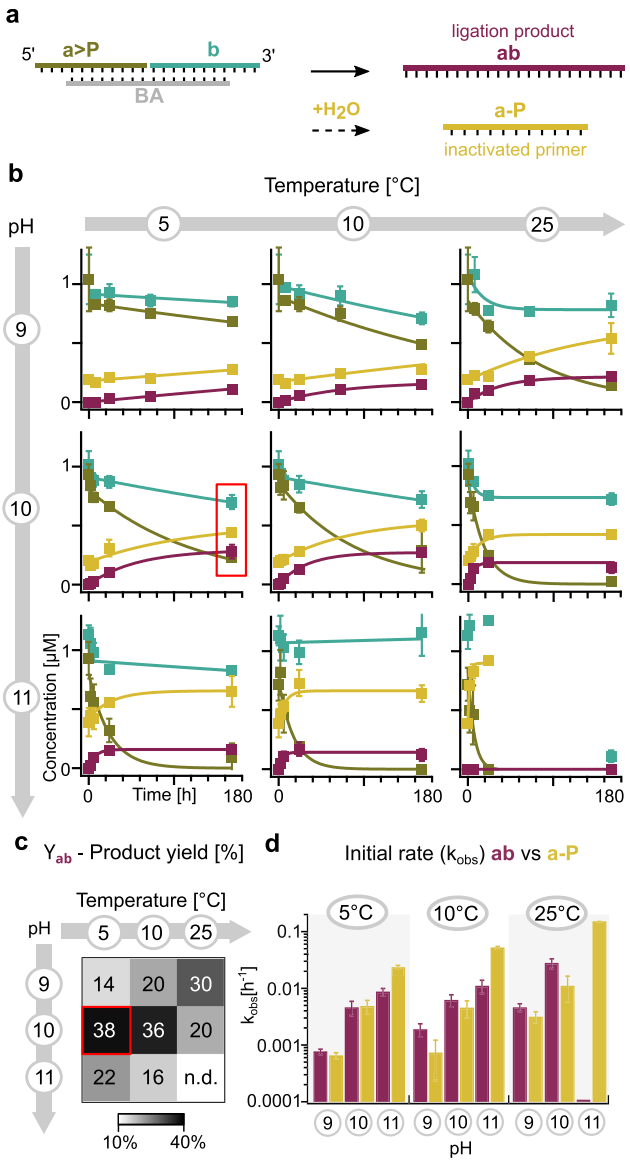


Figure 2. Kinetic study of temperature and pH influence in ligation reaction. (a) Schematics of the ligation reaction. The hydrolysis of the cyclic phosphate competes with the ligation reaction. (b) Screening the concentration of primers and product over a 7 day period for varied pH (9−11) and temperature (5, 10, 25 °C). The highest yield at 7 days is highlighted in red, for pH 10 and 5 °C. At both high temperature and pH (25 °C and pH 11), there is additional hydrolysis of the backbone, particularly after 7 days (bottom, right−most graph). The full lines correspond to the exponential fit of the data as a guide to the eye. (c) Product yield obtained (%) at 7 days for all of the conditions tested. Maximum obtained yield was for pH 10 at 5 °C (red square). This reported yield was calculated for the limiting concentration of *a* > *P*. (d) Observed initial pseudo-first order rate of product and inactive primer formation for all the conditions (Supporting Information S8). Reactions contained 1 μM primers, 1 μM template, 50 mM CHES, with varied pH, and 1 mM $MgCl_2$. Concentrations were measured with HPLC UV detection at 260 nm. Data are represented as mean ± standard deviation of three independent replicates.

to the observed partial conversion. However, the addition of excess primer *a* did not improve the yield (Supporting Information S10), indicating product inhibition, confirmed by our estimate that the product strands have an off-rate of about 40 days at 5 °C. Comparable yields have been reported for ligation with phosphorimidazolide activation under similarly low Mg^{2+} concentrations.[8]

Since the formation of the primer−primer−template complex through base-pairing proceeds at much faster time scales ($k_{on} \approx 1\ \mu M^{-1}\ s^{-1}$, literature values for similar-sized oligonucleotides[34−36]) than that of the nucleophilic attack on the cyclic phosphate, it can be assumed that the reaction is of pseudo-first order (Supporting Information S8.1). Thus, first-order kinetic rate constants (k_{obs}) for product formation and hydrolysis of cyclic phosphate were fitted to the data, see Figure 2d (Supporting Information S8, Figure S8.1). The obtained rates are between 0.001 and 0.03 h^{-1}, which are in the same order of magnitude for the ligation of native RNA at 20 mM $MgCl_2$ with 2-Me imidazolide chemistry.[9]

A few salient features of the plots in Figure 2b−d are the rates and yields of ligation (*ab*) and hydrolysis (*a*−*P*). Both the rates of ligation and hydrolysis increase with an increase in either pH or temperature, keeping the other parameter constant. Figure 2d shows that while the observed ligation rate mostly increases from pH 9−10, the inactivation rate increases from pH 10−11. At higher pH, the ligation kinetics is slightly faster; however, the final yield drops due to the competing inactivation rate. The rate of ligation is in general higher than the rate of the hydrolysis at 5 and 10 °C, while 25 °C favors the inactivation. It has been reported that low temperatures reduce the entropic cost of the ligation reaction and shift the reaction equilibrium from hydrolysis to ligation.[23] The maximum 7 day yield obtained is at pH 10 and 5 °C. Significant RNA backbone hydrolysis is observed when both temperature and pH are maximal (25 °C and pH 11).

The attack of the 5′-OH on the cyclic phosphate (Figure 1b) can form either a 2′−5′ or a 3′−5′ phosphodiester bond (Figure 3a,b, respectively). Previous studies on the polymerization and ligation with cyclic phosphates have reported varying ratios of 3′−5′ and 2′−5′ linkages, depending largely on the experimental conditions. For example, dry-state polymerization resulted in a natural linkage enrichment ratio of 2:1,[13,37] while an aqueous state (with 0.5 M diamine, pH 8, and 0 °C) was reported to lead to at least 97% of 2′−5′.[30]

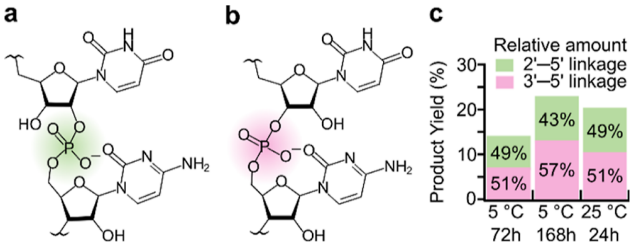


Figure 3. Linkage analysis of the reaction product *ab* through digestion with nuclease P1. Scheme of the ligation site with a 2′−5′ linkage (a, green) and a 3′−5′ linkage (b, pink). (c) Total product yield obtained for three condition sets, with the corresponding relative amounts of 2′−5′ and 3′−5′ linkage. The ligation with 2′,3′-cyclic phosphates does not exhibit regioselectivity as both linkages are equally represented for the studied conditions. After the reactions with 10 μM primers, 10 μM template, 50 mM CHES, pH 10, and 1 mM $MgCl_2$, they were digested with nuclease P1 (Supporting Information S1.5). The concentrations of the samples before and after digestion were measured with HPLC UV detection at 260 nm (Supporting Information S2.1 and S5). Data represent the mean of independent duplicates.

https://doi.org/10.1021/jacs.3c10813
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Templated cleavage and ligation at 25 °C, pH 9, and 5 mM $MgCl_2$ in an aqueous solution was reported to also show a predominance of 2′−5′ linkages (about 95%).[29] Conversely, templated ligation in the eutectic phase resulted in an excess of 3′−5′ linkages.[38] For the templated ligation reaction described here, we found no significant regioselectivity under the tested conditions (Figure 3c). This difference in comparison to previous studies is potentially due to the different systems and conditions tested. To investigate this, the reaction was quenched by ethanol precipitation, and the samples were digested with nuclease P1 following the manufacturer's protocol (Supporting Information S1.5). Nuclease P1 specifically lyses the 3′−5′ linkages, which in this case would digest all the strands *a*, *b*, and *BA* completely but digest *ab* either completely or result in a UC dimer.

The concentration of total product predigestion and UC dimer postdigestion was determined using HPLC UV absorbance (Supporting Information S2.1 and S5). We observed that both types of linkages were formed equally (Figure 3c), indicating that the reaction is not regioselective. However, a slight enrichment of the canonical linkage over time for the 5 °C conditions can be seen, possibly due to the favored hydrolysis of 2′−5′ linkages, particularly in double-stranded RNA in alkaline solutions.[39] The presence of noncanonical linkages, however, does not render the product strands obsolete. Such mixed backbone RNA has been demonstrated to still fold into functional structures.[40] The stability of the RNA duplex has been documented to be reduced for strands fully composed by 2′−5′ linkages in comparison with that of RNA with canonical linkages.[41,42] For this system, however, one single 2′−5′ linkage at the ligation site would likely not have a considerable destabilizing effect as the duplex could accommodate the structural disruption.[40] Furthermore, Supporting Information S7, Figure S7.3a shows that *ab* with a 2′−5′ linkage can still template the formation of *BA* through >*P* mediated ligation, highlighting the possibility of a replication cycle.

To answer the question if the ligation reaction is sequence-specific, we conducted the reaction with each of the 16 different nucleotide combinations (four each on the 3′ end of *a* and the 5′ end of *b*) at the ligation site (Figure 4a). Additionally, we tested two different templates, one with GA and the other with UA at the position complementary to the ligation site. Except for the ligation site, the remainder of the sequence was fully complementary to the template. This is similar to the approaches undertaken for the fidelity calculation for ligases.[43−45] These reactions were carried out at pH 10 and 5 °C for 168 h, i.e., conditions with the highest yield of ligation among the ones tested (Figure 2c). Figure 4b,c shows that of all the combinations of the primers tested, the highest yield of *ab* is obtained for the sequence with the correct nucleotides at the ligation site (marked in red, CU for the template GA and AU for the template UA). However, mismatched ligations did occur, albeit with much lower relative yields, which were especially reduced for the template UA.

Interestingly, G/U wobble pairing of the 5′ U of primer *b* at the ligation site led to a high relative yield (78%) compared to that of the complementary primers for the template GA. When considering one single mutation at the ligation site either on *a* or *b*, the reaction yield drops on average by 91 or 82%, respectively, relative to the nonmutated complex. We term this the ligation fidelity for one mutation. If we consider two mutations at the ligation site, the average experimental yield drops to 12% (template AG) and 5% (template AU), and thus, the ligation fidelity for the respective template is 88 and 95% for two mutations.

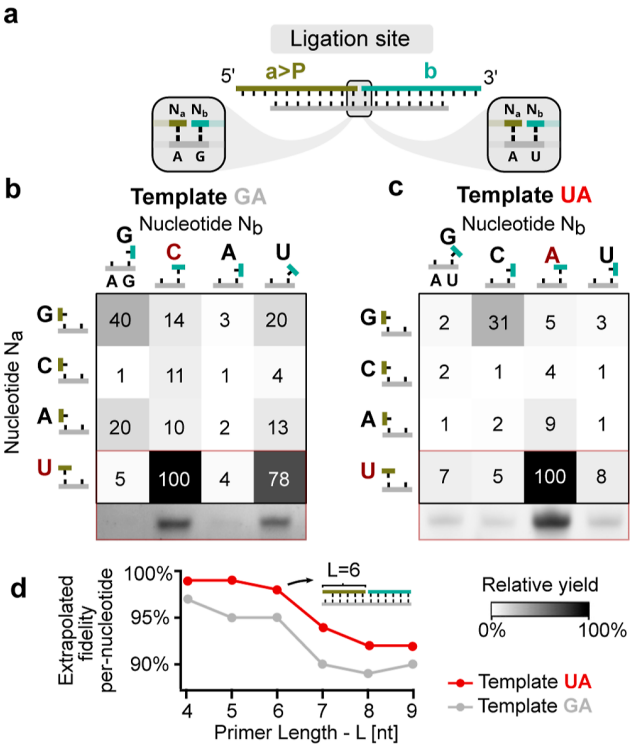


Figure 4. Ligation site specificity for two different template sequences. After 7 days, the yields were measured for reactions using primers that had each of the four possible nucleotides at the 3′ end of *a* (N_a) and the 5′ end of *b* (N_b). This resulted in a total of 16 different primer combinations being evaluated. (a) Schematics of the ligation sites and the two templates tested. The templates differed only at the dimer complementary to the ligation site with either 5′ GA 3′ (b) or 5′ UA 3′ (c). The maximum yield obtained in both cases was for the correct combination of complementary primers (nucleotides highlighted in red). Guanine/uracil (G/U) wobble pairing is represented as a tilted nucleotide in the cartoon representation. For most combinations, one single mutation at the ligation site reduced its relative yield considerably or prevented ligation, even though the other primer is fully bound to the template. The snippet below the heat map in (b,c) corresponds to the PAGE of the bottom row, showing the ligated product *ab*. Reactions were performed with 10 μM primers, 10 μM template, 50 mM CHES, pH 10, and 1 mM $MgCl_2$ for 7 days at 5 °C. Data are represented as mean of independent triplicates. (d) The fidelity of ligation was extrapolated to a per-nucleotide replication fidelity using primers of varying length using single-mutation-sensitive binding calculation of the primers with NUPACK (Supporting Information S9.1). The fidelity decreased for longer primers. Also, a G at the ligation site leads to lower fidelity due to the G−U wobble pairs (Supporting Information S9.3).

A prebiotic replication through ligation would take place from a diverse pool of oligonucleotides consisting of different sequences of varied lengths. In such a scenario, the likelihood of unstable primer−template complexes is high. Two contributions of the nucleotides surrounding the ligation site can be identified as having a significant effect on the stability of the complex, the amount of mismatches, and the length of the binding region. To compare the performance of the ligation with a base-by-base replication, we calculated the per-nucleotide replication fidelity as detailed in Supporting

https://doi.org/10.1021/jacs.3c10813
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Information S9. This corresponds to the minimum fidelity that a base-by-base replicator would require, for each incorporation, to create the same number of errors within the ligated strand. We combined thermodynamic analysis, using NUPACK, for one or two mutations in the ligating strands outside of the ligation site with the experimental ligation errors obtained by mutating both nucleotides at the ligation site (Figure 4b,c). These extrapolated experimental−theoretical per-nucleotide fidelities reached 89−92% for the tested system and 95−98% for primer length of 6-mer. Unlike shorter strands where a single point mutation would destabilize the complex, longer primers are more tolerant to single point mutations at cold temperatures, setting a fidelity-based length limit to the ligating strands. This is the reason for the increase in fidelity for shorter strands. Interestingly, the optimal lengths for ligation are reached by dry oligomerization from 2′,3′-cyclic nucleotides at the same alkaline pH.[15]

After understanding that the nonenzymatic ligation with 2′,3′-cyclic phosphates is a reliable copying mechanism, we aimed to investigate the potential for elongation. This would have been an important characteristic of a potential prebiotic replication mechanism as it would establish a link between nontemplated nucleotide condensation and the faster replication by long ribozymes, with tens or hundreds of nucleotides.[46−48]

We explored the possibility of bridging these two oligonucleotide length regimes by designing splint strands with short (4- or 8-mer) binding regions that can both cross-template and ligate (Figure 5a). Each system has two strands (labeled *c* and *d*) with >*P*. The strands are designed such that the 5′ half of the strand *d* is a reverse complement to the 5′ half of *c*, and the same goes for the 3′ halves such that they bind and form a long network of *ccc*... bound to *ddd*... (Figure 5a, see Supporting Information S3 for sequence information). The formed secondary structure allows for multiple ligations, resulting in homopolymers of *c* and *d*. Figure 5b shows that up to 5 concatenations ($n = 6$) could be detected for both the 16-mer system (Figure 5b) and 8-mer system (Figure 5d) resulting in 96- and 48-mer RNA, respectively. Figure 5c,e shows the gel quantification of the respective gels in Figure 5b,d. For the shorter system, the yield was reduced by about an order of magnitude for each additional concatenation and was generally lower than that for the 16-mer system, which was likely due to the slower kinetics in a frozen state at −20 °C. Specifically for the case of the 16-mer at 5 °C, 5.1 μM of the strands was incorporated into the concatemers, and 4.9 μM remained. The remaining strands *c* and *d* contain either the active or inactive phosphate group as the two species do not resolve through PAGE. We propose that for these conditions, as for the system in Figure 2, the main limitation to the yield is the hydrolysis of the cyclic phosphate.

It is interesting to note that the maximum yield for the 16-mer system was obtained at 5 °C, whereas for the 8-mer system, it was at −20 °C. We believe this to result from the low duplex stability of the 4-mer duplex region at 5 °C. We tested this by studying the ligation of shorter systems (Supporting Information S6). Systems with 4- and 6-mer binding regions were designed to have the same sequence at the ligation site and the same GC content as the strands used in Figures 1 and 2. While there was a significant reduction in yield for both systems, we found that for the 6-mer system, the yields were higher at 5 °C, whereas for the 4-mer system, −20 °C was more favorable (Supporting Information, Figure S6.1). Additionally, the obtained yields at 7 days were very low (about 3%) for the latter. This suggests that to stabilize the duplex for short oligonucleotides, a compromise between the slower rate of ligation and the low probability of duplex formation must be made. We demonstrate that the ligation reaction is robust since the only requirement for the templated ligation via 2′,3′-cyclic phosphate, besides the formation of a duplex, seems to be an alkaline pH, needing no additional molecules or external activations. Also, the shown formation of long RNA, bridges the length gap toward a regime where long, functional ribozymes can evolve.

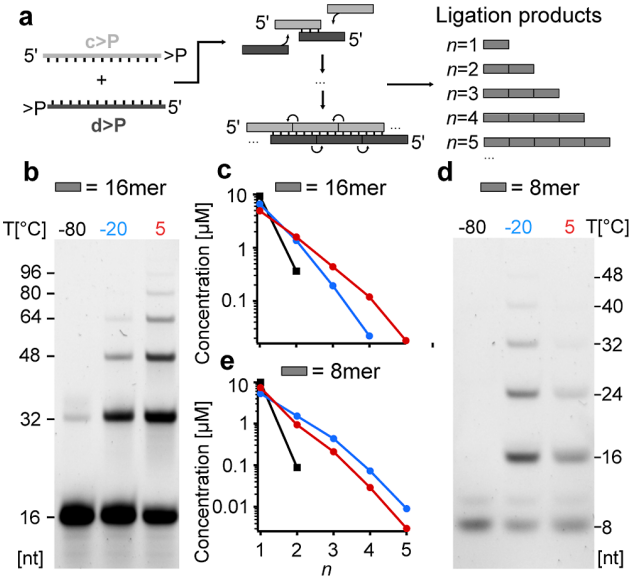


Figure 5. Assembly of long RNA via splinted ligation of 2′,3′-cyclic phosphate containing oligonucleotides. (a) Schematics of sequence design. Two strands labeled *c* and *d* with complementary subregions and not corresponding to the complete reverse complement, bind to form a long chain with repeating units of each. Both *c* and *d* contain 2′,3′-cyclic phosphate, and the ligation can yield all possible length multiples of the initial strands. For the case where *c* and *d* are 16-mer long, denaturing PAGE of ligation reaction at −80, −20 °C (frozen), and 5 °C (b) revealed products of up to five concatenations. The concentration of each ligation product (up to $n = 5$), obtained from the SYBR Gold fluorescence analysis, is plotted in (c). Similarly, for an initial strand size of 8-mer, denaturing PAGE (d) and product concentration (e) are shown. The optimal temperature for splinted ligation depends on oligomer length as 5 °C yields a higher concentration for the 16-mer, while −20 °C is better for 8-mer. Reactions were performed with 10 μM primers, 10 μM template, 50 mM CHES, pH 10, and 1 mM $MgCl_2$ for 7 days. Data are represented as the mean of three independent replicates.

CONCLUSIONS

A nonenzymatic replicator chemistry on the early Earth should have the capacity, under plausible conditions, to elongate strands and undergo further replication steps all while being highly accurate and processive. We demonstrate with this work that ligation with >P RNA fulfills these criteria.

First, we show that the template-directed replication mechanism only requires salts for stabilizing the duplex and alkaline pH, making the ligation with >P RNA robust and reproducible in both aqueous and frozen solutions. For the aqueous case, a 38% yield was observed in contrast to the previously reported yield of 16% for similar reactions. It was found that a combination of high pH and low temperature

https://doi.org/10.1021/jacs.3c10813
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

promotes ligation over the hydrolysis of the cyclic phosphate moiety. Such conditions are thought to be plausible on early Earth, where the fainter sun contributed to a cold surface temperature that would still allow liquid water.[49] Additionally, Hadean oceans were potentially alkaline due to the sequestering of CO_2 in carbonate minerals,[50] and alkaline conditions present in freshwater volcanic lakes[31,32] have been proposed to foster early metabolism.

However, a fraction of the >P still hydrolyzes, which contributes to its incomplete conversion. While the yield could be further improved by adding reagents that aid in the recyclization of the monophosphate moiety, such as diamidophosphate in combination with imidazole,[23] this would increase the complexity of the system. Moreover, this reaction has low salt requirements (1 mM $MgCl_2$, Supporting Information, Figure S4.3), ensuring RNA backbone integrity and compatibility with strand separation. It can also proceed in a wide pH range, even unbuffered (Supporting Information S4).

The elongation of short RNA was demonstrated with splinted systems that yielded up to six-copy concatemers of short RNA strands of either 16- or 8-mer, resulting in long RNA on the scale of 100-mer (Figure 5). This length range approaches the average length of replicating ribozymes,[46−48] representing a significant step toward assembling functional RNAs by plausible means. Even very short >P RNA fragments (with 4-mer base-pairing regions) ligate under frozen conditions (Supporting Information S6), establishing a bridge from the single nucleotide condensation reactions, yielding very short RNA strands, to a regime where templated ligation reactions could dominate.

Furthermore, we evaluated the copying accuracy with varying nucleotides at the ligation site. An experimentally measured fidelity of at least 82% was obtained upon screening all possible single base mutations at either the 3′ or 5′ end of the ligation site (Figure 4d), with the remaining primer being entirely complementary. To compare the ligation fidelity with a base-by-base replication, we extrapolated a per-nucleotide fidelity of 89−92% for adding eight nucleotides or a fidelity of 95−98% when adding six nucleotides.

Contrary to previous studies on >P, we found that under the tested conditions, the reaction was not regioselective, producing equal amounts of 2′−5′ and 3′−5′ linkages at the ligation site (Figure 3). While approximately half of the linkages were noncanonical, we argue this does not diminish the applicability of the reaction in a prebiotic context. Strands with 2′−5′ linkages have been shown to fold into functional structures,[40] and these noncanonical linkages have also been demonstrated to be more labile than 3′−5′ and have potential for interconversion.[39,51] Furthermore, Supporting Information, Figure S6.3 shows that product *ab* with a 2′−5′ linkage at the ligation site could still template the reverse ligation reaction. The noncanonical linkage in the template at the ligation site did not impede ligation, paving the way for exponential replication cycles.

Strand separation, driven by nonequilibrium environments with thermal, salt, or pH oscillations would allow for the implementation of a ligation chain reaction, similar to that reported by Edeleva et al., with the added benefit of not generating deleterious side-products by a prebiotically implausible EDC.[7] An air−water interface with continuous feeding would allow for both the denaturation and replenishment of activated primers, without necessarily high temperature which could degrade the cyclic phosphate and the RNA backbone.[6] This suggests that such scenarios could provide a niche, where the ligation reactions by 2′,3′-cyclic phosphate could evolve toward a ribozymatic replicator.

Considering these results, ligation with >P is an interesting framework to produce diverse pools of long RNA that could undergo molecular evolution. We show that the system described in the current study enables the generation of long RNA, with high fidelity. This was demonstrated for a range of lengths, sequence combinations, reaction conditions, and temperatures, suggesting that ligation of RNA with >P holds a central position in the general conception of the RNA world.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.3c10813.

Experimental procedures and quantification, nucleic acid sequences, buffer and salt conditions screening, control experiments, kinetic constant fitting, and per-nucleotide fidelity extrapolation parameters and expressions (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Dieter Braun − *Department of Physics, Center for Nanoscience, Ludwig-Maximilians-Universität München, 80799 Munich, Germany;* orcid.org/0000-0001-7751-1448; Email: dieter.braun@lmu.de

Authors

Adriana Calaça Serrão − *Department of Physics, Center for Nanoscience, Ludwig-Maximilians-Universität München, 80799 Munich, Germany;* orcid.org/0000-0003-0878-2185

Sreekar Wunnava − *Department of Physics, Center for Nanoscience, Ludwig-Maximilians-Universität München, 80799 Munich, Germany;* orcid.org/0000-0001-8793-2676

Avinash V. Dass − *Department of Physics, Center for Nanoscience, Ludwig-Maximilians-Universität München, 80799 Munich, Germany; Department of Physics and Astronomy, McMaster University, Hamilton, Ontario L8S4M1, Canada;* orcid.org/0000-0003-4787-375X

Lennard Ufer − *Department of Physics, Center for Nanoscience, Ludwig-Maximilians-Universität München, 80799 Munich, Germany*

Philipp Schwintek − *Department of Physics, Center for Nanoscience, Ludwig-Maximilians-Universität München, 80799 Munich, Germany;* orcid.org/0000-0002-6440-5918

Christof B. Mast − *Department of Physics, Center for Nanoscience, Ludwig-Maximilians-Universität München, 80799 Munich, Germany;* orcid.org/0000-0002-3338-6275

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.3c10813

Author Contributions

[§]A.C.S and S.W. contributed equally to this paper.

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/jacs.3c10813
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

■ ACKNOWLEDGMENTS

The authors acknowledge the support from the European Research Council (ERC Evotrap, grant number 787356), the CRC 235 Emergence of Life (project-ID 364653263), the CRC 392 Molecular Evolution in Prebiotic (project-ID 521256690) and the Excellence Cluster ORIGINS (under Germany's Excellence Strategy—EXC-2094−390783311) funded by the Deutsche Forschungsgemeinschaft (DFG), the Simons Foundation (327125), and the Center for Nano-Science (CeNS).

■ REFERENCES

(1) Gilbert, W. Origin of Life: The RNA World. *Nature* **1986**, *319* (6055), 618.
(2) Zhang, S. J.; Duzdevich, D.; Ding, D.; Szostak, J. W. Freeze-Thaw Cycles Enable a Prebiotically Plausible and Continuous Pathway from Nucleotide Activation to Nonenzymatic RNA Copying. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119* (17), No. e2116429119.
(3) Li, L.; Prywes, N.; Tam, C. P.; O'Flaherty, D. K.; Lelyveld, V. S.; Izgu, E. C.; Pal, A.; Szostak, J. W. Enhanced Nonenzymatic RNA Copying with 2-Aminoimidazole Activated Nucleotides. *J. Am. Chem. Soc.* **2017**, *139* (5), 1810−1813.
(4) Vogel, S. R.; Deck, C.; Richert, C. Accelerating Chemical Replication Steps of RNA Involving Activated Ribonucleotides and Downstream-Binding Elements. *Chem. Commun.* **2005**, *39*, 4922−4924.
(5) Sosson, M.; Pfeffer, D.; Richert, C. Enzyme-Free Ligation of Dimers and Trimers to RNA Primers. *Nucleic Acids Res.* **2019**, *47* (8), 3836−3845.
(6) Salditt, A.; Karr, L.; Salibi, E.; Le Vay, K.; Braun, D.; Mutschler, H. Ribozyme-Mediated RNA Synthesis and Replication in a Model Hadean Microenvironment. *Nat. Commun.* **2023**, *14* (1), 1495.
(7) Edeleva, E. V.; Salditt, A.; Stamp, J.; Schwintek, P.; Boekhoven, J.; Braun, D. Continuous Nonenzymatic Cross-Replication of DNA Strands with in Situ Activated DNA Oligonucleotides. *Chem. Sci.* **2019**, *10* (22), 5807−5814.
(8) Wachowius, F.; Holliger, P. Non-Enzymatic Assembly of a Minimized RNA Polymerase Ribozyme. *ChemSystemsChem* **2019**, *1*, 1−4.
(9) Zhou, L.; O'Flaherty, D. K.; Szostak, J. W. Template-Directed Copying of RNA by Non-enzymatic Ligation. *Angew. Chem.* **2020**, *132*, 15812−15817.
(10) Joyce, G. F. Non-enzymatic template-directed synthesis of RNA copolymers. *Orig. Life Evol. Biosph.* **1984**, *14* (1−4), 613−620.
(11) Prywes, N.; Blain, J. C.; Del Frate, F.; Szostak, J. W. Nonenzymatic Copying of RNA Templates Containing All Four Letters Is Catalyzed by Activated Oligonucleotides. *Elife* **2016**, *5*, No. e17756.
(12) Zhou, L.; O'Flaherty, D. K.; Szostak, J. W. Assembly of a Ribozyme Ligase from Short Oligomers by Nonenzymatic Ligation. *J. Am. Chem. Soc.* **2020**, *142* (37), 15961−15965.
(13) Verlander, M. S.; Lohrmann, R.; Orgel, L. E. Catalysts for the Self-Polymerization of Adenosine Cyclic 2′,3′-Phosphate. *J. Mol. Evol.* **1973**, *2* (4), 303−316.
(14) Verlander, M. S.; Orgel, L. E. Analysis of High Molecular Weight Material from the Polymerization of Adenosine Cyclic 2′, 3′-Phosphate. *J. Mol. Evol.* **1974**, *3* (2), 115−120.
(15) Dass, A. V.; Wunnava, S.; Langlais, J.; von der Esch, B.; Krusche, M.; Ufer, L.; Chrisam, N.; Dubini, R. C. A.; Gartner, F.; Angerpointner, S.; Dirscherl, C. F.; Rovó, P.; Mast, C. B.; Šponer, J. E.; Ochsenfeld, C.; Frey, E.; Braun, D. RNA Oligomerisation without Added Catalyst from 2′,3′-Cyclic Nucleotides by Drying at Air-Water Interfaces**. *ChemSystemsChem* **2023**, *5* (1), No. e202200026.
(16) Powner, M. W.; Gerland, B.; Sutherland, J. D. Synthesis of Activated Pyrimidine Ribonucleotides in Prebiotically Plausible Conditions. *Nature* **2009**, *459* (7244), 239−242.
(17) Lohrmann, R.; Orgel, L. E. Prebiotic Synthesis: Phosphorylation in Aqueous Solution. *Science* **1968**, *161* (3836), 64−66.
(18) Gibard, C.; Bhowmik, S.; Karki, M.; Kim, E. K.; Krishnamurthy, R. Phosphorylation, Oligomerization and Self-Assembly in Water under Potential Prebiotic Conditions. *Nat. Chem.* **2018**, *10* (2), 212−217.
(19) Lafontaine, D. A.; Beaudry, D.; Marquis, P.; Perreault, J.-P. Intra- and Intermolecular Nonenzymatic Ligations Occur within Transcripts Derived from the Peach Latent Mosaic Viroid. *Virology* **1995**, *212* (2), 705−709.
(20) Buzayan, J. M.; Gerlach, W. L.; Bruening, G. Non-Enzymatic Cleavage and Ligation of RNAs Complementary to a Plant Virus Satellite RNA. *Nature* **1986**, *323* (6086), 349−353.
(21) Li, Y.; Breaker, R. R. Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2′-Hydroxyl Group. *J. Am. Chem. Soc.* **1999**, *121* (23), 5364−5372.
(22) Breslow, R..; Labelle, M.. Sequential General Base-Acid Catalysis in the Hydrolysis of RNA by Imidazole. *J. Am. Chem. Soc.* **1986**, *108* (10), 2655−2659.
(23) Song, E. Y.; Jiménez, E. I.; Lin, H.; Le Vay, K.; Krishnamurthy, R.; Mutschler, H. Prebiotically Plausible RNA Activation Compatible with Ribozyme-Catalyzed Ligation. *Angew. Chem. Int. Ed.* **2021**, *60* (6), 2952−2957.
(24) Mutschler, H.; Holliger, P. Non-Canonical 3′-5′ Extension of RNA with Prebiotically Plausible Ribonucleoside 2′,3′-Cyclic Phosphates. *J. Am. Chem. Soc.* **2014**, *136* (14), 5193−5196.
(25) Mutschler, H.; Wochner, A.; Holliger, P. Freeze-Thaw Cycles as Drivers of Complex Ribozyme Assembly. *Nat. Chem.* **2015**, *7* (6), 502−508.
(26) Lutay, A. V.; Chernolovskaya, E. L.; Zenkova, M. A.; Vlassov, V. V. The Nonenzymatic Template-Directed Ligation of Oligonucleotides. *Biogeosci. Discuss.* **2006**, *3* (3), 243−249.
(27) Lutay, A. V.; Chernolovskaya, E. L.; Zenkova, M. A.; Vlasov, V. V. Nonenzymatic Template-Dependent Ligation of 2',3′-Cyclic Phosphate-Containing Oligonucleotides Catalyzed by Metal Ions. *Dokl. Biochem. Biophys.* **2005**, *401* (1−6), 163−166.
(28) Mutschler, H.; Taylor, A. I.; Porebski, B. T.; Lightowlers, A.; Houlihan, G.; Abramov, M.; Herdewijn, P.; Holliger, P. Random-Sequence Genetic Oligomer Pools Display an Innate Potential for Ligation and Recombination. *Elife* **2018**, *7*, No. e43022.
(29) Lutay, A. V.; Zenkova, M. A.; Vlassov, V. V. Nonenzymatic Recombination of RNA: Possible Mechanism for the Formation of Novel Sequences. *Chem. Biodivers.* **2007**, *4* (4), 762−767.
(30) Renz, M.; Lohrmann, R.; Orgel, L. E. Catalysts for the Polymerization of Adenosine Cyclic 2′,3′-Phosphate on a Poly (U) Template. *Biochim. Biophys. Acta, Nucleic Acids Protein Synth.* **1971**, *240* (4), 463−471.
(31) Toner, J. D.; Catling, D. C. A Carbonate-Rich Lake Solution to the Phosphate Problem of the Origin of Life. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117* (2), 883−888.
(32) Toner, J. D.; Catling, D. C. Alkaline Lake Settings for Concentrated Prebiotic Cyanide and the Origin of Life. *Geochim. Cosmochim. Acta* **2019**, *260*, 124−132.
(33) Lutay, A. V.; Grigoriev, I. V.; Zenkova, M. A.; Chernolovskaya, E. L.; Vlassov, V. V. Nonenzymatic Recombination of RNA by Means of Transesterification. *Russ. Chem. Bull.* **2007**, *56* (12), 2499−2505.
(34) Kinjo, M.; Rigler, R. Ultrasensitive Hybridization Analysis Using Fluorescence Correlation Spectroscopy. *Nucleic Acids Res.* **1995**, *23* (10), 1795−1799.
(35) Wetmur, J. G.; Davidson, N. Kinetics of Renaturation of DNA. *J. Mol. Biol.* **1968**, *31* (3), 349−370.
(36) Cisse, I. I.; Kim, H.; Ha, T. A Rule of Seven in Watson-Crick Base-Pairing of Mismatched Sequences. *Nat. Struct. Mol. Biol.* **2012**, *19* (6), 623−627.
(37) Usher, D. A.; Yee, D. Geometry of the Dry-State Oligomerization of 2′,3′-Cyclic Phosphates. *J. Mol. Evol.* **1979**, *13* (4), 287−293.
(38) Vlassov, A. V.; Johnston, B. H.; Landweber, L. F.; Kazakov, S. A. Ligation Activity of Fragmented Ribozymes in Frozen Solution:

https://doi.org/10.1021/jacs.3c10813
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Implications for the RNA World. *Nucleic Acids Res.* **2004**, *32* (9), 2966−2974.
(39) Rohatgi, R.; Bartel, D. P.; Szostak, J. W. Nonenzymatic, Template-Directed Ligation of Oligoribonucleotides Is Highly Regioselective for the Formation of 3′-5′ Phosphodiester Bonds. *J. Am. Chem. Soc.* **1996**, *118* (14), 3340−3344.
(40) Sheng, J.; Li, L.; Engelhart, A. E.; Gan, J. H.; Wang, J. W.; Szostak, J. W. Structural Insights into the Effects of 2 ’-5 ’ Linkages on the RNA Duplex. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111* (8), 3050−3055.
(41) Giannaris, P. A.; Damha, M. J. Oligoribonucleotides Containing 2′,5′-Phosphodiester Linkages Exhibit Binding Selectivity for 3′,5′-RNA over 3′,5′-ssDNA. *Nucleic Acids Res.* **1993**, *21* (20), 4742−4749.
(42) Wasner, M.; Arion, D.; Borkow, G.; Noronha, A.; Uddin, A. H.; Parniak, M. A.; Damha, M. J. Physicochemical and Biochemical Properties of 2‘,5‘-Linked RNA and 2‘,5‘-RNA:3‘,5‘-RNA “Hybrid” Duplexes. *Biochemistry* **1998**, *37* (20), 7478−7486.
(43) Lohman, G. J. S.; Bauer, R. J.; Nichols, N. M.; Mazzola, L.; Bybee, J.; Rivizzigno, D.; Cantin, E.; Evans, T. C. A High-Throughput Assay for the Comprehensive Profiling of DNA Ligase Fidelity. *Nucleic Acids Res.* **2016**, *44* (2), No. e14.
(44) Wu, D. Y.; Wallace, R. B. Specificity of the Nick-Closing Activity of Bacteriophage T4 DNA Ligase. *Gene* **1989**, *76* (2), 245−254.
(45) Luo, J. Improving the Fidelity of Thermus Thermophilus DNA Ligase. *Nucleic Acids Res.* **1996**, *24* (15), 3071−3078.
(46) Jaeger, L.; Wright, M. C.; Joyce, G. F. A Complex Ligase Ribozyme Evolved in Vitro from a Group I Ribozyme Domain. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (26), 14712−14717.
(47) Horning, D. P.; Joyce, G. F. Amplification of RNA by an RNA Polymerase Ribozyme. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (35), 9786−9791.
(48) Shechner, D. M.; Grant, R. A.; Bagby, S. C.; Koldobskaya, Y.; Piccirilli, J. A.; Bartel, D. P. Crystal Structure of the Catalytic Core of an RNA-Polymerase Ribozyme. *Science* **2009**, *326* (5957), 1271−1275.
(49) Rosing, M. T.; Bird, D. K.; Sleep, N. H.; Bjerrum, C. J. No Climate Paradox under the Faint Early Sun. *Nature* **2010**, *464* (7289), 744−747.
(50) Kadoya, S.; Krissansen-Totton, J.; Catling, D. C. Probable Cold and Alkaline Surface Environment of the Hadean Earth Caused by Impact Ejecta Weathering. *Geochem., Geophys., Geosyst.* **2020**, *21* (1), No. e2019GC008734.
(51) Jarvinen, P.; Oivanen, M.; Lonnberg, H. Interconversion and Phosphoester Hydrolysis of 2’,5′- and 3′,5′-Dinucleoside Monophosphates: Kinetics and Mechanisms. *J. Org. Chem.* **1991**, *56* (18), 5396−5401.

https://doi.org/10.1021/jacs.3c10813
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Acknowledgements

I would like to thank Dieter for giving me the opportunity to study and research such an interesting topic. Thank you for trusting me and giving me the freedom to explore new ideas. This is really the fun part about doing research.

I am grateful for all the collaborations and teaching opportunities I had while working in this lab. Science should be done as network effort and you made that possible. Firstly, thank you to Giacomo and Christoph Weber for all the interesting discussions about phase separation. It was always interesting to hear the 'theory' perspective. Thank you to all the bachelor, master and working students I supervised over the years: Max, Lennard, Zsófia, Ekaterina, Dorothea, Philipp, Felix, Lara and Éléonore. I learned a lot in the process. Thank you to Andres Jäschke and Christian for the talks about the origin of the genetic code. It continues to be one of the most interesting and simultaneously challenging questions on the origins of life.

Thank you to all the members of the Braun lab across the years. This lab has something special: the people. Never once I felt there was a competitive feeling between colleagues. You became my friends.

In particular, thank you to Sree, Juliette, Philipp and Martina for all the endless breaks and evenings. You were like a family during corona times. Looking forward to our next trip together. Sree, you are the most curious person I know. Staying in the lab late to look at the ligation project with you was easy because it never felt like work. Looking forward to see your barista career thrive. Juliette, you are the most kind person and never turned your back on anyone that needed help. I hope you find the courage to do exactly as you want after the PhD. Looking forward to visit you in a remote island with piles of hand-written notebooks. Martina, I miss our talks. I still believe they belong in a podcast...Thank you for bringing a burst of energy to our lab and the long afternoons cleaning the back lab, I always looked forward to them. Philipp, working here was worth it if only to have met you. You brought sunshine to the lab even in the greyest winter days and made me the happiest I have ever been. Looking forward to the many years to come.

Thank you to Christof, for the infinite knowledge, clarity and the bubble of influence you naturally have around you. Thank you, Alex F. for the funniest side comments. You always see the perfect punchline. Felix, working with you on the Bst project really felt like how teamwork should be. Ever since you are here the whole lab is laughing more. To Annalena, you were an example to me on how to 'PhD', you are so resourceful and determined. I actively took your example. To Max, for our Isar picnics and the infinite tech knowledge. To Saroj, the best cook I know. Thank you for patiently explaining me indian spices and amyloid structures. You have such a calming energy. Paula, for always being motivated for a party and the long sunny days in San Sebastian together. Chrissy for our late night talks in the Stone Village. Thomas for the calming talks about the end of the PhD and to Bobby for bringing the Gen Z slang to the lab and having the best dance moves.

Finally, thank you Bernhard for being one of the most knowledgeable and interesting people I know, while simultaneously so kind. From day one when we met in the Braun lab I knew we

would be friends. Now I know we will stay friends. To Stijn, for all the great moments across the years. You always give me a new perspective. Looking forward to summer with both of you. Thank you to Finn and Katrin for making Munich feel like home.

Obrigada aos meus amigos e família. Longe da vista mas perto do coração. Obrigada Mãe e Pai pelo suporte e motivação. Obrigada Mano por teres suscitado a minha curiosidade há muitos anos atrás. Obrigada Patrícia pelos momentos em família e pelas conversas que põem muito em prespectiva. Fico feliz de poder estar presente para ver o Daniel, o Lucas e a Paloma a crescer e genuinamente orgulhosa e curiosa pelos novos desenvolvimentos de cada vez que vou a Portugal. Obrigada Rita por seres o meu pilar e uma companheira para a vida.