Inaugural-Dissertation

zur

Erlangung der Doktorwürde

der

Gesamtfakultät

für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht – Karls – Universität

Heidelberg

vorgelegt von

M.Sc. Christian Schmitt

aus Heidelberg

Tag der mündlichen Prüfung: 18.06.2025

Investigating the Origin of the Genetic Code

Gutachter:

Prof. Dr. Andres Jäschke

Prof. Dr. Dieter Braun

Abstract

In extant life, tRNAs facilitate protein biosynthesis by decoding genetic information with their anticodons and by providing matching amino acids bound to their 3′ end. In turn, proteins are needed to attach amino acids to their cognate tRNAs through aminoacylation. Yet, in a prebiotic scenario, likely no proteins but only RNA existed. To explain the emergence of protein biosynthesis and the origin of the genetic code, John J. Hopfield proposed in 1978 a "testable Hypothesis" about a potential self-aminoacylating tRNA precursor. These "Hopfield folds" might have exhibited a folding pattern that positioned the anticodon loop close to the 3′ end. Thereby, anticodons could have also functioned as chemical sensors to facilitate binding of only their cognate amino acids. If such Hopfield folds had indeed been the precursors of extant tRNAs, it would emphasize a stereochemical origin of the genetic code and molecular vestiges of the sequences and structures of Hopfield folds might still be present in extant tRNAs.

To test this hypothesis, I successfully developed a multi-step procedure that involved the synthesis of various aminoacyl-5′-adenylates and L-lysyl-5′-cytidylate. These were used in aminoacylation experiments with an RNA library that resembled Hopfield folds with randomized anticodon loops to identify the most reactive and selective sequences through the use of Illumina library preparation and sequencing techniques. Furthermore, the sequencing data could be validated through additional individual characterization of the four most active and selective sequences identified.

I could show that the tested Hopfield folds exhibited various degrees of self-aminoacylation activity based on the sequence of their anticodon loop. Only a small fraction of the initial pool of sequences was reactive which indicated that aminoacylation had to be specifically catalyzed. In contrast to Hopfield's hypothesis, not only the single stranded anticodon loop but also the double stranded region close to the 3′ end and more distant parts in the 5′ terminal region were significantly involved in the realization of self-aminoacylation. However, none of the most active and selective sequences found shared notable resemblance to extant anticodons. Furthermore, the adenylate or cytidylate group had the most significant effect on the selection of reactive sequences. Amino acid side chains seemed to cause steric hindrance that correlated with their mass but was mitigated by moieties that could facilitate attractive interaction with RNA like amino and hydroxy groups. The influence of amino acid L- and D-conformation was various and sequence-dependent. An optimum at pH 5 and almost no self-aminoacylation activity at neutral pH could be observed and lower concentrations of activated amino acids increased aminoacylation specificity.

Kurzzusammenfassung

In heutigem Leben ermöglichen tRNAs die Proteinbiosynthese, indem sie mit ihren Anticodons genetische Informationen dekodieren und an ihrem 3'-Ende gebundene passende Aminosäuren bereitstellen. Im Gegenzug werden Proteine benötigt, um Aminosäuren durch Aminoacylierung an ihre entsprechenden tRNAs zu binden. In einem präbiotischen Szenario gab es jedoch wahrscheinlich keine Proteine, sondern nur RNA. Um die Entstehung der Proteinbiosynthese und den Ursprung des genetischen Codes zu erklären, schlug John J. Hopfield 1978 eine „überprüfbare Hypothese " über einen möglichen selbstaminoacylierenden tRNA-Vorläufer vor. Diese „Hopfield folds" könnten ein Faltmuster aufgewiesen haben, das die Anticodon-Schleife nahe dem 3'-Ende positionierte, sodass diese als chemischer Sensor fungieren und so nur die Bindung passendender Aminosäuren zulassen konnte. Dies würde für einen stereochemischen Ursprung des genetischen Codes sprechen, und molekulare Überreste der Sequenzen und Strukturen der Hopfield folds könnten in heutigen tRNAs noch vorhanden sein.

Um diese Hypothese zu testen, habe ich erfolgreich ein mehrstufiges Verfahren entwickelt, das die Synthese verschiedener Aminoacyl-5'-adenylate und L-Lysyl-5'-cytidylat beinhaltete. Diese wurden in Aminoacylierungsexperimenten mit einer RNA library verwendet, die Hopfield folds mit randomisierten Anticodon-Schleifen entsprach, um die reaktivsten und selektivsten Sequenzen durch den Einsatz von Illumina library preparation und Sequenzierungstechniken zu identifizieren. Darüber hinaus konnten die Sequenzierungsdaten durch zusätzliche individuelle Charakterisierung der vier aktivsten und selektivsten Sequenzen validiert werden.

Ich konnte zeigen, dass die getesteten Hopfield folds abhängig von der Sequenz ihrer Anticodon-Schleife unterschiedlich hohe Selbstaminoacylierungsaktivitäten aufwiesen. Nur ein kleiner Teil des anfänglichen Pools von Sequenzen war reaktiv, was darauf hindeutete, dass die Aminoacylierung spezifisch katalysiert werden musste. Im Gegensatz zu Hopfields Hypothese waren nicht nur die einzelsträngige Anticodon-Schleife, sondern auch die doppelsträngige Region in der Nähe des 3'-Endes und weiter entfernte Teile in der 5'-terminalen Region maßgeblich an der Realisierung der Selbstaminoacylierung beteiligt. Keine der aktivsten und selektivsten Sequenzen, die gefunden wurden, wies jedoch eine auffällige Ähnlichkeit mit existierenden Anticodons auf. Darüber hinaus hatte die Adenylat- oder Cytidylatgruppe den größten Einfluss auf die Auslese reaktiver Sequenzen. Die Seitenketten der Aminosäuren schienen abhängig von ihrer Masse sterisch zu hemmen, was durch chemische Gruppen wie Amino- und Hydroxygruppen, die eine attraktive Interaktionen mit RNA erlauben, abgeschwächt wurde. Der Einfluss der L- und D-Konformation der Aminosäuren war sequenzabhängig unterschiedlich ausgeprägt. Es konnte ein Optimum bei pH 5 und fast keine Selbstaminoacylierungsaktivität bei neutralem pH beobachtet werden. Niedrigere Konzentrationen von aktivierten Aminosäuren erhöhten die Spezifität der Aminoacylierung.

Content

1 Introduction

1.1 Proteins

1.1.1 Structure and functions

In all extant life, proteins are made of organic molecules of at least 2 carbon atoms composed of a carboxylic acid group formed by their C1-atom and an amino group linked to their C2-atom. As the carboxylic acid and amino moiety are connected to the same carbon atom, they are referred to as *alpha amino acids* (Figure 1 A).

Extant life utilizes a total of 22 different additional residues or side chains that are also linked to the C2-atom even though a plethora of isomers or similar compounds would be chemically feasible. These residues differ in their chemical properties like hydrophobicity, acidity, aromaticity, etc. resulting in a range of total molecular weight of 75 (glycin) to 204 (tryptophan) or 255 (pyrrolysine) g/mol. If a residue causes the C2 atom to be stereogenic, solely the L-enantiomer is biologically relevant. As only these 22 L-alpha-amino acids are being used for the production of proteins, they are designated as *proteinogenic*. In the following, the term *amino acid* standardly refers to these proteinogenic L-alpha-amino acids.

Amino acids are being connected to each other through the formation of peptide bonds between the carboxy-group of one amino acid and the amino-group of another one. Through this mechanism, stable linear polypeptide chains with specific sequences can be created, harboring up to tens of thousands of amino acids [1]. This sequence is called *primary* structure and possesses a free amino group at one end which forms its *N-terminus*, and a free carboxylic acid group at the other end which forms its *C-terminus* (Figure 1 B). Depending on their primary structure, these polypeptide chains often form typical helical, planar, or other structures or are completely unstructured; these typical structural elements make up the *secondary structure*. The entirety of three-dimensional structures that a certain polypeptide chain exhibits is called *tertiary structure*. If two or more polypeptide chains interact with each other, they may form *quaternary structures*. All biologically relevant polypeptides are called *proteins* which realize the majority of metabolic functions of a cell. They can e.g. serve as scaffolds for larger complexes, form pores or motors. Proteins that harbor a catalytically *active center* are called *enzymes* and may need to interact covalently or non-covalently with *cofactors* like small molecules or metal ions to exhibit their function. Proteins can also be modified to tune their chemical properties like glycosylation to increase their solubility. Extant

life depends to the highest degree on the manifold capabilities of proteins; without them, life as we know it would be unfeasible.

The relatively small size and chemical complexity of amino acids are advantageous for the formation of complex three-dimensional structures, but polypeptide chains cannot serve as templates for their own replication. All biological proteins are therefore produced by a process called *translation*.

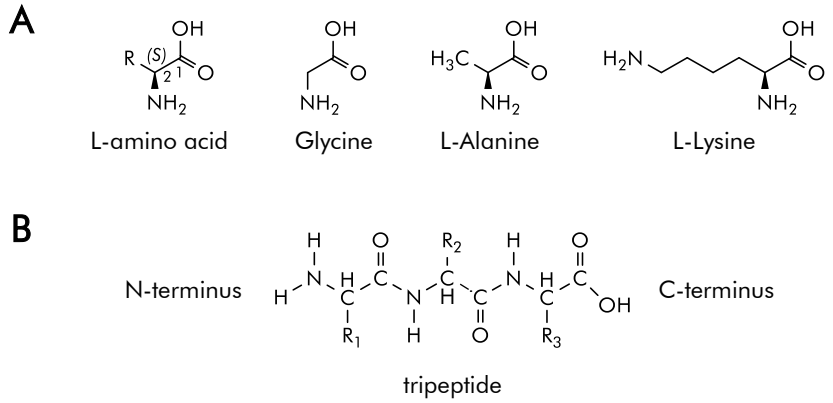


Figure 1. **(A)** General structure of biologically relevant amino acids alongside the three most relevant amino acids for this thesis. **(B)** The chemical structure of the displayed tripeptide is exemplary for all polypeptides as up to several thousands of amino acids can be linked like this.

1.2 DNA and RNA

1.2.1 Structure

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) share many similarities. Both consist of 4 different monomeric units that form polymers. Each monomer consists of one out of four nitrogenous nucleobases namely adenine (A), cytosine (C), guanine (G), and thymine (T) for DNA or uracil (U) for RNA. These nucleobases are coupled to the 5-carbon sugar *ribose* at its C1 atom via a beta-N-glycosidic bond. In the case of RNA, regular ribose is used whereas DNA contains deoxyribose which lacks the hydroxy group at its C2 atom (Figure 2). The C5 atom of the ribose can be linked to a chain of 0 to 3 phosphoryl groups. Nucleobase and ribose without any phosphoryl groups are called *nucleosides*, nucleoside monophosphates are called *nucleotides* and constitute the monomeric unit of DNA and RNA. If the ribose is to be specified, the terms *deoxynucleoside* or *ribonucleoside* may be used; the same applies for *deoxynucleotide* and *ribonucleotide*. In any case, nucleoside triphosphates constitute the chemically activated form of nucleosides and are used for the biological synthesis of novel genetic material. Especially guanosine triphosphate and adenosine triphosphate are furthermore used in manifold metabolic processes to form reactive intermediates.

To form polymers (strands), the hydroxy group at the C3 atom of a nucleotide's ribose can be linked to a phosphoryl group of another nucleotide forming an ester bond. These connections can be repeated to form chains of up to billions of nucleotides (nt). One end of the chain harbors a free hydroxy group linked to the C3 atom of the ribose; this end is therefore called *3' end* or *3' terminus*. The other end contains either a free or phosphorylated hydroxy group bound to the C5 atom of the ribose; this end is therefore referred to as *5' end* or *5' terminus*. In biology, both DNA and RNA strands are synthesized in 5' to 3' direction and analogously, sequences are regularly displayed in 5' to 3' direction, too.

DNA is typically organized as a double stranded (ds) helix with both strands being antiparallel so that each end of the helix shows both a 3' and a 5' end (Figure 2 A). Eukaryotic genomes are organized in this linear form whereas bacterial genomes are circular. Arranged like this, the nucleobases can form base pairs (bp). Each nucleobase has a specific set of hydrogen donor and acceptor sites that can form hydrogen bonds. Under physiological conditions, A and T can form two hydrogen bonds and C and G can form three (Figure 2 A). A and T as well as C and G are therefore *complementary* to each other and generally referred to as Watson-Crick base pairs as they bind stronger than every other permutation of A, C, G, and T.

The formation of base pairs or the formation of a double strand from single strands is called *hybridization* or *annealing*; their separation is called *denaturation* or *melting*.

In RNA, instead of thymine, uracil is being used which merely lacks a methyl group but otherwise behaves similar to thymine (Figure 2 B). RNA can form the same helical structures like DNA and both DNA:DNA, DNA:RNA, and RNA:RNA double strands are possible. In most biological applications though, RNA is present as a single strand (ss) that forms complex structures through intramolecular hybridization based on its sequence. Thereby, not only the classical Watson-Crick base pairs can be observed but also less stable interactions like the non-Watson-Crick base pairs G-U or A-A. Furthermore, other non-canonical interactions may occur in various geometries that may also include a nucleobase's Hoogsteen or sugar edge [2]. Due to its internal 2′ hydroxy groups, RNA is also more prone to hydrolysis, especially under alkaline conditions compared to DNA.

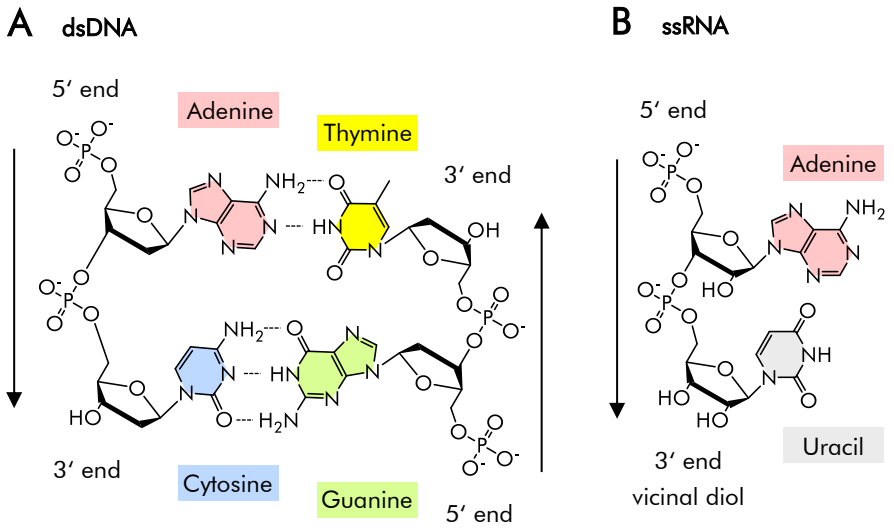


Figure 2. (A) Molecular structure of double stranded DNA in the shape of two hybridized dinucleotides. All four canonical nucleobases forming the two most stable pairs are displayed; dotted lines indicate hydrogen bonds. Arrows indicate regular 5′ to 3′ direction as can be typically observed in this double stranded helical structure. Longer polynucleotides behave similar. **(B)** Molecular structure of a single stranded RNA dinucleotide. RNA uses Uracil instead of Thymine as in the case of DNA; Uracil merely lacks a methyl group but is otherwise identical to thymine. The other difference is the use of ribose instead of deoxyribose which grants RNA a vicinal diol at its 3′ end as well as internal 2′ hydroxy groups.

1.2.2 Replication

When comparing the structure of proteins with that of DNA and RNA, one major difference is the double helical structure and the base pairing mechanism of DNA and RNA as polypeptide chains exhibit neither of both. This enables DNA and RNA sequences to be copied relatively easily. In general, DNA replication is realized in 3 steps: First, double stranded DNA has to be denatured so that in the second step, a small RNA or DNA oligo (primer) can be annealed to a specific site on the now single stranded DNA molecule that is supposed to serve as template. In the third step, the primer is being elongated by a DNA polymerase which moves over the template in 3′ to 5′ direction whilst elongating the primer in 5′ to 3′ direction with the nucleotides complementary to the template strand. Biological organisms conduct these steps at their body temperature and through the aid of additional proteins like primases and helicases. Polymerase chain reaction (PCR) – a widespread analogous laboratory technique – facilitates a thermostable DNA polymerase and uses a sequence of specific elevated temperatures between ~50 to 98 °C to achieve denaturation, annealing, and elongation within one thermal cycle. These steps can be repeated as often as desired, leading to an exponential replication of the initial material as per cycle the DNA present in the sample is roughly doubled. Also, primers can be used that bind properly with their 3′ region to the template but possess an overhang at their 5′ region leading to the addition of specific sequences to the resulting product.

Some viruses use reverse transcriptases to use RNA as template for the production of DNA. Analogous *in vitro* techniques are available which employ enzymes derived from the Moloney Murine Leukemia Virus (M-MLV) that depend on a DNA primer and can use both RNA and DNA as template. These enzymes are not thermally stable and the production of DNA is being conducted in one set of thermal denaturation, annealing of the primer and its extension. The resulting product is usually called *copy DNA* (cDNA) and can be extended through the use of 5′ overhang primers the same way as in the case of PCR.

Other viruses employ RNA-dependent RNA polymerases that synthesize RNA from an RNA template. Besides that, RNA is mostly produced from DNA templates through a process called *transcription*. Such RNA polymerases are usually able to denature double stranded DNA templates on their own sufficiently to reveal enough single stranded DNA to serve as template and transcription start site. Both *in vivo* and *in vitro* reactions therefore function and room temperature or around 37 °C. Other than DNA polymerases, RNA polymerases do not rely on primers, they merely initiate the synthesis of a novel RNA strand at specifically recognized sequences called *promoters*.

Another key process of maintaining and handling genetic information both *in vivo* and *in vitro* is the mechanism of ligation. Manifold ligases exist which catalyze the conjunction of the 3′ end of one strand with the phosphorylated 5′ end of another strand or the same strand with the latter resulting in a circular product. For in vitro purposes, ligases for the ligation of all kinds of combinations were made commercially available; including ligations of DNA:DNA, RNA:RNA, DNA:RNA, as well as RNA:DNA. Ligases usually rely on the use of ATP to form a chemically activated intermediate through adenylation. In this case, adenosine monophosphate is being attached via its phosphate group to the phosphorylated 5′ end of the strand that is about to be ligated with another one's 3′ end. This results in the formation of an anhydride bond between the two phosphoryl moieties releasing pyrophosphate in the process. Anhydride bonds are generally reactive so in this case, their chemical energy is used to form the ester bond – which contains less chemical energy – between the hydroxy group of the 3′ end and the phosphoryl group of the 5′ end of the strands to be ligated.

As a completely abiotic approach, DNA and RNA can also be chemically synthesized through the use of phosphoramidite building blocks in solid-phase synthesis. In this regard, most relevant for this thesis is the limitation of the maximum length of oligonucleotides produced this way. At the time of this thesis, the maximum length of RNA that could be obtained was 60 nt and that of DNA was 200 nt. Other than that, this chemical synthesis progresses in 3′ to 5′ direction which causes highest fidelity in the 3′ region and an increased error rate in the 5′ region depending on the length of the oligonucleotide. Lastly, the synthesis is being conducted in a step- or cycle-wise manner so that one nucleotide at a time is being added. During a step, a mixture of activated nucleotides can be used to achieve the incorporation of different nucleotides at one position leading to a mixture of sequences in the final product. This process is usually biased though, leading to distributions other than 1:1:1:1 for each nucleotide.

1.2.3 Functions

DNA serves as the main medium for the storage of information in a cell and some viruses in the form of a genome whose main function is to regulate the production of proteins. Depending on the organism, more or less regulatory and other sequences are present in the genome that are never transcribed. Usually, only a small fraction of an organism's genome actually encodes blueprints for the production of proteins. Such areas give rise to coding RNA or messenger RNA (mRNA). Other areas of the genome give rise to non-coding RNAs that either serve regulatory purposes or are directly involved in the production of proteins [3]. As

mentioned before, this process is called translation as a sequence of nucleotides is being translated into a sequence of amino acids. Most RNA present in extant cells serves the realization of this process. 80 to 90% of RNA mass is made up of ribosomal RNA (rRNA) [4] which together with ribosomal proteins form the large and small subunits of ribosomes which are the central piece of translation.

The second most abundant RNA by mass with 10 to 15% is transfer RNA (tRNA) [4]. These are single stranded RNAs of about 76 to 90 nt in length [5] whose two-dimensionally projected secondary structure resembles the shape of a cloverleaf (Figure 15 A) whereas its tertiary structure is similar to the capital letter *L*. With their 3′ hydroxy moiety, they can form a reactive ester bond with the carboxylic acid moiety of an amino acid. At this state, the 3′ end is aminoacylated and the tRNA is referred to as *charged*. Positions 33 to 37 make up the anticodon loop with position 34, 35, 36 resembling the actual anticodon. The remaining structures like the TΨC loop or D loop realize structures that are both compatible to properly enter the ribosome and to be recognized by enzymes called aminoacyl-tRNA-synthetases [6,7]. tRNAs also typically carry non-canonical bases like e.g. inosine in some anticodons which facilitates wobble base pairing with C, U, and A to ultimately let one anticodon decode several codons [8] (Figure 4). Other modified nucleotides facilitate structural changes or interaction with other structures [9].

The entirety of the translation process is accompanied by the activity of several proteins called translation factors; they aid in the initiation, elongation, and termination of the process [10-12]. For simplicity reasons, their structure and functions will not be further elucidated but only the main components will be displayed. The first step of translation is the transcription of a protein coding gene into mRNA. This mRNA is then bound by both the small and large ribosomal subunit. The ribosome moves in 5′ to 3′ direction until it reaches the first AUG codon. This defines the reading frame as from this position onwards, tRNAs with the matching anticodon can enter the ribosome to deliver their bound amino acid. As AUG refers to methionine, it is always the first amino acid which protein biosynthesis starts with. The ribosome has three sites, each one covering one codon. They facilitate tRNAs to enter the ribosome (aminoacyl site), to contribute their bound amino acid to the ongoing protein biosynthesis (peptidyl site), and to leave the ribosome again uncharged (exit site). Thereby, protein biosynthesis proceeds in tandem with the movement of the ribosome along the mRNA [13]. This process is terminated when the ribosome reaches a stop codon which does not encode an amino acid but causes the release of the newly synthesized protein. Which codon matches which amino acid is called the genetic code and is nearly identical in all organisms [14,15] (Figure 4).

One group of enzymes produced like this are aminoacyl-tRNA-synthetases (aaRS). A cell contains at least as many different aaRS as it uses amino acids for protein

biosynthesis. These aaRS recharge the corresponding tRNAs with their according amino acid once their bound one was used up during translation [6]. tRNAs are recognized by their corresponding aaRS by several structural features which only in some cases include the anticodon loop. The mechanism that assigns which tRNA can bind to which aaRS is sometimes called the *second genetic code* but this term is not consistently established [16].

aaRSs bind free amino acids and ATP from the cytosol to synthesize an intermediate aminoacyl-5'-adenylate that is kept bound onto the enzyme to protect these labile compounds from hydrolysis. In a second step, these adenylates are used to aminoacylate the 3' end of a depleted tRNA releasing AMP in the process. At this point, the recharged tRNA is released from the aaRS and can serve its purpose in translation again [7].

There are two classes of aaRS; class I enzymes initially conduct aminoacylation at the terminal 2' OH of their cognate tRNAs followed by a subsequent transesterification reaction that results in a 3' aminoacylated tRNA whereas class II enzymes directly aminoacylate the 3' end of their cognate tRNA [7]. Due to the amino acids they are responsible for, class II aaRS appear to have arisen earlier in evolutionary terms [17,18].

1.2.4 Illumina high throughput sequencing

Sequencing describes the act of determining which nucleotides in which order a given DNA or RNA molecule consists of. The first widespread technique was established in the 1970s and was named by one of its inventors as *Sanger sequencing*. Through this technique, up to 1000 consecutive nucleotides of a DNA molecule could be determined per *run* which included chain termination PCR followed by electrophoresis [19].

Even though read length, fidelity, and cost per run render Sanger sequencing competitive until today if only a very limited amount of different molecules is to be sequenced, other methods were developed that allowed for parallel sequencing of millions of molecules in parallel. Hence, these novel approaches were termed *high throughput sequencing* with one of the most successful ones being developed by the company *Illumina*.

Illumina provides both the machines (sequencers) and the necessary reagents to conduct fully automated sequencing runs. The sequencers differ from each other in technical details but the underlying working principles are nearly identical. All sequencers can be viewed as a combination of a microfluidics device and a

fluorescence microscope with the former handling the addition of buffers, substrates, enzymes, etc. to realize each desired reaction and the latter being relevant during the actual sequencing process. The core part of each sequencer is its so called *flow cell*, a structure of two flat parallelly aligned glass surfaces that enclose a volume. The distance between these glass plates is typically below 1 mm and its area varies between 0.5 cm^2 and several cm^2 depending on the sequencer. The inside of these glass plates is coated with two species of covalently bound oligonucleotides of known sequence. These oligonucleotides are attached with their 5′-end onto the glass plates leaving their 3′-ends free [20].

The genetic material that is to be sequenced is typically composed of a mixture of millions to billions of different DNA molecules in one solution; such a mixture of genetic material is called a *library*. All of these molecules though must possess sequences that are complementary to the attached oligonucleotides in order to facilitate hybridization. Once hybridized, the oligos act as primers and a DNA polymerase can elongate the bound oligos with the complementary sequence of the DNA that is to be sequenced; this first replication event binds the sequences of interest covalently to the flow cell. The original DNA molecules are washed away at this point. The elongated single stranded oligonucleotides can now bend over to hybridize with their 3′-ends to a bound oligonucleotide of the other kind which can then serve as a primer for the synthesis of the complementary strand. Upon denaturation, both the forward and the reverse sequence is present on the flow cell. This process of bending over, DNA synthesis, and denaturation can be repeated several times to give rise to locally confined clusters of DNA strands of identical sequence and is thereby called *bridge amplification*. Initial developments used unstructured flow cells on which size, shape, and number of clusters could only indirectly be influenced mostly by the concentration of the initial material given to the flow cell. At the time of this thesis, patterned flow cells started to be introduced which possessed wells of defined size, shape, and number which made this *cluster generation* more robust and controllable.

Following cluster generation, the actual sequencing process starts by cleaving one kind of the bound oligonucleotides near their initial 3′ end. This causes the release of one kind of strand from each cluster while its reverse complementary strand remains intact. The loose strands are washed away. Next, a defined set of primers is given to the clusters to hybridize to specific complementary sequences so that ahead of their 3′ end lies the sequence of interest. At this point, the sequencing by synthesis process starts by filling the flow cell with suitable buffer and DNA polymerase to incorporate a single fluorescently labeled nucleotide. These nucleotides are reversibly modified at their 3′ end to enable the incorporation of only one nucleotide at a time. Later on, this modification as well as the fluorophore can be cleaved off to enable elongation again; this approach is therefore known as *cyclic reversible termination* (Figure 3) [21].

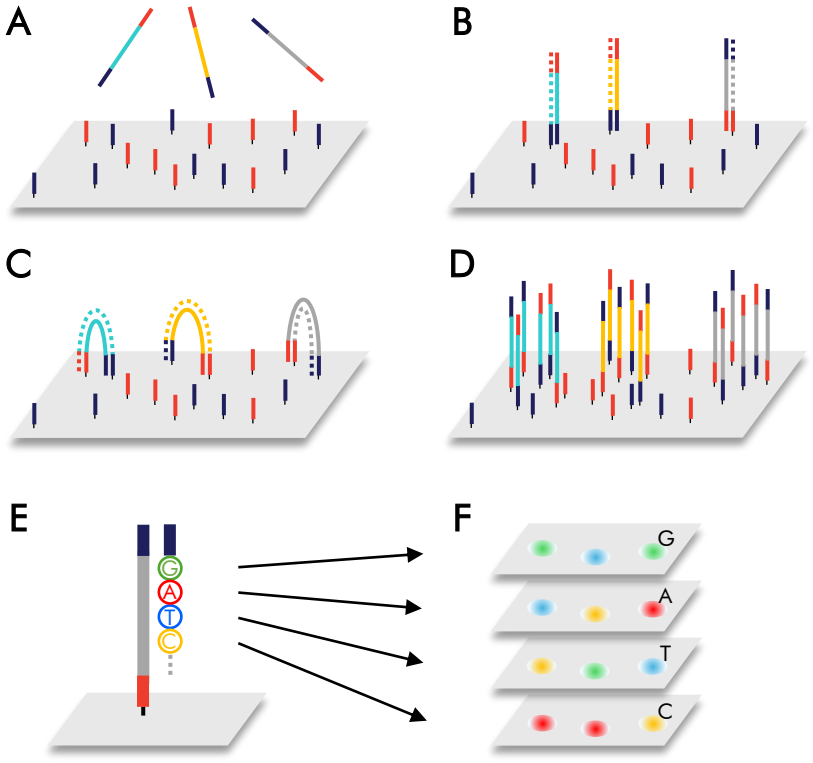


Figure 3. Schematic view of the Illumina high throughput sequencing process on an unpatterned flow cell (grey plate). **(A)** Denatured, ssDNA with sequences complementary to the bound oligonucleotides is introduced into the flow cell. **(B)** After annealing, the bound oligonucleotides serve as primers for DNA synthesis. **(C)** The original strands were washed away and the bound newly synthesized strands can bow and hybridize with their 3′ end with another bound oligonucleotide which can again serve as primer. **(D)** Upon several repetitions of denaturing, bowing, and DNA synthesis, local clusters of identical sequence can be produced. This is therefore referred to as bridge amplification. **(E)** During sequencing, other primers are introduced and DNA polymerase is conducted through cyclic reversible termination. Here, 4-channel chemistry is displayed. **(F)** During each cycle, the flow cell undergoes fluorescent microscopy to scan the fluorescent behavior of each cluster thereby determining the last incorporated base.

Illumina deployed three different approaches to the fluorescent readout called 4-, 2-, or 1-channel chemistry with the 4-channel chemistry being the oldest one. In this case, each of the four nucleotides is labeled with another fluorophore which are detected using four different excitation and emission wavelengths. Through this fluorescent signal, the nucleotide that was incorporated last can be determined in

each cluster in each cycle; this process is called *base calling*. This necessitates the use of a complex fluorescent microscope setup to scan the entire flow cell for all four fluorophores. In an attempt to simplify this setup, the number of fluorophores was reduced, and intermediate steps were introduced that changed the fluorescent behavior of the lastly incorporated nucleotide by cleaving off or adding a fluorophore onto certain nucleotides. This gave rise to the 2- and 1-channel chemistry that used only 2 or 1 fluorophore, respectively.

In an extreme case, the entire sample that is to be sequenced consists of the same sequence. In this case, the sample is of the lowest complexity and all clusters would show the same color in each cycle. If on an unpatterned flow cell two clusters were directly adjacent to each other, it would be impossible to identify each individual one. It is therefore beneficial to employ samples of higher complexity so that over the course of especially the first few cycles, all clusters show different fluorescent behavior than their neighboring ones. In the case of patterned flow cells, the use of complex samples enables the exclusion of clusters generated from two or more different template molecules leading to higher quality data. As the first cycles also serve to calibrate the optical system, it may lead to errors, or the run being aborted if the initial sample complexity is too low as the machine may interpret the readout as erroneous.

The entire sequencing process is composed of up to four sections with the first one being the sequencing of the sample in one direction (read 1 or R1). As single sequencing runs are oftentimes used to sequence multiple samples at once, indices are used to identify which sample each cluster was generated from. Two different indices of various lengths can be used with the first one being sequenced right after the product of read 1 was washed away. Therefore, another primer is being introduced that hybridizes with a defined sequence closer to the 5′ end of the bound DNA strand with the index ahead of its 3′ end. Sequencing is being conducted analogously to read 1. Subsequently, the product of this first index read (index read 1) is being washed away and the previously cleaved off complementary strands of the bound DNA are being regenerated through bridge amplification. Depending on the sequencer, during or after this bridge amplification the other index is being sequenced using the bound oligo as a primer or employing a dissolved primer, respectively (index read 2). Lastly, the region of interest can be sequenced from the other end using another primer analogous to read 1.

If the region of interest is sequenced in just one read, it is called *single-end sequencing*; using both reads is called *paired-end sequencing*. Paired-end sequencing results in longer effective read length if the two reads have little overhang or a higher confidence in the base calls in the case of extended overhang. The use of indices is optional but if included, higher complexity is advantageous in their case, too.

After the sequencing is finished, general metrics allow for a quick assessment of how successful the sequencing run was. *Occupation* summarizes how densely the flow cell was populated with clusters. *Reads passing filter* describes the fraction of reads with acceptable quality. A general trend and a main factor for optimization is the concentration of the input sequencing library. If it is too low, then reads passing filter may be high but as the occupation tends to be low in this case, not much data is generated. A too highly concentrated library tends to result in higher occupation but as the flow cell becomes more crowded, clusters increasingly merge with one another which gives ambiguous results reducing the fraction of reads passing filter which in turn also lowers the generation of useful data.

1.2.5 Library preparation

The process of preparing the desired genetic material for Illumina sequencing is called *library preparation*. As input material, usually DNA or RNA extracted from biological samples or synthetic libraries are being used. Depending on the input material, different steps like reverse transcription or overhang PCR are employed to generate the final dsDNA library that can be sequenced. As the read length of most Illumina sequencers is limited to 150 bp, the sequence of interest usually does not exceed around 275 bp if paired-end sequencing is desired. The sequence of interest must also be flanked at least on one side with a sequencing primer binding site and on both sides with complementary sequences to the bound oligonucleotides within the flow cell. A second sequencing primer binding site as well as the use one to two indices is optional (Figure 16).

In case only a fraction of a sequencing run's capacity is required for a single sample, several samples can be prepared using different indices and mixed prior to sequencing; an approach known as *multiplexing*. The resulting sequencing data can afterwards be demultiplexed using the index reads to assign the reads of each cluster to the sample they originated from.

1.3 Emergence of life

1.3.1 Defining life

In contrast to non-living matter, life exhibits typical phenomena like using cells as the smallest, self-regulated unit of organization. Although chemically diverse amongst species, all cells share a membrane that separates the exterior environment from their inner cytoplasm. Across this cell membrane, multiple gradients like different concentrations of ions are actively maintained to propel other functions like motility [22] and the production of ATP [23]. Such disequilibria and other physico-chemical mechanisms are constantly and actively maintained by a cell's metabolism. It consumes energy from the environment and releases entropy into it [24]. As the breakdown of the metabolic network usually results in the irreversible death of a cell, the only forms of life without metabolism are dormant forms like spores [25] and viruses [26] but these also rely on the mechanisms of living cells to fulfill their life cycle and to maintain their population [27,28]. Another characteristic of life is its ability to reproduce itself. Cells therefore duplicate themselves into copies with mostly the same abilities. The duplication of the genome is seldomly perfect; mutations in the genotype may lead to alterations of the phenotype. In combination with natural selection for the fittest organisms, evolution occurs [29]. Life can therefore be defined from abiotic phenomena by its self-regulated cellular organization which exhibits semi-permeable membranes, an ongoing metabolism that keeps the system in constant chemical disequilibrium, as well as the ability of reproduction and Darwinian evolution [24,30,31].

1.3.2 Timeline

Around 4.5 billion years (Ga) ago, the solar system formed, including the Sun and the Earth [32-34]. During that time, the surface of Earth was too hostile to host life or even complex organic molecules so whatever was present before or during the formation event, Earth was sterile at this point. The moon-forming impact and a potential late heavy bombardment phase kept conditions harsh up until around 3.9 Ga ago [33,35-38]. The first biological signatures could be dated back 3.7 Ga ago [33,39,40] and the first fossilized cells 3.4 Ga ago [33,41]. Mostly depending on the severity of the late heavy bombardment phase, Earth provided habitable conditions for 200 to 800 million years until life emerged [33].

Within this period, a complex molecular yet prebiotic evolution was likely to have taken place from which the first RNA may have emerged. As RNA can fulfill both

genetic and catalytic functions, theories of a so called RNA-world exist that state that at some point, molecular evolution or actual life was based mostly on RNA before DNA and proteins were established [42]. Eventually, the combination of these prebiotic processes is thought to have caused at least one instance of abiogenesis – the emergence of life from inanimate matter. These first lifeforms likely resembled anaerobic, autotrophic cells that were highly dependent on geochemical processes, like those being found at hydrothermal vents, to fuel their metabolism [43].

From these primordial cells, over time, a strain evolved that all extant life can be traced back to and is therefore referred to as the Last Universal Common Ancestor (LUCA) [44,45]. The use of proteins, the translation apparatus, and the genetic code may have been relatively late developments in the emergence of life, but all were fully available to LUCA. This is the reason why these and all other fundamental mechanisms are highly similar between all bacteria, archaea, and eukaryotes that LUCA eventually evolved into until today.

1.3.3 Origin and evolution of the genetic code

The genetic code of Escherichia coli was deciphered in 1963 [14,46] which was honored by the Nobel prize in 1968 [47]. Except for minor deviations in distinct species and mitochondria, the code is largely universal amongst all life [15,48]. In general, triplets of nucleotides act as $4^3 = 64$ different codons each of which is assigned to one of 20 to 22 amino acids or a termination signal. As some amino acids are encoded by more than one codon, the genetic code is designated as being *degenerate* or *redundant*. Two of the most special cases are the codons of selenocysteine and pyrrolysine in some species through the codons UGA or UAG respectively which otherwise serve as stop codons [49,50].

Within the extant code, multiple regularities could be identified rendering it highly nonrandom. Amongst the most prominent features are the general correlation that amino acids which are more frequently incorporated into proteins and possess a lower molecular weight are encoded by a higher number of codons than less frequent and heavier ones [51-54]. Also, the second or central nucleotide of each codon has the strongest influence on the properties of the encoded amino acid as well as the class of the involved aaRS [55-57]. In contrast, the third nucleotide exhibits in general the smallest influence on which amino acid is being encoded [58]. The code also seems to be organized in a block-like structure (Figure 4) which minimizes the impact of point mutations and erroneous codon recognition during translation as changes or misreadings of single nucleotides within a codon usually result in the incorporation of an amino acid with similar properties as the original one [59-61].

The origin and evolution of the genetic code could not yet be precisely elucidated, but several theoretical approaches have been undertaken to do so. These theories are not mutually exclusive, multiple of the following may – in part – be true in parallel [14].

First, neither the presence of 4 different nucleobases, the use of triplets as codons, nor 20 different amino acids must be taken for granted. All these parameters likely evolved over time from simpler precursors. A hypothetical genome containing only two different nucleobases could have stored information that was not disrupted upon the introduction of additional nucleobases. The use of other than three nucleotides per codon could also be thinkable, however a change in the reading frame would have rendered all information stored in the genome unreadable at that point. It appears therefore more likely that the primordial translation apparatus facilitated three nucleotides per codon from the beginning, but not all three nucleotides decided about which amino acid was to be incorporated. Translation machineries that moved other than three nucleotides per codon may have existed in parallel but eventually, a three nucleotides reading frame emerged as the most successful one [58]. The chronological emergence of amino acids may be debatable but most likely, the least chemically complex and most prebiotically accessible ones were available first followed by amino acids of increasing chemical complexity, molecular weight and evolutionary age of its cognate aaRS. This led to the following "consensus order: G, A, D, V, P, S, E, (L, T), R, (I, Q, N), H, K, C, F, Y, M, W" [18,62].

The genetic code may be seen as a *frozen accident*. But as the code strongly suggests the presence of optimized patterns, it is unlikely that these appeared randomly or all at once. Yet, at some point the genetic code became fixed in such a way that any further deviation resulted in a net loss of fitness. If a codon is being assigned to a different amino acid, multiple proteins may be affected by potentially detrimental alterations in their amino acid sequence as long as the genome does not undergo corrective mutations as well. The extant genetic code might therefore not be the most perfect one but one that could be gradually developed and may still contain minor yet acceptable flaws.

The evolution of the genetic code could have been driven by the concept of error minimization as initially proposed by the *Lethal Error theory* [63]. It basically states that the extant code is the evolutionary result of minimizing the impact of mutations which are more likely to be detrimental than beneficial to result in the extant code's properties described above.

Woese suggested the *Translation Error-Ambiguity* [64] which proposed that the primordial translation machinery was prone to errors as it could neither precisely distinguish codons nor amino acids. Therefore, amino acids of comparable

properties may have been used in groups as such. As its precision increased over time, similar amino acids remained encoded by similar codons.

The *Vocabulary Extension theory* assumes an early emergence of 64 codons and their complete assignment to primordially available amino acids. As more amino acids became evolutionary available, they were stepwise added to the code by occupying codons that were originally assigned to other amino acids with a preference to replace an amino acid by a similar one [58,65].

Lastly, there is also a theory about a potential stereochemical origin of the genetic code. During an RNA world or before the emergence of aaRSs, primordial tRNAs may have acted as their own aaRSs by binding their cognate amino acid directly with their anticodon. Thereby, the initial assignment of the genetic code would have been based on stereochemical interactions and the most suitable amino acids for such interactions would have been evolutionary preferred. An initial assignment like this may have also not been necessary for all primordial codons and amino acids as already short and simple polypeptides can have beneficial properties. A general difficulty in such a scenario would be to avoid the bound amino acid subsequently interfering with the interaction between the anticodon and the codon during translation [58,65].

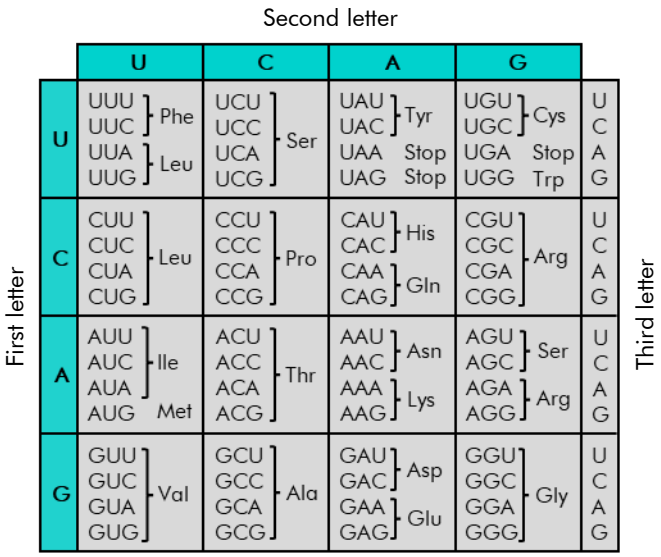


Figure 4. The genetic code as it is used by most extant organisms [48]. As indicated, the codons are sorted by their first, then second, then third letter. As the first and second letter have the strongest influence on the identity of the encoded amino acid, they appear sorted when arranged like this.

1.4 State of research

1.4.1 Hopfield's "testable hypothesis"

In 1978, John Joseph Hopfield, a physicist and Nobel Prize laureate in 2024, published a "testable hypothesis" about the stereochemical origin of the genetic code [66]. He described it as an "evolutionary paradox" that enzymes like aaRSs need correctly charged tRNAs to be produced and functional aaRS to charge tRNAs; a system that is far too complex to emerge at once.

At that time, it was well known that tRNAs could change their conformation depending on whether they were free in solution or interacting with the ribosome [66-73]. He therefore proposed a hypothetical alternative folding structure of primal tRNA that forms a stem loop at its acceptor stem which results in the anticodon loop being in close proximity to the 3'-end where the amino acid is being bound to (Figure 15 B). Based on its sequence, the anticodon could thereby influence aminoacylation and hydrolysis rates. Even though he did not introduce this term, this alternative structure will from here on be called *Hopfield fold*. After folding the sequences of 20 different tRNAs in *Escherichia coli* like this, the resulting six base pairs at the 3'-end were analyzed and revealed significantly elevated counts of correctly paired bases compared to random sequences (Figure 15 B). This nonrandomness may be interpreted as fossilized remnant from when tRNA actually formed this structure.

Hopfield also mentioned the conserved purine next to the anticodon and being the closest single stranded nucleotide to the 3' end. Only the alpha amino group of an L-amino acid but not of a D-amino acid could be able to interact closely with this purine, guiding it to the terminal 2' hydroxy group. As terminal 2' aminoacylation is not uncommon for tRNA, this could explain why L-amino acids were favored by evolution and why there is no tDNA as it lacks 2' hydroxy groups.

Assuming that also in this primordial system, aminoacylation and translation was spatially separated, some time had to pass until an aminoacylated Hopfield fold reached an active ribosome. During this time, Hopfield suggested, correctly charged Hopfield folds may undergo less hydrolysis compared to folds that carried an amino acid that did not match their anticodon. Through simulations he showed that even little differences in stability lead to a significant proofreading effect over time when incorrectly charged Hopfield folds were nearly completely hydrolyzed while some correctly charged ones were still present. Upon reaching the ribosome, the Hopfield fold may change its shape through interactions so that both the anticodon became available for codon recognition and the 3'-end for peptidyl

transfer, potentially realizing coded peptide synthesis without the involvement of enzymes.

1.4.2 SELEX

Beyond its functions in extant biology, RNA can also be structured to bind to specific target molecules or to catalyze certain reactions with the former being called aptamers and the latter ribozymes. A key methodology for the creation of artificial aptamers and ribozymes was established in 1990 and became known by *Systematic Evolution of Ligands by Exponential Enrichment* (SELEX). This approach uses DNA libraries that contain sequences for their effective handling but are besides that largely randomized. Such a DNA library then serves as template for in-vitro transcription which generates an equally randomized RNA library. Through different selection mechanisms, reactive *ribozymes* or selectively binding *aptamers* can be selected or enriched *in vitro* from this initial pool. By reverse transcribing such an enriched RNA library back to DNA again, a template library is being generated that is enriched in the selected sequences. By repeating these steps iteratively, singular best performing sequences could eventually be depicted [74-77].

1.4.3 Self-aminoacylating ribozymes

Through SELEX, Yarus and coworkers could repeatedly demonstrate the selection of ribozymes that could realize critical steps of translation without the involvement of enzymes. In one of their first instances [78], they used an initial pool of $1.7*10^{14}$ different RNA molecules with 50 randomized nucleotides flanked by constant regions; as $4^{50} = 1.3*10^{30}$, not all possible sequences were represented. From this library, they were able to select 35 different highly reactive RNA sequences that could aminoacylate their 3′-end using phenylalanyl-5′-adenylate as substrate. The aminoacylation conditions were set to pH 7 and 0 °C and included moderate concentrations of different mono- and divalent cations. The selection from the randomized library was conducted through the use of Naphthoxyacetyl-N-hydroxysuccinimide ester (NHS ester) after aminoacylation which covalently linked a hydrophobic tag to the amino group of the amino acid. This caused a substantial shift in subsequent HPLC purification through which reacted and non-reacted RNAs could be separated from each other. As part of further investigations and verifications, they also subjected selected ribozymes to oxidation through periodate. Periodate readily reacts with vicinal diols just like in the case of RNA 3′-ends to form aldehydes [79,80]. If a ribozyme was subjected to periodate oxidation before

aminoacylation, no product could be obtained, which suggested that the site of self-aminoacylation was the 3'-end. If oxidation was conducted after aminoacylation, the 3'-end was protected from oxidation through the bound aminoacyl moiety. So, the selected RNA conducted self-aminoacylation in a way similar to extant aaRSs and tRNAs [78]. However, these RNA sequences possessed little similarities to extant tRNA or any obvious precursor.

In a following project, the selection and characterization of both an unspecifically and specifically self-aminoacylating ribozyme could be demonstrated. With a length of around 70 nt the specific ribozyme lay well within the range of extant tRNAs [81].

So far, the 5' and 3'-terminal regions of the DNA and RNA libraries was kept constant for simpler handling during reverse transcription, PCR, and in vitro transcription. This, however, inhibited the optimization of the 3' terminal region which may have a major influence on the reactivity of the selected RNAs. In a later approach, Yarus and coworkers therefore introduced a ligation step after selection which allowed the 3' region of the RNA library to be randomized, too. It could be observed that a relatively high number of different structures showed activity which suggests that self-aminoacylation was a rather accessible task for RNA to achieve [82].

Further stepwise optimization of these ribozymes eventually led to the development of a ribozyme consisting of only 5 nucleotides with the additional ability to *act in trans* and to facilitate formation of peptidyl-RNA. [82-84] Similar motifs could be identified in rRNA and were positively tested for activity even though they were constrained by the structure of their surrounding sequences [85].

1.4.4 Kinetic sequencing

In 2019, Chen and coworkers utilized an approach similar to classical SELEX but in combination with Illumina high throughput sequencing. Hence, they termed their approach *sequencing to measure catalytic activity paired with in vitro evolution* (SCAPE). They used an RNA library consisting of 21 randomized nucleotides flanked by constant regions. Their initial pool size was therefore $4^{21} = 4.4 * 10^{12}$ different sequences which they reacted with a 5(4H) oxazolone derived from O-methyl-tyrosine and a biotin containing moiety. They chose using an oxazolone as these compounds were considered to be more prebiotically plausible than adenylates [86-90]. Upon aminoacylation, sequences carried both a tyrosyl and a biotin moiety so reacted sequences could be purified from unreacted ones through interaction with streptavidin. After five rounds of classical SELEX, the pool was

reduced to ~10^4 to 10^5 sequences of elevated reactivity. This pool was reacted with the oxazolone at different concentrations and the reacted fraction was purified and subsequently sequenced. Read numbers could therefore be used to quantify the number of reacted sequences from which, in combination with a standard for normalization, the kinetic rate constants of all members of this pool could be determined. Therefore, this step was termed kinetic sequencing (k-seq). Gel shift assays of singular sequences served as control experiments and verified the data obtained by sequencing. Eventually, an evolutionary network of self-aminoacylating RNA sequences could be described with its members exhibiting comparable activity to the selected ribozymes from Yarus and coworkers which used adenylates for self-aminoacylation [90].

1.4.5 Abiotic adenylate synthesis

Adenylates can be synthesized chemically by multiple different approaches described in literature. One of the most established procedures, which was summarized by Paul Berg in 1958 and employed by Yarus and coworkers as described above, used N, N'-dicyclohexylcarbodiimide (DCC) as coupling agent to synthesize aminoacyl-5'-adenylates from amino acids and adenosine monophosphate in a one-step reaction. The material was subsequently washed with various organic solvents and then dried, typically resulting in around 50% yield. This product can be further purified by column chromatography; originally, cation adsorption was employed which quantitatively removed excess adenosine monophosphate from the solution and eluted the final adenylate product [91].

In 1979, Armstrong et al. reviewed the synthesis of aminoacyl-5'-adenylates and revealed several flaws of using DCC as coupling agent. Not only was the according synthesis cumbersome and time consuming, it also resulted in around 20% of the product consisting of an unwanted isomer in which the amino acid was bound to the 2' or 3' hydroxy group of the AMP through an ester linkage. This isomer had the same mass and similar chromatographic properties as the desired adenylate, but it was much less reactive [92,93]. Therefore, they developed a superior synthetic approach which used N-t-Boc protected amino acids that were reacted with sec-butyl-chloroformate first to yield an intermediate mixed anhydride which in a second step was reacted with AMP to yield protected aminoacyl-5'-adenylate. The protecting group could subsequently be cleaved off rapidly and quantitatively using trifluoroacetic acid (Figure 5); combined with several washing steps in anhydrous organic solvents, alanyl-, phenylalanyl-, and methionyl-5'-adenylate could be obtained with yields of around 50% in regard to AMP as the limiting factor. Successful synthesis was proven through elemental analysis, infrared, and ^{1}H-NMR

spectroscopy; no major impurities like the aforementioned ester could be observed. In conclusion, the proposed method was faster, more robust, and applicable to a wider variety of amino acids than all other methods reviewed [92].

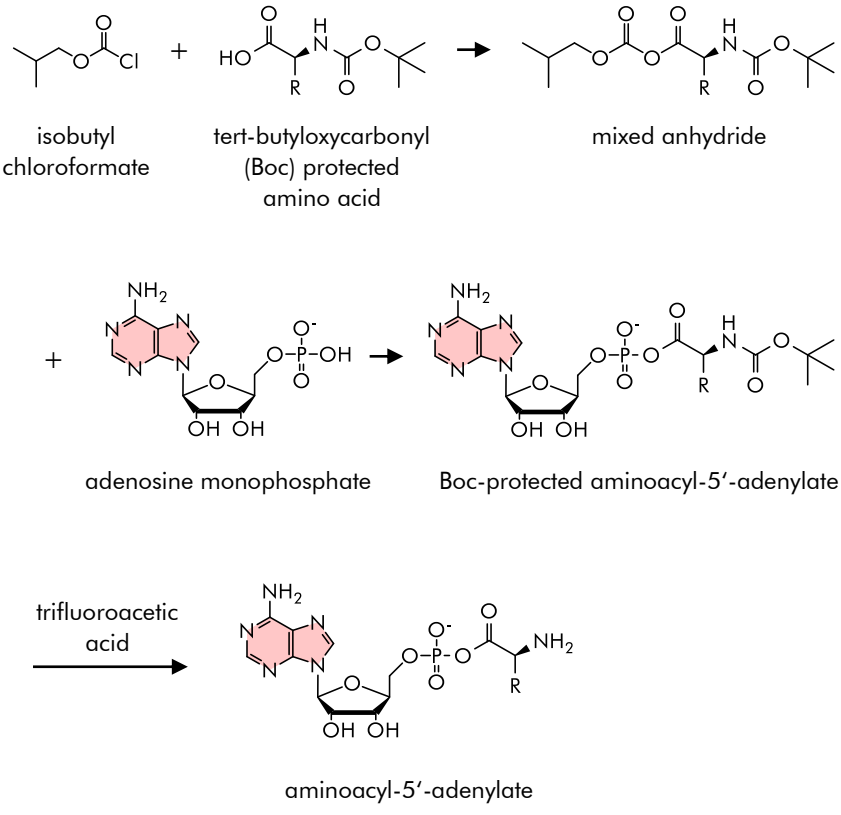


Figure 5. Adenylate synthesis scheme as published by Armstrong et al. in 1979 [92]. Originally, the isomeric sec-butyl chloroformate was used to form a mixed anhydride with an amino acid protected with a tert-butyloxycarbonyl (Boc) group at its amino moiety. In a subsequent step, this mixed anhydride is brought into contact with adenosine monophosphate (AMP) to form still Boc-protected aminoacyl-5'-adenylate. In the last step, the Boc group was cleaved off through the use of trifluoroacetic acid.

1.4.6 APB-PAGE

In 1985, an affinity electrophoresis method was developed that facilitated the copolymerization of **a**cryloylamino**p**henyl**b**oronic acid into **p**oly**a**crylamide gels used for **g**el **e**lectrophoresis (APB-PAGE). This method covalently fixed a boronic acid moiety to the polyacrylamide gel matrix. As vicinal diols, like native 3'-ends of

RNA, form transient diesters with boronic acid, their movement was slowed by this interaction during electrophoresis (Figure 6). It was shown that native tRNA and tRNA with 3'-ends oxidized by periodate but otherwise exhibiting the same length and sequence could be effectively separated using this technique [94].

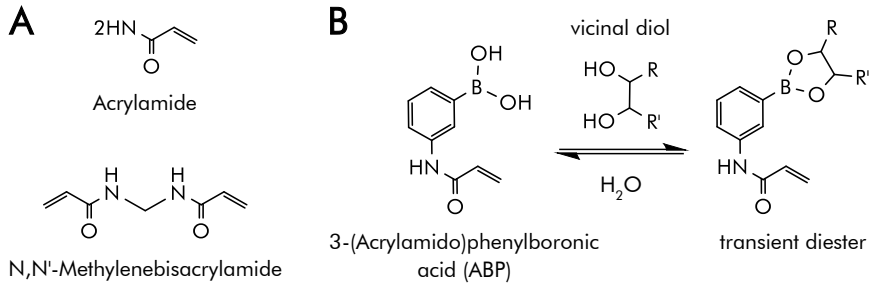


Figure 6. (A) Monomeric components to form polyacrylamide gels. Lower amounts of the bisacrylamide crosslink linear polymer strands formed by the acrylamide. **(B)** 3-(Acrylamido)phenylboronic acid (ABP) is able to copolymerize with the components named in (A). This fixes boronic acid residues within the gel matrix which in turn can form transient diesters with vicinal diols ultimately slowing down molecules that contain such diols during electrophoresis.

2 Motivation and objectives

In 1978, John Hopfield proposed a "testable hypothesis" about the stereochemical origin of the genetic code. It claimed that Hopfield folds may have been precursors of extant tRNAs but refolded to position the anticodon loop close to the 3' end. In extant biology, the 3' end of tRNAs is aminoacylated with the cognate amino acid by proteins but aminoacylated tRNAs are also necessary to produce such proteins. Hopfield folds may solve this evolutionary dilemma by being capable of self-aminoacylation with specific amino acids mediated by their anticodon acting as chemical sensor in a sequence-dependent manner. These as well as other sequences and structures may have left molecular vestiges that persisted ever since and may be found in extant tRNA sequences and the genetic code.

Hopfield's hypothesis combined with the repeated demonstration of the simplicity and robustness of self-aminoacylating ribozymes by Yarus and coworkers, the kinetic sequencing as employed by Chen and coworkers, as well as the availability of chemical synthesis routes for activated amino acids, gave rise to the research question of this thesis. Based on previous proposals by Prof. Jäschke, I planned to use an RNA library that closely resembled Hopfield folds including the native 3' terminal CCA and to randomize the usual seven nucleotides that form the anticodon-loop. This Hopfield fold RNA library would therefore contain $4^7 = 16{,}384$ different sequences which would fall well within the range of sequences submitted to kinetic sequencing by Chen and coworkers, rendering previous SELEX obsolete. To ensure statistical significance, it was desired to obtain 100 reads for each possible sequence. This would lead to $1.6*10^6$ reads per sample which combined with the capacity of Illumina sequencing in 2019 of 10^6 to 10^8 reads per run would result in the parallel analysis of tens to hundreds of samples in a work- and cost-efficient manner.

The Hopfield fold RNA library was meant to be mixed with activated amino acids like adenylates as used by Yarus and coworkers or oxazolones as used by Chen and coworkers under various conditions like concentration, temperature, pH, additives, etc. followed by oxidation with periodate. As demonstrated by Yarus and coworkers, 3'-terminal aminoacylation protected the RNA from being oxidized and alkaline treatment could be used to hydrolyze aminoacylated RNA [95]. This would eventually yield a mixed RNA library of which some part carried native 3' ends which resembled sequences that successfully underwent aminoacylation and another part with oxidized 3' ends which did not. Thereby, a labile partly aminoacylated RNA library would be converted into a stable partly oxidized RNA library. This *oxidative fixation* would thereby simplify handling and minimize losses of aminoacylated sequences of interest through hydrolysis (Figure 7).

The following planned steps included the separation of RNA molecules with native 3′ ends from molecules with oxidized 3′ ends. To achieve this, two main approaches were considered. Either APB-PAGE could be employed to separate native from oxidized RNA through affinity electrophoresis or ligation could be used as oxidized 3′ ends could not form ester bonds with 5′ phosphoryl groups.

Eventually, a library preparation procedure was planned to be established to convert partly oxidized Hopfield fold RNA libraries into DNA libraries compatible with Illumina sequencing. It was desired to establish a procedure with a minimum number of total steps, especially intermediate purifications, with maximum yields of each step to introduce as little biases as possible. During library preparation, missing sequencing primer binding sites as well as indices for multiplexing were to be added. The region of interest, i.e. the randomized anticodon loop, was though flanked by constant regions, which harbored minimal complexity. As described before (Chapter 1.2.4), low complexity is highly detrimental for Illumina sequencing. To circumvent this problem, another region of randomized nucleotides was to be introduced during library preparation in such a way that it was the first nucleotides to be sequenced. Randomized sequences exhibit the highest complexity and by using several different lengths for this region, a frameshift could be realized between the clusters increasing overall complexity during sequencing of the constant regions, too. The introduction of another randomized sequence after aminoacylation could also be used to identify PCR duplicates as multiple assemblies of the same pair of one of 16,384 different anticodon loops with one of a larger number of secondary randomized sequences was rather unlikely.

In the obtained sequencing data, after demultiplexing, exclusion of PCR duplicates, and application of possible other filters, the obtained abundance of each sequence would serve as proxy for its reactivity. DNA or RNA of suitable sequence and known concentration may be used as spike-in for normalization. From this data, the kinetics of all the 16,384 sequences may be assessed to elucidate the plausibility of Hopfield's testable hypothesis. Or to state it in Hopfield's own words: "The level of discrimination and which bases are important in amino acid-anticodon interactions must be determined experimentally." [66]

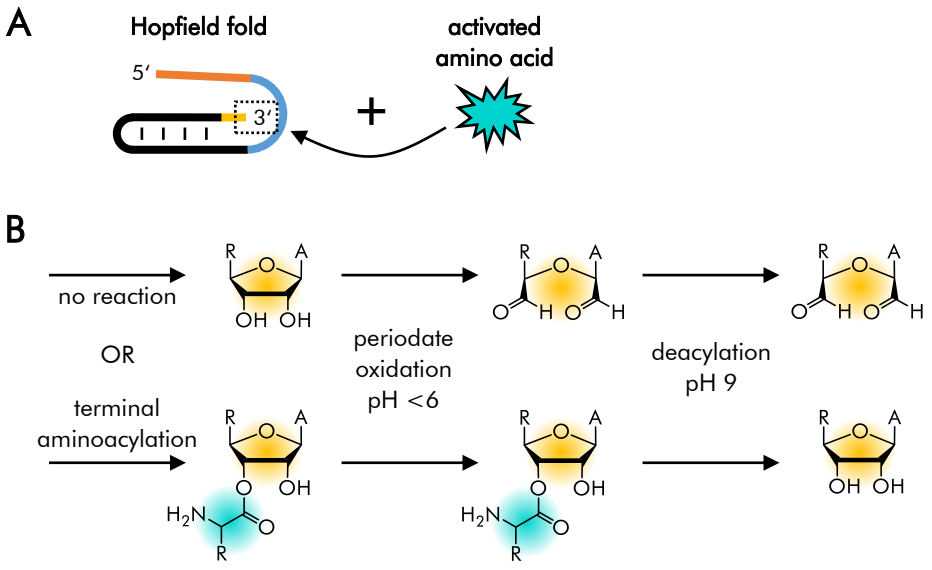


Figure 7. (A) Aminoacylation of a Hopfield fold RNA molecule with a chemically activated amino acid. Solely terminal 2′ and 3′ aminoacylations were relevant for this project. **(B)** Aminoacylation and oxidative fixation of the 3′ end as indicated in (A); R represents the rest of the RNA and A the nucleobase adenine. In the first step, either no reaction or terminal aminoacylation is indicated. Next, under acidic conditions and through the use of periodate, only native 3′ ends undergo oxidation; aminoacylated ones undergo no further change. After quenching of excess periodate, alkaline conditions are realized which cause bound amino acids to be cleaved off, leaving native 3′ ends behind. Oxidized 3′ ends undergo no change in this step.

3 Results and discussion

3.1 Method development

3.1.1 Adenylate synthesis

Setup

Primarily adenylates, rather than oxazolones, were chosen as activated amino acids for this thesis for several reasons. Although aminoacyl-5′-adenylates are less prebiotically plausible than oxazolones, they are utilized by extant aaRS and are synthetically and analytically more accessible.

At the onset of my work on the chemical synthesis of aminoacyl-5′-adenylates, I collaborated with Prof. Dr. John Sutherland (MRC Laboratory of Molecular Biology) and his research group. Longfei Wu, a member of the group, advised me to follow the approach of adenylate synthesis published by Armstrong et al. [92] (Chapter 1.4.5). Based on his recommendation, I purchased free AMP salt and titrated it with tetrabutylammonium hydroxide solution to form the tetrabutylammonium-salt of AMP (TBA-AMP), contrasting with the originally prepared trioctylammonium-AMP salt. I also replaced sec-butyl chloroformate with its isomer, isobutyl chloroformate, due to its ready availability and near-identical structure.

Since the anhydrides formed during the synthesis were labile to hydrolysis, it was necessary to work under anhydrous conditions. I achieved this for all solvents and reactants through extensive lyophilization or molecular sieves. However, to simplify the process, I omitted the use of an anhydrous atmosphere and conducted all syntheses under ambient conditions. To increase the number of experimental attempts, I scaled down the original recipe tenfold and utilized molecular biology-grade plasticware instead of conventional chemical glassware. The quality and yield of the reactions were assessed primarily using a combination of mass spectrometry and ^{31}P-NMR spectroscopy. A detailed list of all modifications I made to the original process is provided in Table 1.

Category	Armstrong et al.	This thesis
Reaction scale	1-fold (~10 g)	0.1-fold (~1 g)
Labware	Chemistry grade glassware	Molecular biology grade plasticware
Anhydrous conditions	Rotary evaporation only	Lyophilization, use of molecular sieves
Ratio amino acid to AMP	2:1	4:1
Incubation temperature	Mostly 4 °C	Mostly ambient temperature
Organically soluble AMP salt	Tri-n-octylammonium salt of AMP	Tetra-n-butylammonium salt of AMP
First step activation compound	Sec-butyl chloroformate	Isobutyl chloroformate
Analysis	^{1}H-NMR, infrared spectroscopy	^{31}P-NMR, MS
Yield of alanyl-5′-AMP	60%	Near quantitative
Yield of phenylalanyl-5′-AMP	43%	Near quantitative
Yield of methionyl-5′-AMP	42%	0%

Table 1. Summary of all modifications made to the original adenylate synthesis procedure described by Armstrong et al [92].

Iterative optimization

During my laboratory work, different incubation times, temperatures, and amino acid-to-AMP ratios were iteratively tested, and techniques to ensure anhydrous reaction conditions were established to optimize the synthesis of aminoacyl-5′-adenylates. The optimization process was guided by two main priorities. First, canonical L-α-amino acids that could be protected with compatible protecting groups and yielded high amounts of aminoacyl-5′-adenylates were prioritized. The second priority was the aminoacylation yields achievable using these adenylates. Amino acids that performed well in both categories received additional optimization efforts, resulting in near-quantitative yields for lysyl-, alanyl-, and glycyl-5′-adenylates, as these exhibited the best overall performance and acceptable stability. Conversely, less work was invested in optimizing adenylates derived from amino acids that could not be effectively converted, yielded poorly during aminoacylation, were unstable, or exhibited solubility issues (Table 2).

Amino acid (adenylate)	Yield (protected)	Yield (deprotected)	Estimated half-life at pH 5	Aminoacylation yield
L-/D-Lysine	100%	93%	1.5 h	Very high
L-Lysine-CMP	100%	67%	0.5 h	High
L-/D-Alanine	100%	89%	1.5 h	High
Glycine	100%	93%	1.5 h	High
L-Valine	100%	67%	12 h	Medium
L-Isoleucine	100%	100%	24 h	Medium
L-Phenylalanine	-	64%	5 h	Low
L-Proline	-	52%	1 h	Low
L-Leucine	-	59%	5 h	Low
L-Tryptophan	100%	88%	0.5 h	Low
L-Histidine	76%	56%	Seconds	-
L-Tyrosine	76%	53%	Solubility issues	-
L-Threonine	9%	0%	-	-
L-Serine	-	0%	-	-
L-Asparagine	100%	0%	-	-
L-Methionine	100%	0%	-	-

Table 2. Aminoacyl-5′-adenylate yield of different amino acids. Most recent yields as measured by ^{31}P-NMR = c(adenylate) / (c(adenylate) + c(AMP)). In the cases of lysine and alanine, both enantiomers behaved similarly.

In contrast to the results reported by Armstrong et al., methionyl-5′-adenylate could not be successfully obtained. Quantitative turnover was achieved up to the deprotection step; however, all deprotection attempts resulted in complete hydrolysis of the adenylate. Modifications to incubation time, temperature, and the use of different deprotection agents, such as 4 M HCl in dioxane or Reagent H [96],

yielded no success. Reagent H, developed to protect sulfur-containing moieties from oxidation, was tested extensively by Suria Morales Guzmán during a research internship. She hypothesized that this reagent might be less harsh than pure trifluoroacetic acid for deprotection, but no intact methionyl-5′-adenylate was obtained. Attempts to dry trifluoroacetic acid with molecular sieves resulted in a rapid reaction that formed a thick brownish mass, rendering it unusable for deprotection purposes. No further attempts were made to optimize the deprotection process, as the observed losses were eventually deemed acceptable.

Adenylates derived from amino acids that contained hydroxy groups exhibited poor or no yields, likely due to solubility issues, as synthetic attempts remained cloudy or formed precipitates over time. Interestingly, the synthesis of prolyl-5′-adenylate uniquely caused the reaction mixture to turn a vibrant red immediately after the addition of isobutyl chloroformate, reverting to colorless upon deprotection. No practical or theoretical explanation for this phenomenon could be identified, as no conjugated π-electron systems—typical of organic dyes—were formed at any stage.

Later, the adenylates of D-lysine and D-alanine were synthesized, displaying behavior identical to their L-enantiomers. Additionally, the synthesis of lysyl-5′-cytidylate was successfully achieved with near-quantitative turnover prior to deprotection and 63% yield post-deprotection. This process followed a similar protocol to that used for adenylates but required dimethylformamide (DMF) as the sole solvent during synthesis, as other solvents caused strong precipitation upon addition of TBA-CMP. DMF was also essential for the synthesis of arginyl-5′-adenylate, as the starting material, asparagine, was protected with both a Boc group at the amino moiety and a 9-xanthenyl protecting group at the amide moiety, the latter impeding dissolution in dioxane.

Under my supervision, Jennifer Sauerland tested the influence of different incubation temperatures on adenylate yield during her internship but no striking effect could be observed. It was therefore concluded to conduct syntheses at room temperature from then on [97]. The same applied to Suria Moralez Guzmán but we were testing different incubation times and came to the conclusion that the originally proposed were already optimal except for valine and isoleucine for which incubation times had to be increased due to reduced reaction rates [98].

Analysis

Mass spectrometry consistently revealed that the techniques employed predominantly produced the expected adenylate and free AMP, as these were the most prominent signals detected across multiple samples. Additional peaks indicated the presence of tetrabutylammonium as the main contaminant, along

with traces of cyclic AMP (cAMP), which was absent in the AMP starting material (data not shown).

During the first step of the reaction, isobutyl chloroformate was used in slight excess relative to the protected amino acid. This excess may have reacted with the free amino group of the subsequently introduced AMP, potentially explaining the repeated formation of trace molecules that were 100 u heavier than expected (Figure 9, Supplementary Figure 1 & 2).

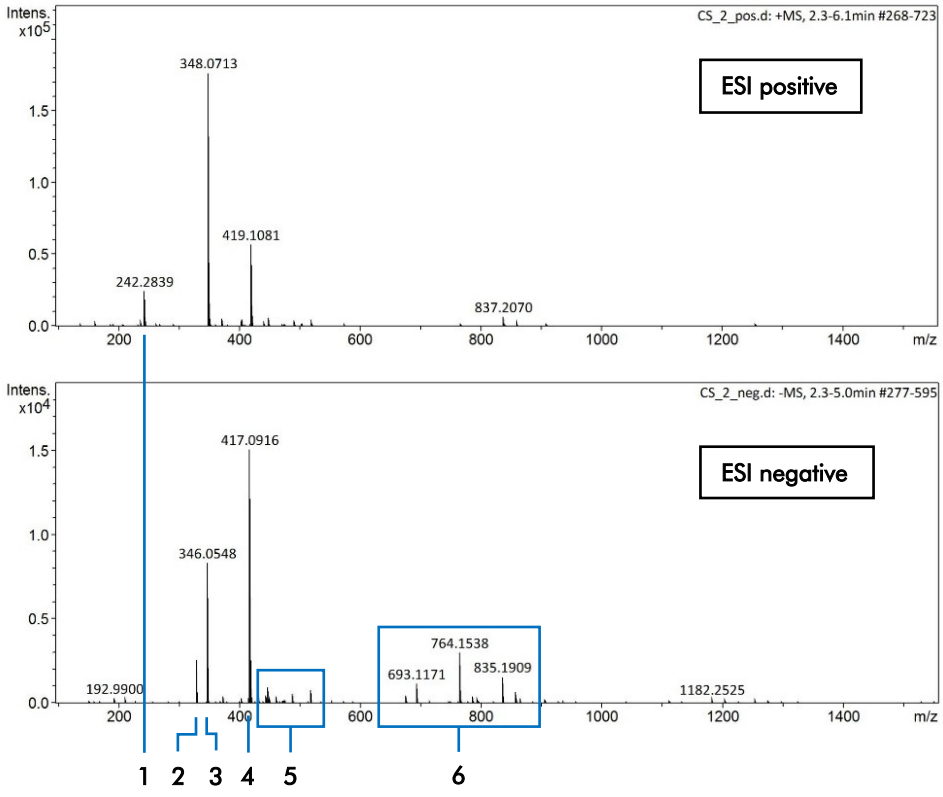
Following mass spectrometric analysis, which indicated the absence of major impurities in the final adenylate product, the purity was quantified using ^{31}P-NMR spectroscopy. The analysis focused on the relative content of aminoacyl-AMP and free AMP. Free AMP generated a signal near 0 ppm, while aminoacyl-5′-adenylates produced a signal at approximately 8 ppm. These signals were integrated, and the purity was calculated ratiometrically using the formula provided in Figure 8.

$$I(x) = Integral\ of\ signal\ x \qquad \%\ purity(aa\text{-}AMP) = \frac{I(aa\text{-}AMP)}{I(aa\text{-}AMP) + I(AMP)} * 100$$

Figure 8. Formula for the calculation of adenylate purity based on ^{31}P-NMR measurements.

This approach was used to monitor the purity of the adenylate product after each synthetic process and to assess its stability during aminoacylation experiments. Aminoacyl-5′-adenylates were ultimately synthesized with near-quantitative purity, as confirmed by ^{31}P-NMR spectroscopy and distinct signals in mass spectrometry, as described above.

2′- or 3′-aminoacyl esters of AMP were the most likely undesired side products. The mass of an aminoacyl-5′-adenylate and that of a 2′- or 3′-aminoacyl ester was identical. Their ^{31}P-NMR spectra were not, as in the case of the ester, the phosphoryl group remained unchanged compared to AMP. As in the spectra of later adenylate products a clear signal corresponding to the adenylate and none for AMP could be observed, the presence of 2′- or 3′-aminoacyl esters could thereby be excluded. Mass spectrometry showed no peak corresponding to aminoacyl-5′-adenylates carrying additional aminoacyl esters at their 2′- or 3′-hydroxy groups, excluding their presence, too.



#	Compound	m/z calculated		m/z found	
		$[M+H]^+$	$[M-H]^-$	ESI pos.	ESI neg.
1	Tetrabutylammonium	242.47 [M]		242.284	-
2	cAMP	330.213	328.199	-	328.0
3	AMP	348.227	346.213	348.071	346.055
4	Ala-AMP	419.306	417.292	419.108	417.092
5	Isobutylformiate-adducts	+100 of previous		mostly found	
6	Combinations of previous signals i.e. typical artifacts				

Figure 9. Exemplary mass spectrometric analysis using electrospray ionization (ESI) of the L-Ala-AMP product. Analyses of different adenylate products gave similar results matching the molecular weight of the used amino acid (Supplementary Figure 1 & 2).

Chromatographic purification

Before achieving near-quantitative purity directly after synthesis, efforts were made to purify certain aminoacyl-5'-adenylates using reverse-phase column chromatography. Substantial work in this regard was conducted by Suria Moralez

Guzmán and Jennifer Sauerland during their research internships [97,98]. Chromatographic purification was attempted with various aminoacyl-5′-adenylates available at that time, but only L-Leu-AMP, L-Phe-AMP, and L-Ile-AMP produced chromatograms with two distinct peaks, allowing effective separation of the components. The first peak was primarily attributed to free AMP, while the second peak corresponded to the adenylate. The material of the second peak was collected and lyophilized and typically showed a far increased purity compared to the input material (Figure 10).

In contrast, attempts to purify L-Lys-AMP, L-Ala-AMP, and Gly-AMP yielded only a single peak, making separation unachievable. The hydrophobicity of the amino acid appeared to play a critical role in successful separations, as amino acids involved in effective purifications were markedly more hydrophobic than those where separation failed. Given the labor-intensive nature of these purifications, their limited applicability, and the eventual increase in adenylate synthesis yields, further efforts toward purification techniques were discontinued.

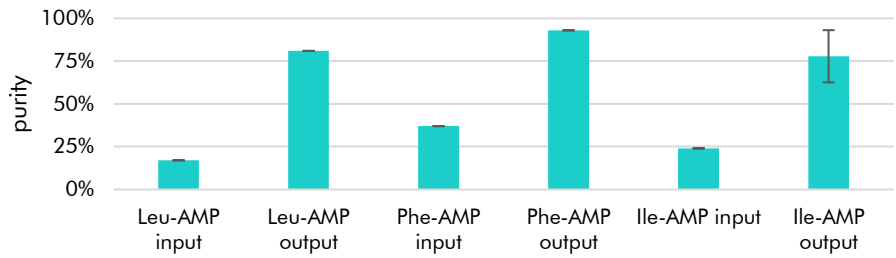


Figure 10. Chromatographic purification of different L-aminoacyl-5'-adenylates. In all cases, only a single sample served as input or was obtained except for Ile-AMP; one larger batch served for a total of seven purification attempts. Its graph shows the mean value and standard deviation. The purifications of Leu-AMP and Phe-AMP were conducted by Suria Moralez Guzmán, and the Ile-AMP samples were purified by Jennifer Sauerland, both under the supervision of the author of this thesis [97,98].

Long-term stability

After achieving near-quantitative yields for L-Lys-AMP, L-Ala-AMP, and Gly-AMP, a larger batch of each compound was produced to facilitate all corresponding aminoacylation experiments throughout the project. During laboratory work, data on the long-term stability of these adenylates under storage conditions were collected. The adenylates were stored as lyophilized powders under an anhydrous argon atmosphere at –20 °C and exhibited a degradation rate of approximately 1% per month (Figure 11).

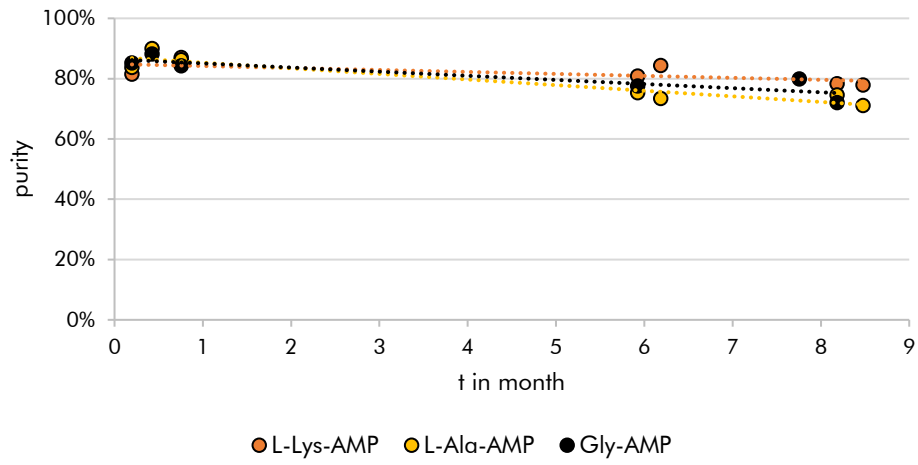


Figure 11. Long-term stability of selected adenylates under storage conditions (dry powder, anhydrous argon atmosphere, -20 °C). The average loss appeared to be around 1% per month.

3.1.2 Aminoacylation

To determine the optimal conditions for the aminoacylation of RNA Hopfield folds using aminoacyl-5′ adenylates, several factors were carefully considered and balanced. The rationale behind each component is explained below. In general, the experimental design prioritized simplicity to maintain prebiotic plausibility while maximizing aminoacylation efficiency.

<u>Amount of RNA</u>

The amount of RNA used per experiment was constrained by several factors. At the lower end, reduced quantities led to increased relative losses and handling challenges. At the upper end, excessive RNA input caused overloading effects during subsequent library preparation steps, as the employed enzymes were only effective within specific concentration ranges. Upscaling these reactions would have led to unnecessary product excess and increased costs. To balance these considerations, 1 μg of RNA per aminoacylation reaction was determined to be an appropriate quantity.

Buffer reagents

After synthesis, aminoacyl-5'-adenylates were obtained as lyophilized powders, most of which were readily soluble in water or aqueous buffers. Upon dissolution, the pH decreased significantly due to the acidic nature of the adenylates. Precise pH control was essential for the subsequent reaction with RNA-based Hopfield folds to achieve aminoacylation, as pH strongly influenced RNA folding. The lower limit of the acceptable pH range was determined by RNA denaturation, with 4.45 corresponding to the pK_a of cytosine [99]. The upper limit was constrained by the decreasing stability of the adenylates, which were highly unstable at pH 7 or above (Figure 41 B).

Acetic acid was selected as the primary buffering agent due to its simple chemical structure and pK_a of 4.75, allowing effective buffering within the pH range of 3.75 to 5.75. This range was considered prebiotically plausible. Furthermore, acetic acid did not interfere with RNA folding, the aminoacylation reaction, or ^{31}P-NMR measurements.

In preliminary experiments, phosphoric acid was employed to adjust buffers within the pH range of 1 to 3. This buffer system was compatible with subsequent oxidative fixation but incompatible with ^{31}P-NMR spectroscopy. In other cases, 2-(N-morpholino)ethanesulfonic acid (MES), with its pK_a of 6.15, was used to buffer reactions at pH 6 or 7. Prior to oxidative fixation, the pH needed to be lowered, as higher pH values inhibited oxidation with periodate. ^{31}P-NMR spectroscopy was fully compatible with MES.

Amount of adenylate

Initially, I selected 1 mg of adenylate per 1 μg of RNA, resulting in a mass ratio of 1000:1, to maximize aminoacylation yields. This amount also represented the practical lower limit of material required for NMR measurements. Reaction volumes were set at a minimum of 20 μl, as relative losses due to evaporation and pipetting errors were acceptable within this range. Both 1 mg of adenylate and 1 μg of RNA dissolved completely in 20 μl of water, and 500 mM acetate buffer effectively maintained the pH within the desired range.

However, as 500 mM buffer systems are unlikely to have been present in prebiotic scenarios, during her internship, Suria Moralez Guzmán and I explored the possibility of neutralizing the adenylates post-synthesis. This involved dissolving the adenylates in water, adjusting the pH with diluted sodium hydroxide, and subsequently lyophilizing the solution. Regardless of the pH adjustments, large portions or all of the adenylate were hydrolyzed when tested after the lyophilization process was finished. Consequently, this approach was not pursued further. If

successful, it could have enabled the use of less concentrated buffer systems while still maintaining the desired pH [98].

Divalent cations

Mg^{2+} cations can heavily influence RNA folding and the proper function of ribozymes and aptamers. They were therefore included in the selection buffers of Yarus and coworkers and Chen and coworkers. I decided to include them as well, adapting the value of 5 mM employed by Chen and coworkers [90,100].

Two-vial design

I figured that for aminoacylation reactions, a larger setup had to be prepared so that 20 µl could be withdrawn after each desired time interval. These 20 µl would immediately undergo oxidative fixation to fixate their state of aminoacylation at that time. In contrast to the originally planned 1 mg of adenylate, which the buffer system was tested for, I eventually had to reduce the amount of adenylate to 300 µg per reaction due to issues in oxidative fixation.

For every aminoacylation reaction, I prepared in a separate vessel another reaction to monitor the initial state and rate of hydrolysis of the used adenylate through ^{31}P-NMR spectroscopy as previously described. Analogous to the aminoacylation setup, this separate reaction shared the same experimental conditions, reagents and was also upscaled so that 20 µl could be withdrawn per NMR measurement. The only differences were that it contained no RNA and the full 1 mg of adenylate per 20 µl as 1 mg was the practical lower limit of adenylate to be used per NMR measurement. The RNA could be saved because I could observe that the presence or absence of RNA did not have an influence on the adenylate's hydrolysis rate (data not shown). The composition of aminoacylation reactions and the parallelly ongoing hydrolysis reaction (NMR setup) are summarized in Table 3.

The repeated withdrawal was necessary because the 20 µl had to be mixed with 480 µl of D_2O to facilitate proper NMR measurements. In this diluted form though, the adenylates were considerably more stable than in the original form (data not shown). Therefore, after each desired time interval, another fresh dilution had to be prepared from the ongoing reaction master mix. At least at the start of every aminoacylation reaction, I conducted one such ^{31}P-NMR measurement to ensure the used adenylate did not undergo hydrolysis prior to the start of the reaction. In some cases, I also conducted further ^{31}P-NMR measurements after suited time intervals to thoroughly monitor the hydrolysis rate of the used adenylate in parallel to the ongoing aminoacylation reaction (Figure 47).

Compound	NMR setup	Aminoacylation setup
RNA	none	1 µg
Adenylate	1 mg	300 µg
Buffer compound	500 mM	500 mM
$MgCl_2$	5 mM	5 mM

Table 3. Chemical composition of adenylate hydrolysis (NMR setup) and aminoacylation reactions. Acetic acid was the main buffer component and used within pH 4.5 and 5.5. Phosphoric acid was used below pH 4.5 and 2-(N-morpholino)ethanesulfonic acid (MES) was used for pH values above 6. In any case, the pH was adjusted with NaOH.

3.1.3 Oxidative fixation

Oxidation

According to the project plan, RNA had to undergo oxidative fixation using periodate directly after aminoacylation (Figure 7). As previously mentioned, the initially planned amount of adenylate had to be reduced from the targeted 1 mg to 300 µg per aminoacylation reaction as during Lina Bingel's internship, we first noticed that too high amounts of adenylate or AMP might quench the periodate so that not all RNA could be oxidized that was supposed to [101]. Shortly after, Jennifer Sauerland and I finished optimizing the oxidative fixation during her internship which resulted in the final form of this procedure [97].

After oxidative fixation, the RNA had to be purified from the salts involved to avoid interference with the subsequent library preparation. Normally ethanol precipitation was employed for this purpose, but this caused strong, unwanted co-precipitation. The use of lithium chloride was a common alternative but, in this case, it did not cause the RNA to precipitate, most likely because this technique was better suited for longer strands. Through trial and error, it was eventually found that dilution with water and the use of isopropanol led to the desired result of only the RNA precipitating. This dilution was to be minimized to maximize the yield of precipitated RNA and a reasonable optimum was found at 4 µmol of periodate per reaction. Depending on the involved amino acid, 300 µg of adenylate corresponded to around 0.74 µmol, but this value had to be doubled to 1.48 µmol as each molecule of adenylate could react with periodate via both its contained ribose and amino acid [79,102]. This resulted in a 2.7-fold surplus of periodate to adenylate. In this context, 1 µg of the RNA 60-mer involved in this reaction corresponded to 52 pmol and could therefore be neglected.

Quenching

To quench excess periodate after the oxidation reaction, I initially tried to use the glycogen that was supposed to be added to the precipitation reaction anyway, but I realized that too large amounts would be required. Since glycogen co-precipitated with the RNA and was quite expensive, it could not be added in large quantities. I therefore shifted to the simplest available diol to be used as a quenching agent, namely pure ethylene glycol, which had also been used for this purpose in other studies [103,104]. Due to its high viscosity, I wanted to minimize the risks of pipetting errors and decided to add 4 µl to the oxidation reaction, resulting in 71 µmol or a 48-fold surplus towards periodate to be added. This also eliminated any risk associated with ethylene glycol absorbing water from the ambient atmosphere, thereby diluting itself over time (Figure 12 A).

Incubation times

Through APB-PAGE, it could be shwon that the employed amounts of periodate and ethylene glycol were sufficient to achieve complete oxidation and quenching within just 30 seconds of incubation each (Figure 12 B & C). Rapid oxidation and quenching were critical to prevent hydrolysis of aminoacylated RNA during these steps, which would have otherwise led to false negative results. To ensure complete turnover, I set both incubation times to 3 minutes each. After quenching, the samples remained stable, and during more complex experiments, such samples were regularly left at room temperature for up to an hour without observable detrimental effects.

Formation of formaldehyde

In retrospect, I realized that the reaction of periodate converted some of the ethylene glycol into formaldehyde, which can react with both proteins and nucleic acids, forming cross-links [105]. This may have impacted the subsequent library preparation procedure, and no other quenching agents were tested that did not release formaldehyde. For reference, the concentration of periodate before quenching was 100 mM; due to stoichiometry, this may have led to a maximum concentration of 200 mM of formaldehyde after quenching. Typical fixation experiments use final concentrations of 1.3 M formaldehyde which is 6.5-fold more [106].

Random crosslinking would have likely increased variation amongst replicate samples, but as only minor variance could be observed in both bioanalyzer readouts, sequencing and densitometry, either crosslinking did not interfere with

library preparation, no crosslinking occurred or crosslinking exclusively occurred in specific sequences affecting their performance consistently.

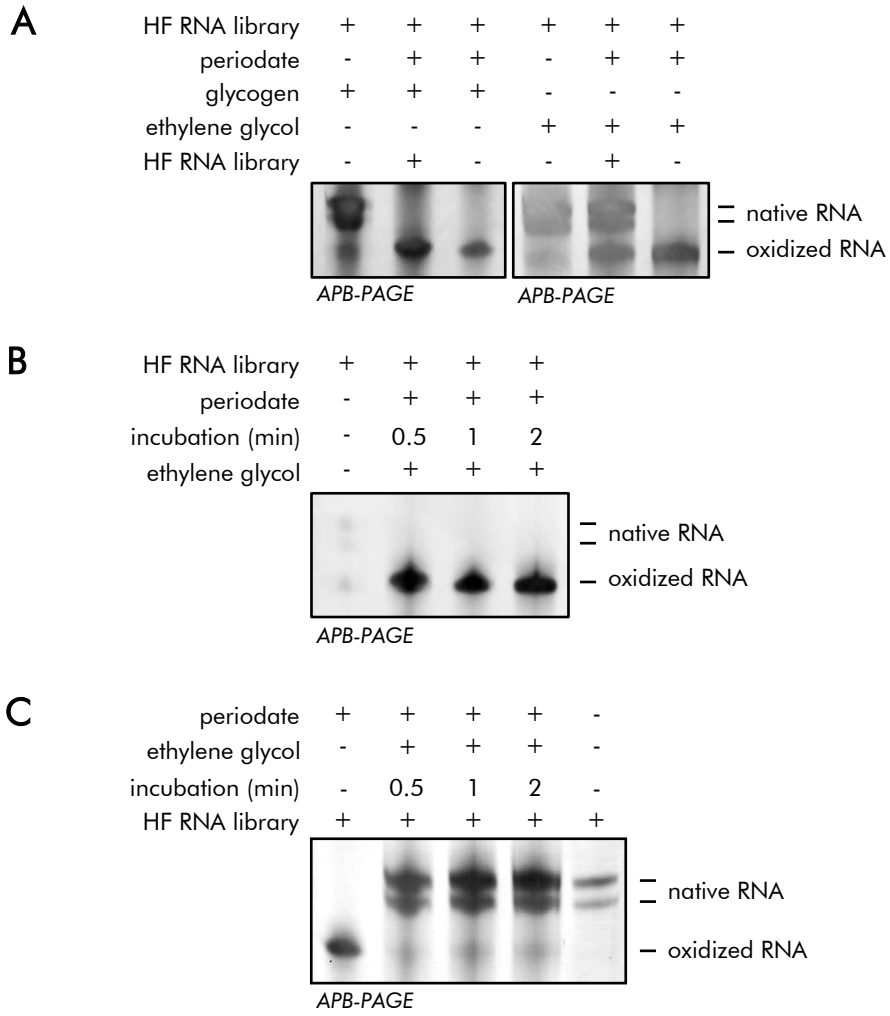


Figure 12. Optimization of periodate oxidation of the Hopfield fold RNA library with subsequent quenching to bottom. **(A)** Comparison of glycogen and ethylene glycol as quenching agents for excess periodate. RNA was added at two time points to the applicable reactions. In this case, the use of glycogen served as an example of insufficient quenching of excess periodate. **(B)** Kinetics of oxidation reaction. **(C)** Kinetics of quenching.

3.1.4 Deacylation

Following oxidation and quenching, all remaining aminoacyl esters on the RNA molecules' 3' ends needed to be cleaved to restore their native diol structure, which was critical for subsequent APB-PAGE and ligation. Deacylation is a standard procedure in tRNA research, typically performed using Tris-HCl buffer at pH 9 [95]. After oxidation and quenching, the reaction was buffered in 44 µl of ~250 mM acetate buffer, usually at pH 5. To ensure a complete shift to the desired alkaline conditions, the reaction was mixed with an equal volume of 1 M Tris pH 9. Measurements with pH paper confirmed that pH 9 was achieved. Incubation at 37 °C for 30 minutes was employed for most amino acids, while valine and isoleucine required 60 minutes based on prior studies of tRNA deacylation [107].

Although RNA is generally prone to hydrolysis at elevated pH and temperature, the Hopfield fold RNA library remained intact under these reaction conditions, as no smear of lower molecular weight was observed in any corresponding PAGE or APB-PAGE (Figure 13, Figure 14). However, an additional upper band appeared after the RNA library underwent oxidation, quenching, and deacylation. This band was absent when only one or two of these treatments were applied and was visible only upon APB-PAGE, not PAGE (Figure 13). This observation suggested the addition of diol-containing moieties to the RNA and ruled out the possibility of hybridization products. Since Tris base contains three hydroxy groups per molecule, it was suspected to be responsible for this effect, potentially retarding RNA migration in APB-PAGE.

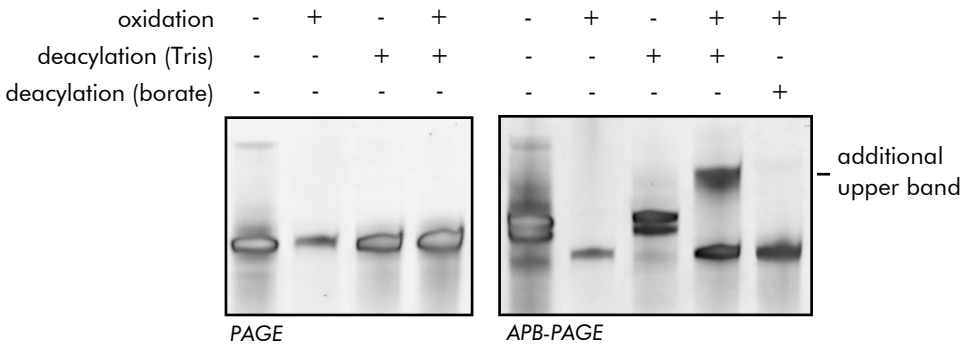


Figure 13. Formation of additional upper band only in the case of APB-PAGE, oxidation, and deacylation with Tris combined. As this effect only applied to oxidized material but not to non-oxidized material of interest (and the two lanes behaved the same as shown in the other gel), this effect was neglected.

To test this hypothesis, I substituted the Tris buffer with 1 M borate buffer at pH 9 under identical deacylation conditions. This prevented the formation of the additional upper band (Figure 13). However, the borate buffer precipitated heavily within 24 hours of preparation, making it significantly more labor-intensive to use than the Tris buffer. Given that the effect occurred only with oxidized RNA, which was not a focus of further investigations, the Tris buffer was retained for deacylation.

The untreated Hopfield fold RNA library consistently formed two bands upon APB-PAGE but only one band upon PAGE. I speculated that one portion of the library harbored a modification influencing its chemical properties. To investigate, I separated the two bands by APB-PAGE, excised and extracted each band, and conducted oxidation and deacylation on each material separately. Both initial bands formed the additional upper band following oxidation, quenching, and deacylation in Tris buffer, demonstrating chemical equivalence in this regard. Furthermore, no interconversion between the two bands was observed under standard buffer conditions at room temperature (Figure 14). These findings confirmed that the 3′ ends of the entire library were intact.

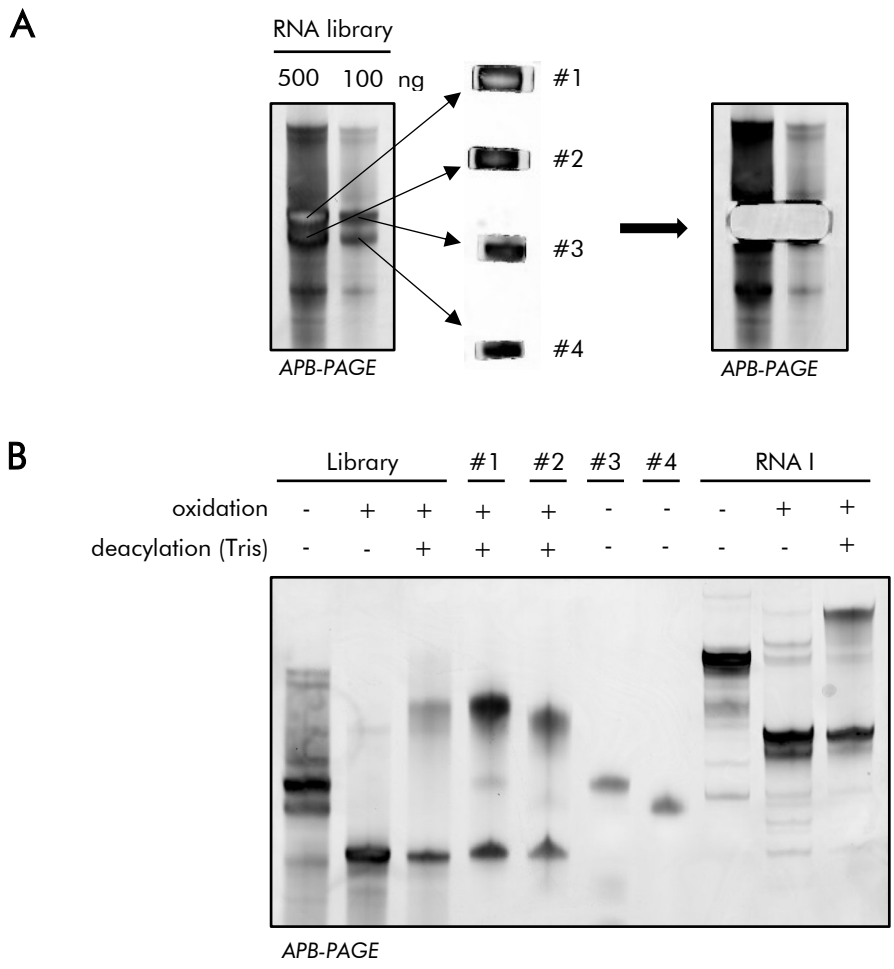


Figure 14. (A) APB-PAGE of 500 ng and 100 ng of Hopfield fold RNA library; subsequently, four snippets were excised to separate the two main bands which were extracted through electroelution. **(B)** Oxidation and deacylation with Tris buffer on the entire Hopfield fold RNA library, its upper and lower main band, as well as another sequence called RNA I.

3.1.5 Library design

I designed the Hopfield fold RNA library under consideration of several factors. To fit to the primary research question, it had to resemble the originally proposed Hopfield fold. Therefore, it had to possess a 3′ terminal region that ended on the nucleotides CCA just like all extant tRNAs did. Secondly, it had to form a stem-loop

structure at its 3′ end that brought a single stranded part mimicking the anticodon loop in close proximity to the 3′ end. In terms of Hopfield's hypothesis, mainly the sequence of the 3′ end as well as this single stranded part in its proximity were relevant. The sequences that formed the stem loop were interchangeable as well as further sequences in the 5′ region.

The Hopfield fold RNA library could either be produced by in-vitro transcription (IVT) or chemical synthesis. The disadvantages of IVT were that especially the 3′ end of the produced RNA tended to be heterogeneous and to ensure that the transcribed RNA library only contains consistent CCA 3′-termini would have made the use of e.g. a hammerhead ribozyme necessary as well as subsequent purification [108-110]. The only real disadvantage of chemical synthesis was its limitation of the maximum product length which was much less of a limitation in the case of IVT. At the time of this thesis, commercially available chemical synthesis of RNA was limited to 60 nucleotides total length. To ensure the integrity and purity of the Hopfield fold RNA library, it was decided to purchase a chemically synthesized RNA library which limited the design to 60 nucleotides.

I tried to mimic Hopfield's proposed original sequence as much as possible by primarily eliminating non-canonically paired nucleotides and by keeping the ratios between the nucleotides as constant as possible. The melting temperature of the stem was designed to be about 50 °C to ensure proper folding during experiments at room temperature whilst not inhibiting reverse transcription which is prone to be blocked by stable secondary structures [111,112]. The loop was shortened from 7 to 4 nucleotides to harbor the unspecific sequence of UAAU. Later on, I also tested an alternative library that contained an UUCG-motif instead which was known to stabilize stem-loop structures [113]. This, however, resulted in markedly decreased yields of the final sequencing library (data not shown). It was not investigated which step was mostly affected by this change, but increased stability generally impedes denaturation which in this case was critical for reverse transcription and PCR. Therefore, exclusively the UAAU loop was used in all Hopfield fold sequences herein.

As previously described (Chapter 2), the stretch of usually seven nucleotides that constituted the anticodon loop were to be randomized, resulting in 4^7 = 16,384 different anticodon-loop sequences.

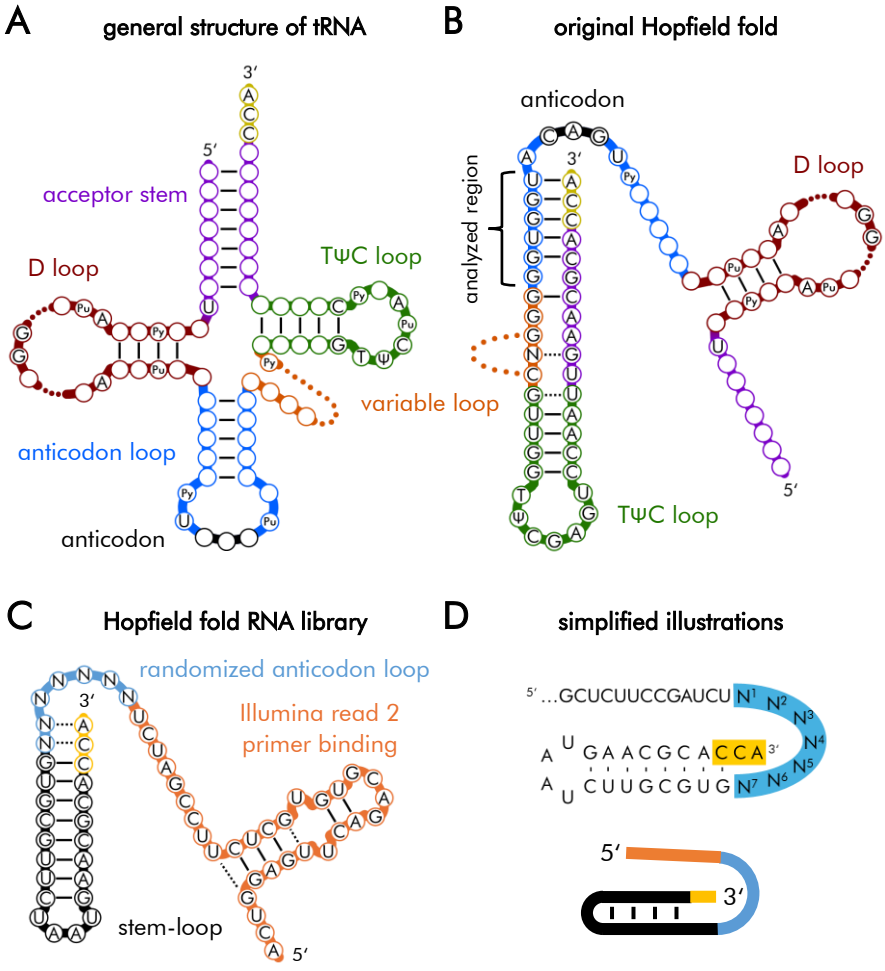


Figure 15. **(A)** Generalized structure of tRNA [114]. **(B)** Structure and sequence of originally published Hopfield fold. Neither the bulge of the variable loop (orange dots) nor the structures 5′ of the anticodon were originally illustrated. According to Hopfield, the analyzed region showed significant enrichment of canonical base pairs in E. coli tRNA sequences [66]. **(C)** Design of the Hopfield fold RNA library derived from Hopfield's proposed structure. **(D)** Same sequence and structure as in (C) but in simplified form.

Hopfield's hypothesis was focused on a structure of 43 nt but I could realize a similar structure including all previously mentioned elements using only 29 nt. The remaining 31 available nucleotides were used to attach an Illumina read 2 sequencing primer binding site onto the 5' end of the randomized anticodon loop. Adding this sequence already into the Hopfield fold RNA library itself notably simplified subsequent library preparation as the Illumina read 2 sequencing primer binding site was planned to be added anyways. At this position, it was located so that the first sequenced nucleotides of read 2 would be the randomized ones of the anticodon loop. By chance, this Illumina sequence formed a comparable structure to the D-loop that would form in the case of an actual tRNA sequence (Figure 15).

Despite their generally comparable structure, the sequence of the Hopfield fold RNA library was significantly shorter and different to the one of the originally proposed Hopfield folds which were based on extant tRNA sequences. Even though the main focus of Hopfield's hypothesis was on the acceptor stem and anticodon loop, other structures like the TΨC loop and the D loop would still be intact if an extant tRNA would be folded into a Hopfield fold (Figure 15). These two loops and other unpaired nucleotides may well continue exhibiting their usual interactions or form new ones. As such considerations as well as the influence of modified nucleotides were neglected by Hopfield's original hypothesis, they were also not considered in the design of the Hopfield fold RNA library.

To facilitate proper Illumina sequencing, more sequences had to be added to the Hopfield fold RNA library as it also had to be converted into a dsDNA library. To achieve this, I conceived two different approaches, both starting with Hopfield fold RNA libraries after oxidative fixation. In one approach, RNA with native 3' ends would be separated from the oxidized ones through APB-PAGE followed by reverse transcription using primers with 5'-overhangs. The other approach was based on ligation of the fixated library without prior purification as only native 3' ends could be ligated whereas oxidized ones could not. Only the ligation product would harbor a primer binding site to enable subsequent reverse transcription. From then on, both approaches would continue identically with overhang-PCR and product purification. In the following, the two approaches will be called by their different selection steps for RNA with native 3' ends namely *APB-PAGE-based library preparation* and *ligation-based library preparation* (Figure 16).

In both cases, the final PCR was to be conducted identically. Using overhang primers, both Illumina P5 and P7 sites were attached as well as i5 and i7 indices if necessary. In the case of the i7 indices, I adopted the sequences provided by NEBNext Adaptors and Primers Set 1 for Illumina [115]. All i5 indices I designed on my own under the considerations to maximize complexity and that at least two mutations were necessary to convert one index into another.

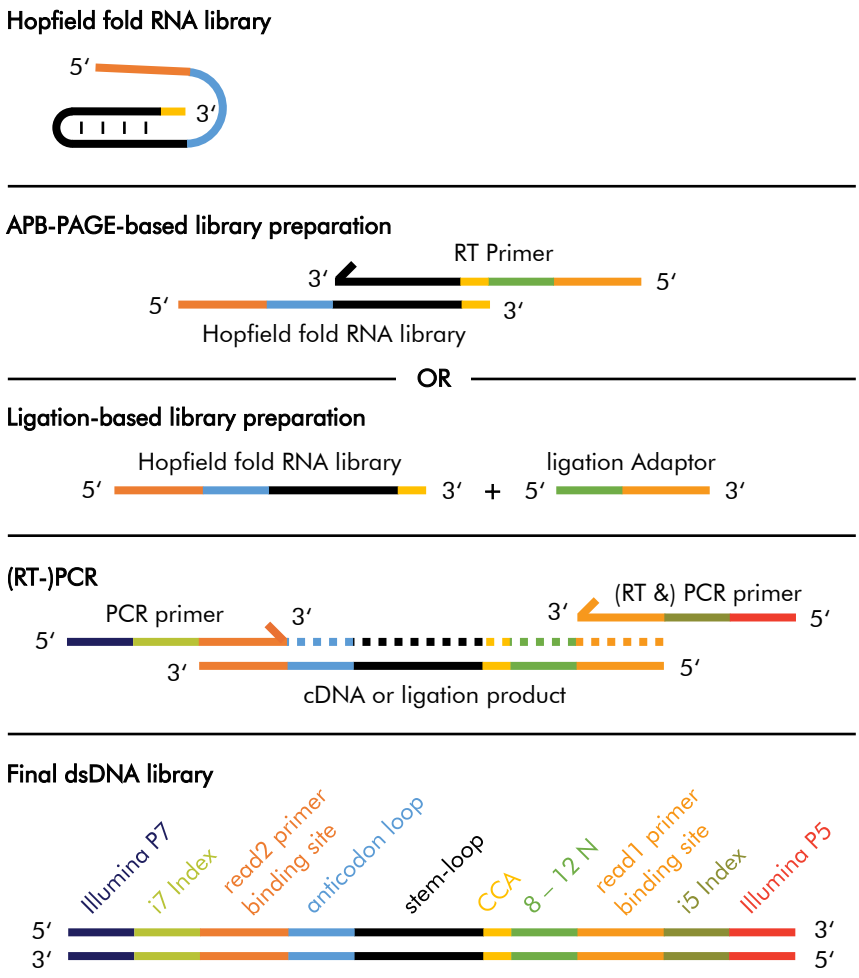


Figure 16. Library preparation procedure. The Hopfield fold RNA library represents the only RNA in the scheme, all other sequences depict DNA. Arrow-shaped lines indicate primers that were elongated upon reverse transcription (RT) or polymerase chain reaction (PCR). See *Final dsDNA library* for color coding; Illumina P5/7 facilitates hybridization with the Illumina flow cell and bridge amplification. The inclusion of indices was optional on both sides; primers with and without different indices were available. Also, 5 different ligation adaptors were available each with 8 to 12 randomized nucleotides (*8 – 12 N*).

3.1.6 APB-PAGE-based library preparation

Optimization of gel elution

In this approach, native and 3' oxidized molecules of partially oxidized Hopfield fold RNA libraries which underwent aminoacylation and oxidative fixation were to be separated from each other through the use of APB-PAGE. I intended to excise the band resulting from the native material and to extract it from the gel cutout. I therefore compared four different elution techniques, two of which extracted the genetic material directly into solution whereas in the case of the other two methods, the material was first transferred from the APB-PAGE cutout into an agarose gel from which it was then excised again and eventually purified.

As one of the most well-established direct extraction methods, I tested the so-called *crush & soak* method [116]. Its main advantage was its simplicity; the APB-PAGE cutout just had to be shredded and incubated in acetate buffer overnight which caused the genetic material to diffuse into the solution. Its main disadvantage lied in difficulties in fully separating the shredded material from the solution again and the increased chance of losing entire samples as pellets formed during ethanol precipitation tended to become loose easiest compared to all other methods tested (Figure 17 C).

Due to these disadvantages, I continued by testing the two indirect methods; in both cases, APB-PAGE cutouts were placed into the wells of ordinary 2% agarose gels in TBE. Upon gel electrophoresis, the genetic material traveled from the APB-PAGE cutout into the agarose gel as a clear band and not as a smear. I then excised this band and extracted the genetic material through the so-called *freeze & squeeze* method [117,118]. For that, I froze the agarose gel cutout and centrifuged it at room temperature over a spin filter to separate solid from liquid material, the latter of which contained the genetic material. After ethanol precipitation, the desired genetic material could be obtained in acceptable purity and quantity. Nonetheless, this method introduced avoidable losses and impurities, and its main disadvantage was its high labor intensity as especially the precise excision of the genetic material from the agarose gel was cumbersome.

As an alternative indirect method, I tested the G-CAPSULE [119] which was a single use device made of plastic and a dialysis membrane a few centimeters in total size. Its plastic part was shaped so that it could punch out the desired band from an agarose gel and contain the resulting piece. Through subsequent electrophoresis, the genetic material was moved from the gel into a small volume of buffer that was enclosed by the plastic, the punched-out agarose gel and the dialysis membrane which the genetic material could not pass through but was concentrated at. After shortly reversing the electrical current to fully transfer the genetic material from the

membrane into the enclosed buffer, this buffer could be extracted by piercing the dialysis membrane and the genetic material could be recovered through ethanol precipitation. In general, this method had no practical advantages to the aforementioned freeze & squeeze method and the single use G-CAPSULES were rather expensive. I therefore did not further use this method.

Eventually, I developed another direct elution method strongly inspired by a published method [120]. I therefore crafted an insert made of Styrofoam, cuvettes, and adhesive tape that – alongside with a regular comb – could be inserted into a solidifying agarose gel (Figure 17 A). After solidifying, the comb and insert were removed which resulted in an agarose gel with regular wells and aligned notches of about 1*1*1 cm in size. I took care to form the bottom of the notches below the bottom of the regular wells so that genetic material coming from these wells could not travel below the notch within the agarose gel. Upon electrophoresis, I first filled the notches with running buffer followed by the chamber. I adjusted the water line of the buffer to match the upper edge of the gel but not to cover it. This prevented genetic material from diffusing out of the notches as during agarose gel electrophoresis, the gel was usually completely submerged in buffer. As I placed cutouts from APB-PAGE gels into the regular wells and applied voltage, the contained genetic material traveled as a band through the agarose gel and into the buffer within the notches. The process took around 10 min and could be monitored through bromophenol blue which traveled at the same speed as the Hopfield fold RNA library (Figure 17 B). As soon as the genetic material was fully inside the buffer of the notch, I stopped the electrophoresis, extracted the ~1 ml of buffer from the notch and purified the genetic material through ethanol precipitation. After removing the depleted APB-PAGE cutouts and refilling the emptied notches, the agarose gel could be reused for further APB-PAGE cutouts. Due to its superior turnover, low labor intensity as well as high yield and purity of the eluted genetic material, I chose this method as the standard for all subsequent extractions from APB-PAGE gels (Figure 17 C).

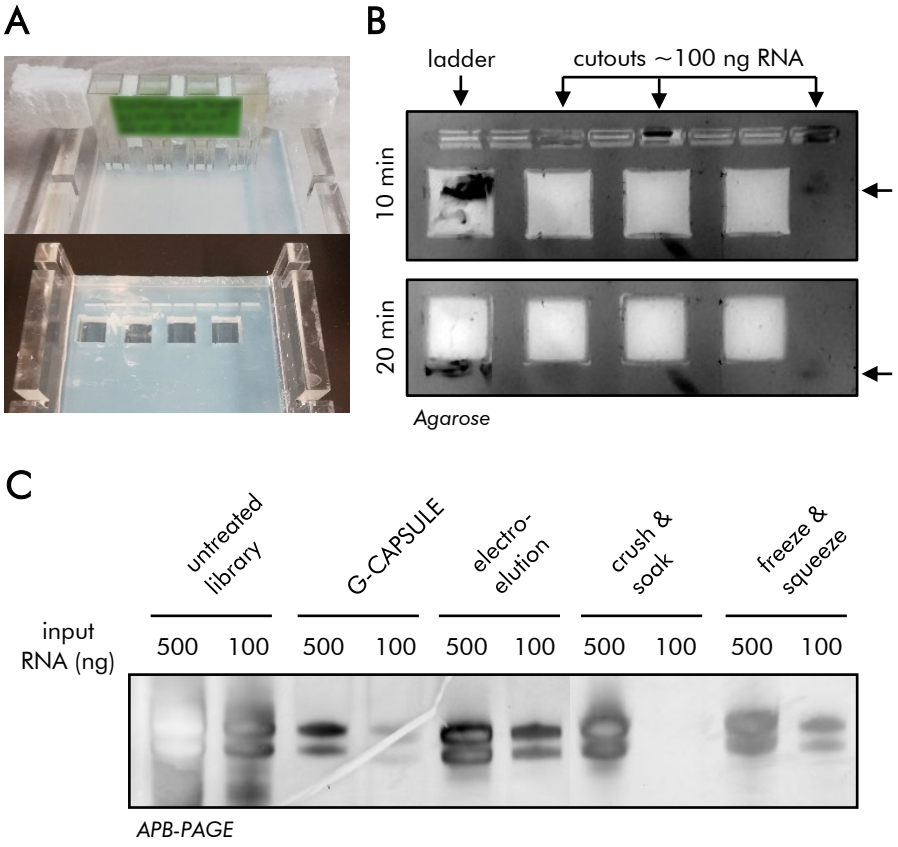


Figure 17. Illustration of the electroelution setup and comparison with other gel extraction techniques. **(A)** Self-crafted insert made of Styrofoam, cuvettes, and tape to form 1*1*1 cm molds into agarose gels alongside regular wells. **(B)** Electroelution of 5 µl of Ultra Low Range DNA Ladder and cutouts of the main bands of Hopfield fold RNA library after APB-PAGE inserted into the indicated wells. Images were taken after 10 and 20 minutes of standard electrophoresis. Arrows on the right-hand side indicate the rightmost material traveling in the shape of a band. **(C)** APB-PAGE of either the untreated library directly or cutouts of the main bands of the library after previous APB-PAGE and elution through the depicted methods. 500 or 100 ng were used as input material for each approach. The 100 ng sample of the Crush & Soak approach was lost; the loss of entire samples was a typical disadvantage of this method.

Semi-quantitative RT-PCR

To assess the efficiency of the entire library preparation process, I conceived to conduct the planned reverse transcription (RT) and polymerase chain reaction (PCR) in a semi-quantitative way during method development. The conceived final

RT-PCR method was planned to undergo a fixed number of thermocycles whereas the semi-quantitative variant was to set up a total of 25 to 30 cycles and to withdraw a sample once every 5 cycles. These samples were eventually analyzed through agarose gel electrophoresis to estimate the minimum number of cycles necessary to form a visible product band. The lower the number, the more efficient the entire library preparation process was. Minimizing the number of cycles also reduced the number of identical copies within the final DNA library as well as reducing the introduction of any reproduction bias.

I could not observe any difference whether I used Hopfield fold RNA library that underwent APB-PAGE, excision of the main band and subsequent electroelution or the same amount of input material directly for library preparation (data not shown). Therefore, this process could be excluded as a major source of variation.

However, even though the amount of input Hopfield fold RNA library was visible on both APB-PAGE and agarose gels, it always took at least 20 cycles of PCR to produce a visible product band (Figure 18 B). This corresponded to 20 duplications which roughly multiplied the input material by one million. I therefore concluded that in between electroelution and PCR either >99% of the genetic material was lost or that the reverse transcription, PCR, or both worked with <1% efficiency. As there was no further cleanup in between electroelution, reverse transcription, or PCR, and as I could never observe major losses of material from general handling, I dismissed that any material was adsorbed, hydrolyzed, or otherwise lost during the process and focused mainly on the RT-PCR itself. As assigned by the manufacturer of the used DNA polymerase, I let the reverse transcription reaction make up 10% of the volume of the PCR. This dilution could have caused a maximum of 4 additional cycles and no major impact on the robustness of PCR was reported by using unpurified RT reactions as template. I therefore focused my attention solely onto the reverse transcription reaction. The cDNA that was supposed to be generated was expected to be 101 nt in length so considerably longer than both the input RNA library and the DNA primer. Yet, I could not observe the formation of any higher product band in PAGE upon reverse transcription. In the contrary, I could clearly observe that a product was repeatedly formed which was apparently shorter than both of the aforementioned RNA and DNA input. Typical for DNA, this product was unaltered by NaOH treatment. Higher concentrations of DNA primer led to correspondingly stronger formation of this short product (Figure 18 A). A study found that betaine could reduce the melting temperature of DNA and RNA and could thereby have beneficial effects on reverse transcription reactions [121]. I therefore decided to test if the addition of betaine could also be beneficial for this reverse transcription however, no positive effects could be observed (Figure 18 A). I also tested different incubation temperatures within the range of 50 °C to 65 °C but none had any effect on the outcome (Figure 18 A). Lastly, I set up a 1-cycle PCR reaction containing only buffer, RT primer, and Q5

polymerase which resulted in the formation of the same short product as could previously be observed in reverse transcription reactions (Figure 18 A). Upon purifying this short product and using it as template in a regular 30-cycle PCR, no expected product was formed (Figure 18 B).

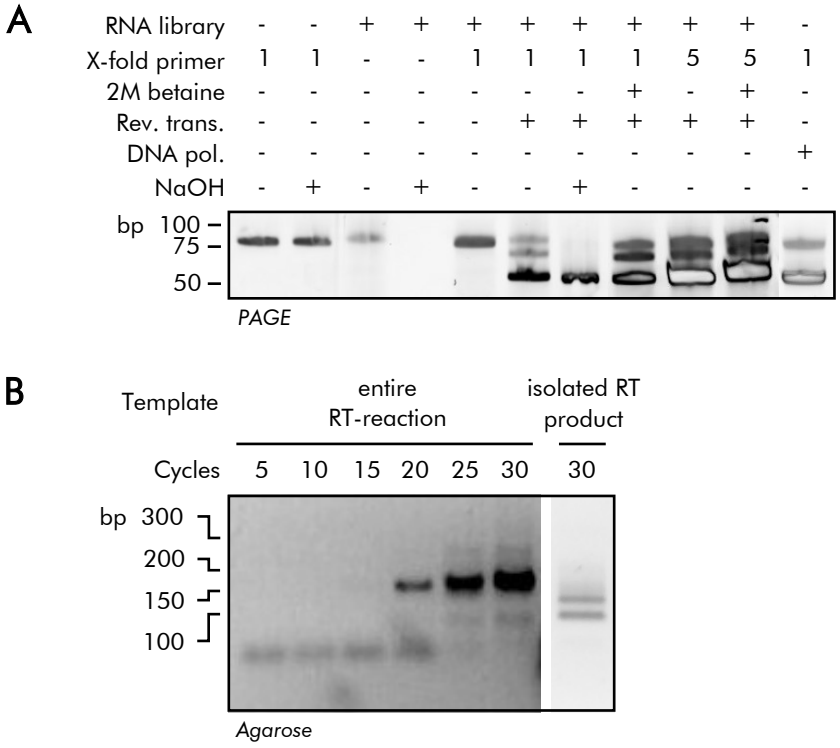


Figure 18. Key findings while optimizing the approach of direct reverse transcription. **(A)** Usual concentrations of Hopfield fold RNA library and reverse transcription primer were used where applicable. Even though the used primer had a length of 63 nt and the RNA library of 60 nt, their corresponding bands formed repeatedly higher than anticipated. The white filling of some bands indicates a surplus of material that could not be fully stained. Reverse transcription reactions including RNA library, 1-fold primer, and reverse transcriptase as depicted in the middlemost lane, looked identical when conducted at 50, 56.4, and 65 °C during elongation or when incubated for 60 min instead of the usual 10 min during elongation. **(B)** At least 20 cycles of PCR were necessary to form visible product bands at around 170 nt which formed only if crude reverse transcription reaction was used as template. If a lower band as depicted in (A) was cut out, purified and used as template, a 30 cycle PCR resulted in the pattern as depicted on the right-hand side.

Putting all this together, I reasoned that the used reverse transcription primer was likely forming a stem-loop at its 3' end just like the one in the Hopfield fold RNA library it was supposed to bind to. This likely facilitated self-priming which caused both the reverse transcriptase (SuperScript IV) and the DNA polymerase (Phusion) to elongate the 3' end of the primer using its single stranded 5' region as template. This may transform the 63 nt long mostly single stranded DNA primer into a 53 bp long double stranded DNA product which may likely have caused the shift towards appearing shorter. In a last attempt, I designed an alternative primer for reverse transcription that was shortened at its 3' end so it could not for undergo self-priming as easily. Even though I tested different incubation temperatures during reverse transcription, no difference could be observed in semi-quantitative PCR as still around 20 cycles were needed to form visible material (data not shown).

To further investigate the PCR product, I ligated it into a plasmid through blunt end ligation and used the product to transform E. coli with. I picked 5 of the many resulting colonies to expand and purify enough plasmid from each sample to submit to sanger sequencing. The results almost completely matched expectations; besides some minor deletions which may happen during cloning, expanding, or sequencing errors, the sequences were as expected. Also, the two randomized regions could be easily identified as they differed between each sample and sample 4 showing mixed signals in its electropherogram (Figure 19).

Despite the positive sequencing results, I decided to no longer pursue the development of the APB-PAGE based library preparation due to its low efficiency and little potential to optimize the reverse transcription reaction. Instead, I continued with the development of the ligation-based library preparation.

In retrospect, aminoacylation reactions with the Hopfield fold RNA library never produced enough desired material to be visible on APB-PAGE which would have been another impeding factor for the ABP-PAGE based library preparation.

Figure 19. Sanger sequencing of the dsDNA library obtained by the APB-PAGE based library preparation approach. **(A)** Alignment of a section of the reads obtained by the five submitted samples showing both randomized regions in the final dsDNA library. Despite some mutations and deletions, the expected sequences could generally be found also beyond the region displayed here. Within the randomized regions, the samples differ the most as was to be expected. **(B)** Electropherogram of sample 4 which in contrast to the other samples seemed to display a mixture of more than one sequence in the randomized regions. This indicated that sample 4 most likely originated from two almost identical dsDNA molecules which differed solely in their randomized region.

3.1.7 Ligation-based library preparation

RNA-RNA ligation

As a first approach to ligation-based library preparation I tried to dimerize the Hopfield fold RNA library through ligation with T4 RNA ligase 1. This experiment needed no further ingredients than were readily available so I could efficiently determine its efficiency. Even though this dimerization made no direct contribution to the planned library preparation, it was nonetheless exemplary for ligations of the 3′ end of the Hopfield fold RNA library with the 5′ end of any other RNA. Also, if such di- or further multimerization could be easily achieved, it could be employed prior to library preparation to increase the number of Hopfield folds being sequenced per read or cluster thereby increasing the cost-effectiveness of the sequencing.

For the actual dimerization experiment, I used a 1:1 ratio of 3′ to 5′ blocked Hopfield fold RNA library. This ensured that only dimers and no further polymers or circular monomers could form. As the library was chemically synthesized, it harbored a hydroxy group at its 5′ end but 5′ phosphoryl groups were necessary for proper ligation. The untreated library could therefore be directly used as the 5′ blocked part. To generate the 3′ blocked part, I first oxidized some Hopfield fold

RNA library with periodate and subsequently 5′ phosphorylated it using T4 polynucleotide kinase under standard conditions. Independent of different incubation times and temperatures or the addition of PEG 8000, no quantitative turnover could be achieved. The combination with highest yield though appeared to be the combination of 25 °C for 5 h with the addition of PEG 8000 (Figure 20). But as even in the latter case still visible amounts of input material were present, I continued testing other methods.

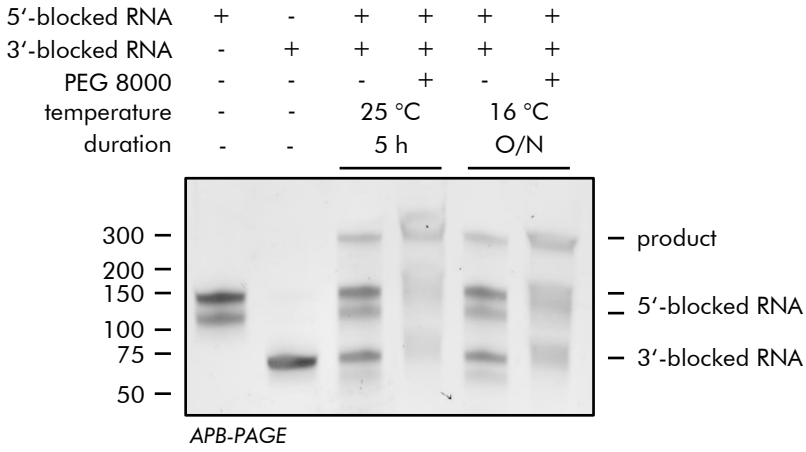


Figure 20. Attempted dimerization of RNA Hopfield folds through RNA-RNA-ligation. 5′ blocked RNA describes untreated RNA Hopfield fold library as it standardly harbored a 5′ OH that could not be ligated. 3′ blocked RNA describes 3′ oxidized and 5′ phosphorylated RNA Hopfield fold library. Standard conditions for T4 RNA ligase 1 mediated ligation were employed. As only poor turnover could be observed, this approach was not further investigated.

Splinted ligation

Next, I tested a ligation approach that was developed by a subgroup of ours [122]. They termed their technique splinted ligation as three strands of DNA or RNA were involved. I adapted this technique to ligate the 3′ end of the Hopfield fold RNA library with the 5′ end of a ssDNA donor strand. The ssDNA splint was designed in such a way so that it hybridized with the terminal regions of the two aforementioned strands to bring their ends that were to be ligated in close proximity to each other. Hybridized like this, T4 DNA ligase was supposed to be able to ligate the Hopfield fold RNA library with the ssDNA donor [123].

Even though no obvious error seemed to have occurred during the setup of the reaction, its yield fell short of expectations (Figure 21). Only two additional faint bads formed that indicated a higher molecular weight than the input material. The main product seemed to possess a molecular weight barely higher than the input DNA donor. Furthermore, the band resulting from the Hopfield fold RNA library did not appear weakened upon ligation. In total, as the turnover of this reaction was apparently very low, I continued exploring other ligation techniques.

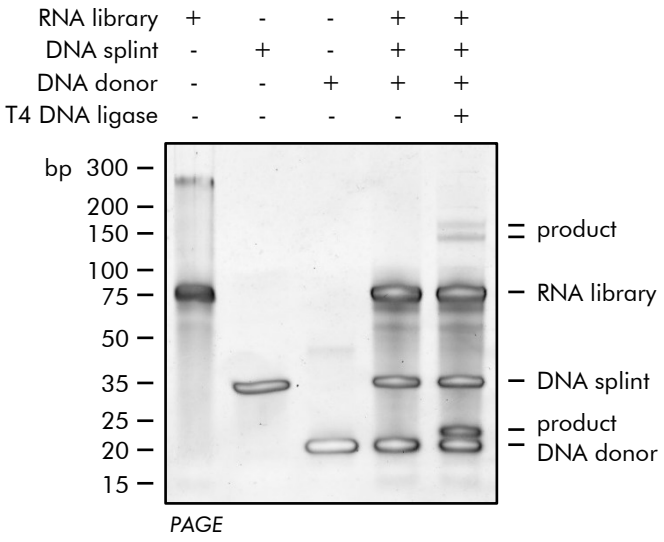


Figure 21. Splint mediated ligation resulting in unacceptably low product yield. However, T4 RNA ligase 1 was falsely used as originally, T4 DNA ligase should have been used instead [122].

Ligation of RNA with 5′-adenylated DNA

As a third option, I started testing a ligation procedure that was previously developed in our own research group in the context of NAD-capture seq [124]. This method made use of two ligases simultaneously namely the T4 RNA ligase 1 which could ligate native 3′ ends of ssRNA to phosphorylated 5′ ends of ssDNA or ssRNA. The other ligase was the T4 RNA ligase 2 truncated K227Q which due to its truncation and mutation was highly specific and effective in the ligation of native 3′ ends of ssRNA with adenylated 5′ ends of ssDNA or ssRNA [123]. The originally published method therefore used chemically synthesized adenosine 5'-phosphorimidazolide (ImpA) to adenylate the 5′ ends of ssDNA strands that were to be ligated [125]; these strands will from here on be called *ligation adaptors*. This

reaction typically resulted in a yield of about 50% leaving half the ligation adaptors with phosphorylated 5′ ends and the other half 5′ adenylated. To inhibit any unwanted ligation of these ligation adaptors' 3′ ends, they were modified with a so called C3-spacer which inhibited ligation with any ligase (Figure 22). To inhibit unwanted ligation of the input RNA's 5′ end, it had to be dephosphorylated which was standardly the case for the chemically synthesized Hopfield fold RNA library.

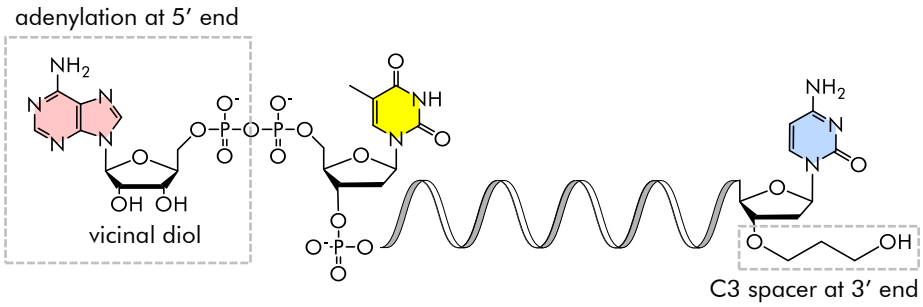


Figure 22. Structure of 5′ adenylated ligation adaptors. The employed T4 RNA ligase truncated KQ could only ligate 5′ preadenylated strands of ssDNA or ssRNA with the native 3′ end of ssRNA. To block any potential side reactions, the 3′ end of the ligation adaptor was modified with a so-called *C3 spacer* which prevented ligation by any ligase. The indicated adenylation was the only occurrence of a vicinal diol within the entire molecule.

In a first attempt, I used this method to ligate the Hopfield fold RNA library with the preadenylated ligation adaptor that was already available for the NAD-capture seq procedure. Even though this ligation adaptor was incompatible with the RT-PCR of the library preparation of this project, I thought it was similar enough to preliminarily investigate this method's overall performance.

Agarose gel electrophoresis after completed ligation revealed the appearance of two bands which I cut out and purified individually. I subsequently treated a part of both of the purified bands with NaOH which should hydrolyze all RNA but no DNA. I could indeed show that the untreated material from the ligation reaction formed bands upon PAGE well above to where the DNA ligation adaptor would have localized. Upon NaOH treatment, the higher bands disappeared and a new band at the height of the ligation adaptor appeared. This indicated that high amounts of the desired RNA-DNA ligation product were formed by the ligation and that the NaOH treatment has hydrolyzed the RNA-part which released the DNA ligation adaptor again (Figure 23).

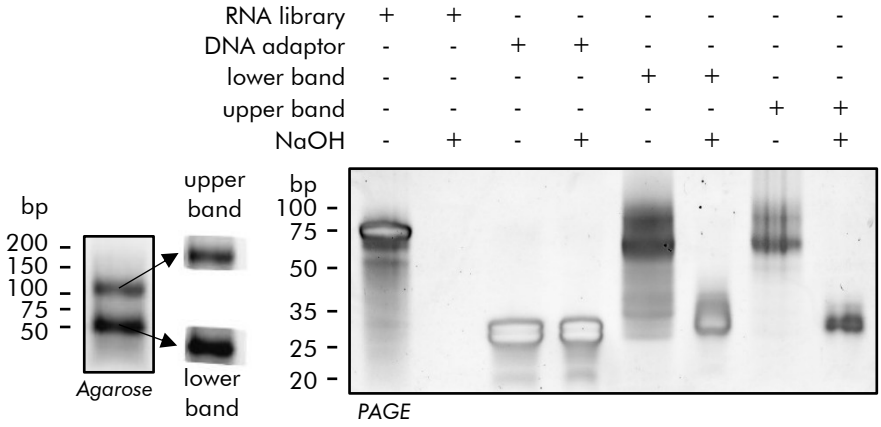


Figure 23. Ligation reaction as described in the in-house developed NAD-capture seq protocol [124]. **(Left)** The reaction was first submitted to agarose gel electrophoresis and the only two resulting bands were cut out and purified through. **(Right)** The input genetic material for this reaction as well as the isolated lower and upper band alongside their products upon treatment with NaOH were submitted to PAGE. As expected, the RNA underwent hydrolysis upon treatment with NaOH whereas the DNA did not. Both the upper and lower band of the ligation reaction seemed to contain no material similar to the adaptor but after hydrolysis a band at a similar height to the adaptor appeared suggesting the isolated ligation product was the desired RNA-DNA hybrid molecule.

As after this preliminary experiment this ligation technique appeared highly promising, I ordered an equally 3′ blocked and 5′ phosphorylated DNA ligation adaptor compatible to the library preparation of this project. I was advised to make the base at the very 5′ end of the ligation adaptor a cytidine as RNA ligases had a preference for this. As several ligation adaptors with randomized stretches of different lengths at their 5′ region were planned to be used, I called this adaptor the C8N adaptor as its 5′ end nucleotide was a C followed by a stretch of eight randomized nucleotides. There was plenty of ImpA still available so I used it corresponding to the established procedure to conduct 5′ adenylation of the C8N adaptor [125]. After the reaction was finished, the adaptor was separated from contaminants like excess ImpA and its hydrolysis products through size exclusion chromatography using NAP-5 columns. As the adaptor was by far the largest molecule in the reaction, it eluted first. The elution could be monitored by the typical absorbance of nucleotides at 260 nm; thereby, also the following peak coming from ImpA and free AMP could be quantified (Figure 24 A). The elution volume coming from the first peak was pooled and analyzed on APB-PAGE which revealed the expected turnover of about 50% as two apparently equally intense bands

appeared. As adenylation incorporated a vicinal diol into the ligation adaptor, which is otherwise free of vicinal diols, both species could be separated from one another (Figure 24 B).

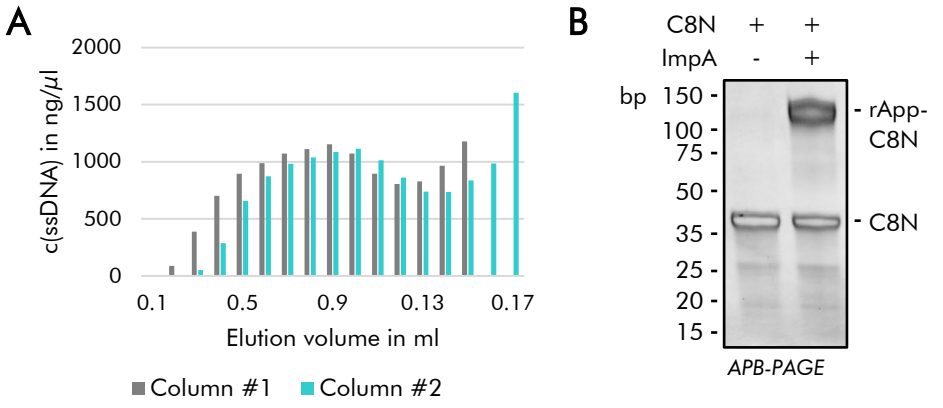


Figure 24. (A) Size-exclusion chromatography on two NAP-5 columns of the DNA ligation adaptor C8N after adenylation with ImpA. Samples from the first peak contained the DNA of interest. **(B)** APB-PAGE of the same DNA before and after adenylation and purification as indicated in (A). The adenylation reaction typically resulted in yields of about 50%.

As the original NAD capture seq protocol handled the RNA attached to beads, the additives DMSO, acetylated BSA, and beta mercaptoethanol had to be added to the ligation reaction. The Hopfield fold RNA library however would always be in solution, so I decided to omit these additives and to rely solely on the T4 RNA ligase truncated KQ combined solely with reactions conditions as indicated by the manufacturer including PEG 8000 as molecular crowding agent to increase yield.

I've set up a corresponding ligation reaction using the Hopfield fold RNA library as well as the previously adenylated C8N ligation adaptor. After the reaction was finished, I analyzed it through PAGE. Even though clearly visible amounts of educts were still present after ligation, an equally intense product band had also formed. I excised the product band along with some other bands for reference, from which I purified the contained material through electroelution. I then treated a part of the purified ligation product with NaOH and analyzed this alongside the untreated ligation product and the other extracted bands through another PAGE (Figure 25 A left). Analogously to the previous NAD capture seq inspired ligation, the ligation product of this ligation was equally sensitive to treatment with NaOH to form a band that corresponded to the ligation adaptor again (Figure 25 A right). These

findings suggested that also this simplified ligation reaction resulted in high yield of the desired RNA-DNA ligation product.

I therefore conducted more such reactions to purify the ligation product through PAGE, excision of the product band and electroelution. I then used this purified ligation product as template in a reverse transcription reaction according to the library preparation procedure. For this reaction, I used both primers with and without the planned overhang. After the reactions were finished, I treated a part of each with NaOH and analyzed these and the crude reaction together with some references through PAGE. The formation of additional bands above the input material could clearly be observed and even though some bands shifted, these elevated bands kept being present upon treatment with NaOH. Also, the bands resulting from the reverse transcription reaction which used primers with overhang were further elevated than those that originated from primers without overhang (Figure 25 B). All this strongly indicated that the desired cDNA was formed.

Lastly, I used the remainders of these reverse transcription reactions as templates for semi-quantitative PCRs. Even though none of the reactions involved were optimized, both of these PCRs showed visible material already after 5 thermocycles, independent of the primers used for the reverse transcription (Figure 26). This point rendered the ligation-based library preparation approach perfectly suitable for this project.

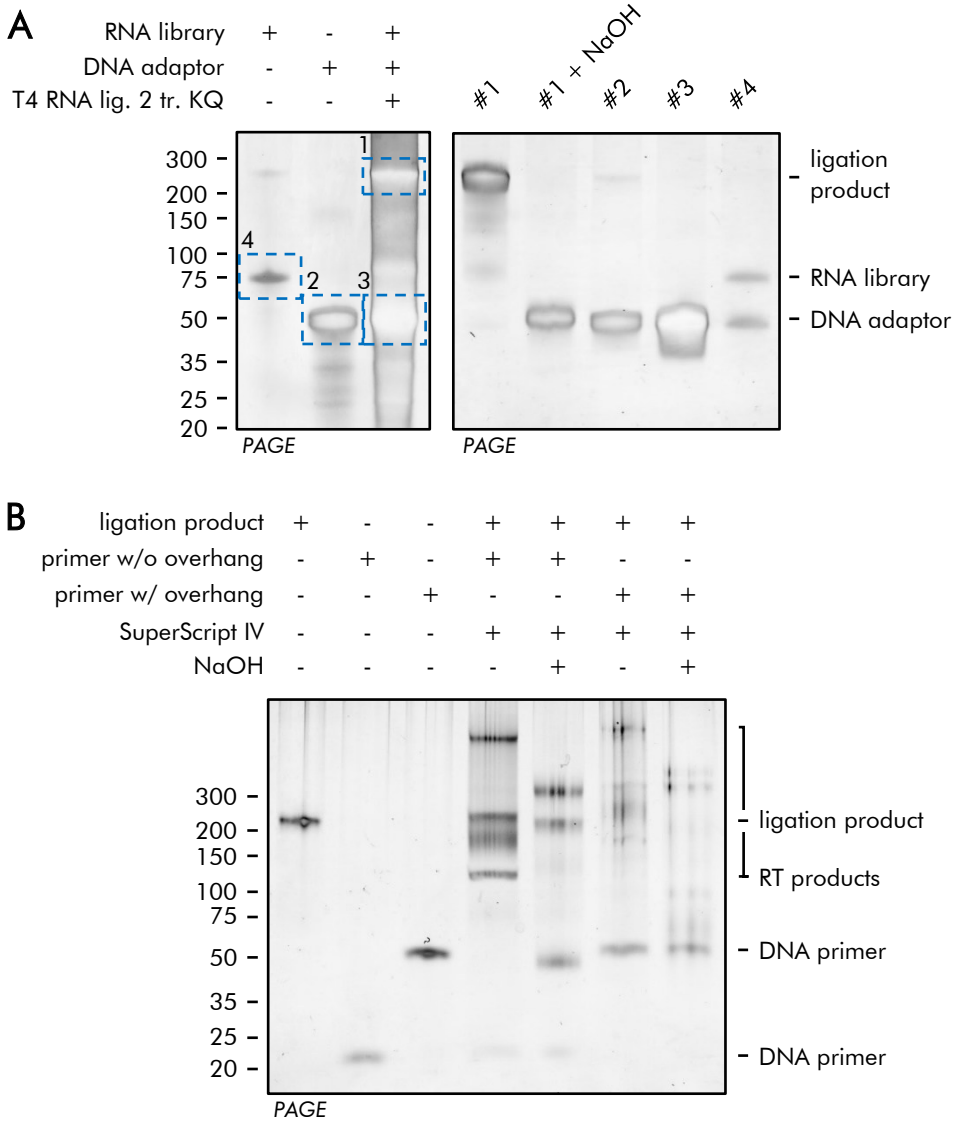


Figure 25. (A) Analysis of ligation reaction. First (left), a ligation reaction was submitted to PAGE and bands of interest were excised and extracted. Second (right), the purified material was again submitted to PAGE including treatment of the ligation product with NaOH. This treatment resulted in material with a similar length than the used DNA ligation adaptor which implied that the ligation product was the desired RNA-DNA hybrid molecule. **(B)** Reverse transcription using purified ligation product similar to #1 in (A) as template. The formation of bands corresponding to a higher molecular weight than the input material could clearly be observed. Elevated bands remained even after treatment with NaOH further indicating towards successful formation of desired cDNA.

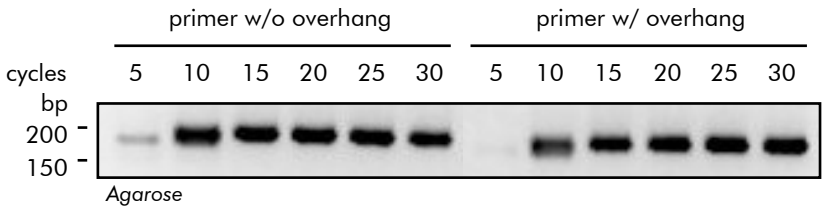


Figure 26. Semi-quantitative PCR using the reverse transcription displayed in Figure 25 B as template. A visible product band formed already after 5 thermocycles. This result appeared independent from the use primers with or without overhang.

So far, only PAGE-purified ligation product were used as template for reverse transcription. In an effort to simplify the library preparation procedure, I tried to use the crude ligation just after PCI extraction and ethanol precipitation. All attempts though caused the semi-quantitative PCR to require about 25 cycles to form the desired product if at all. I suspected that the excess of ligation adaptor, which harbored the binding site for the RT primer, caused problems during the RT reaction resulting in significantly reduced amounts of desired full-length cDNA and increased amounts of truncated cDNA which could serve as template for one of the PCR primers. A study which utilized a similar approach of library preparation employed a combination of 5′ Deadenylase and the 5′ to 3′ exonuclease *RecJf* to digest their adenylated ligation adaptors in between ligation and reverse transcription [126]. Such exonucleases were specific to DNA so the '5-RNA-DNA-3' ligation product was protected from digestion. As 5′ adenylated DNA was protected as well, 5′ Deadenylase had to be employed to hydrolyze the adenylation to reveal a digestible phosphorylated 5′ end.

I tested if RecJf could be used to digest excess ligation adaptor in my case as well. However, the RecJf I was able to obtain seemed to be far less active than the one described in the aforementioned publication as it was unable to digest any visible amount of 5′ phosphorylated ligation adaptor (Figure 27 A). The most plausible explanation for this was that I ordered it from a different supplier. Therefore, I subsequently ordered another 5′ to 3′ exonuclease namely Lambda (λ) Exonuclease which had the same properties as RecJf but turned out to be much more active. I tested the Lambda Exonuclease both in its specialized reaction buffer as well as in ligation buffer and could observe that, depending on the input amount of partly 5′ adenylated ligation adaptor, acceptable amounts of only the 5′ phosphorylated adaptor were digested (Figure 27 A).

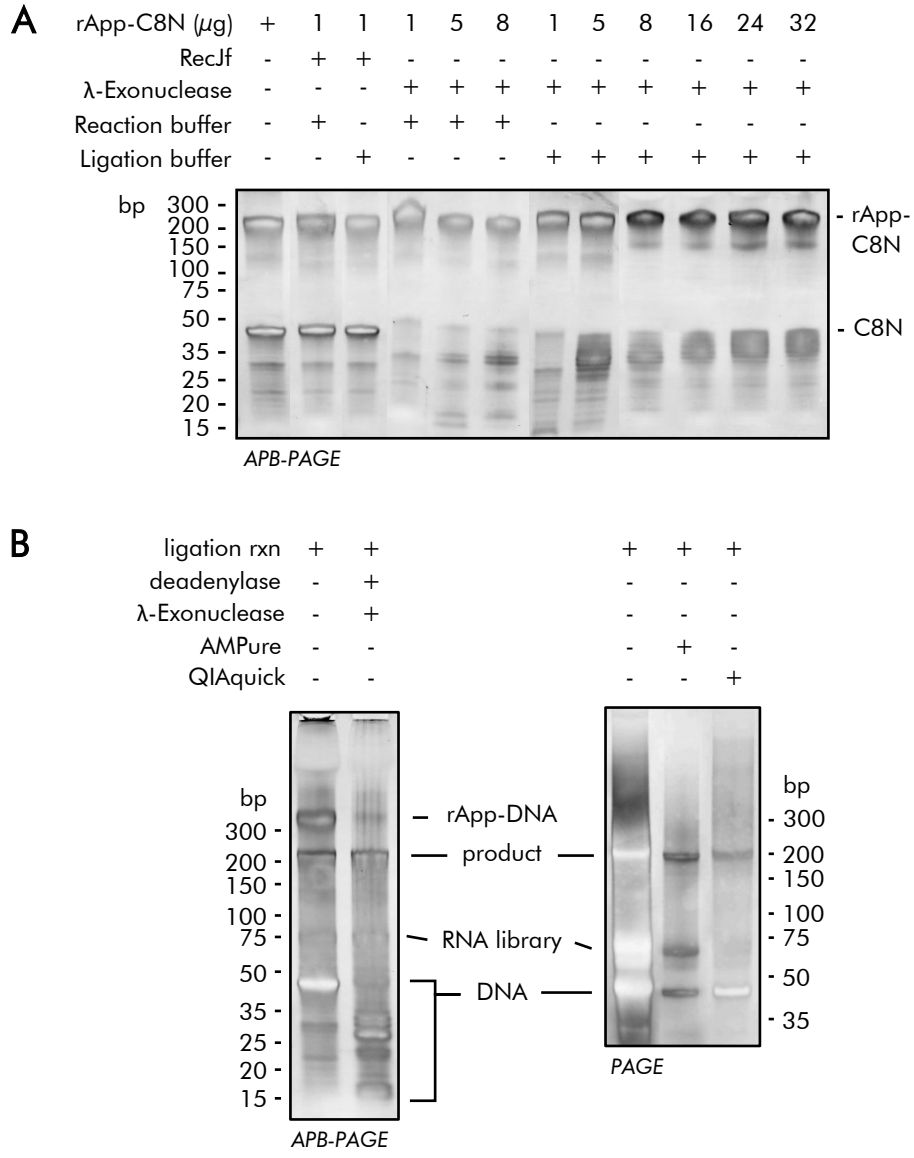


Figure 27. (A) Digestion of partially adenylated C8N ligation adaptor through RecJf and Lambda Exonuclease in either their specialized buffers or ligation reaction buffer. **(B)** (Left) Combined treatment of a finished ligation reaction with 5′ Deadenylase and Lambda Exonuclease to digest all remaining ligation adaptor. (Right) Neither AMPure bead purification nor the QIAquick PCR Purification Kit were able to purify the ligation product from excess ligation adaptor.

I then used both 5′ Deadenylase and Lambda Exonuclease to treat a ligation reaction after it was finished by adding the two enzymes directly to the reaction. I analyzed this ligation before and after the addition of the enzymes through APB-PAGE. I could observe that most of the adenylated ligation adaptor was deadenylated and little full length 5′ phosphorylated adaptor remained but the amount of digestion products still appeared relatively high (Figure 27 B left). I figured that this was the reason why ligation reactions treated with 5′ Deadenylase and Lambda Exonuclease still took around 20 cycles of PCR to form visible product. The loss of only a few bases at the 5′ end of the ligation adaptor could not effectively inhibit hybridization with the RT primer so the digestion had to be conducted more thoroughly.

As it could be conducted without much effort, I also tested if standard PCR purification techniques could purify the ligation product from excess ligation adaptor. However, neither AMPure bead purification nor the spin-column based PCR purification kit by QIAgen (QIAquick) were able to do so as still visible bands corresponding to ligation adaptor formed upon PAGE (Figure 27 B right). These methods alone were therefore not suited to purify the ligation product prior to reverse transcription.

Until this point, I was using the ~1:1 mixture of 5′ adenylated and 5′ phosphorylated ligation adaptor after purification through the NAP-5 columns as input for the ligation reaction. But the 5′ phosphorylated adaptor served no purpose for the ligation reaction and added an additional burden for the exonuclease to digest. I therefore treated some of the ~1:1 mixture with Lambda Exonuclease to digest the 5′ phosphorylated material prior to ligation so that the exonuclease added after ligation only had about half as much material to digest. Over the course of this thesis, I added 4 more ligation adaptors with stretches of randomized nucleotides of different lengths at their 5′ region ranging from 8 to 12 randomized nucleotides. Through sequencing I found no visible preference of a 5′ terminal cytidine for ligation (data not shown) so no other than the initial C8N adaptor carried one. All other adaptors harbored a randomized 5′ terminal nucleotide. Therefore, I named them 8N, 10N, 11N, and 12N with the C8N covering the role of a 9N adaptor. I treated all of them like the C8N adaptor including adenylation using ImpA, purification through NAP-5 columns, digestion with Lambda Exonuclease, and eventual PCI extraction. The resulting input material for ligations showed only traces undigested and partially digested 5′ phosphorylated adaptor and consisted mostly of the desired 5′ adenylated adaptor (Figure 28).

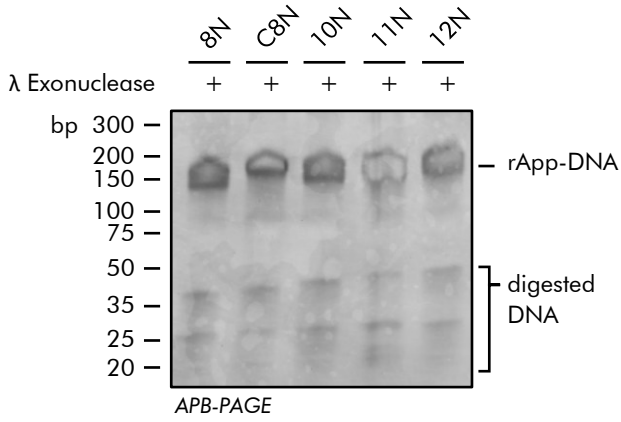


Figure 28. Preparation of pre-adenylated and pre-digested ligation adaptors as used throughout the duration of this thesis.

Generally, RT-PCRs could be set up as a one-step or two-step system. In the one-step approach, RNA input, all primers, reverse transcriptase, and DNA polymerase were put into the same reaction at once buffered by a compromise system that allowed both enzymes to be acceptably active. Specific temperatures first caused the reverse transcriptase to be active followed by the DNA polymerase. In a two-step system, these enzymes act separated from each other in specifically optimized buffers. As already mentioned, I so far used only a part of the crude reverse transcription reaction as template for the PCR. To not lose any material, I opted to use all of the cDNA as template but without intermediate purification. As I was using the reverse transcriptase SuperScript IV for all RT reactions so far, I tested the corresponding one-step RT-PCR system provided by the supplier. However, it turned out to be significantly less efficient compared even to the approach used so far as it took about 10 cycles of PCR to produce visible product bands. I therefore continued testing three two-step approaches. All three started with a standard 20 μl reverse transcription using SuperScript IV and its specialized buffer. I then used the entire 20 μl as template for 50 μl PCR reactions which used buffer and enzyme of the one-step system, of Phusion DNA polymerase, or of Q5 DNA polymerase, respectively. Even though the input of foreign buffer exceeded the recommended 10%, all three reactions produced visible product bands after 5 cycles with the Q5 reaction appearing the most intense one with the least amount of side products (data not shown). I therefore chose this two-step approach with Q5 DNA polymerase as the standard from then on.

While I was developing this library preparation method, I was also conducting aminoacylation experiments with the Hopfield fold RNA library but never obtained visible product bands after oxidative fixation and APB-PAGE. A band consisting of 100 ng of this library has always been easily visible on APB-PAGE so I figured the typical yield of such aminoacylation reactions must be below 10%. I therefore decided to also use correspondingly less ligation adaptor. So far, a two-fold molar excess of ligation adaptor to library has been used i.e. 1.38 μg of adaptor per 1 μg library. This amount was reduced to 100 ng of adaptor per 1 μg library which, due to differences in molecular weight, would be enough to ligate ~14% of the library. To test these adjusted conditions, 10 ng of Hopfield fold RNA library were used as input for three ligation reactions which contained 100 ng, 10 ng, or 1 ng of predigested, adenylated C8N ligation adaptor, respectively. RT-PCR as described above was conducted with 5 cycles of PCR. In parallel, a similar experiment was conducted with 10 ng RNA input and 100 ng ligation adaptor but a reduced number of 2 to 4 cycles of PCR. The low amount of RNA input was meant to simulate 1% of aminoacylation yield to test if the library preparation could handle such low input quantities. Due to the correspondingly low amount of output material to be expected, high sensitivity DNA chips on the Bioanalyzer by Agilent were used for subsequent analysis as these provided higher sensitivity and the data could be quantified more robustly than in the case of PAGE or agarose gel electrophoresis. The samples though had to be purified before they could be analyzed with these chips, so I included purification with AMPure beads as this method was the most efficient and well-established one in the field of library preparation.

It could be shown that 4 cycles of PCR and 100 ng of ligation adaptor were enough to produce sufficient amounts of the desired dsDNA library; using less than 4 cycles was insufficient to do so. Employing 5 cycles produced more material but as long as the amount of dsDNA library was sufficient, I preferred to employ just the minimum number of 4 cycles to reduce the production of duplicates and replication biases as much as possible. Using less than 100 ng of ligation adaptor did not lead to any positive effects (Figure 29). These findings demonstrated that the reduced amount of ligation adaptor made it possible for the 5′ Deadenylase and Lambda Exonuclease to digest as much of the excess adaptor to not negatively interfere with the subsequent RT-PCR.

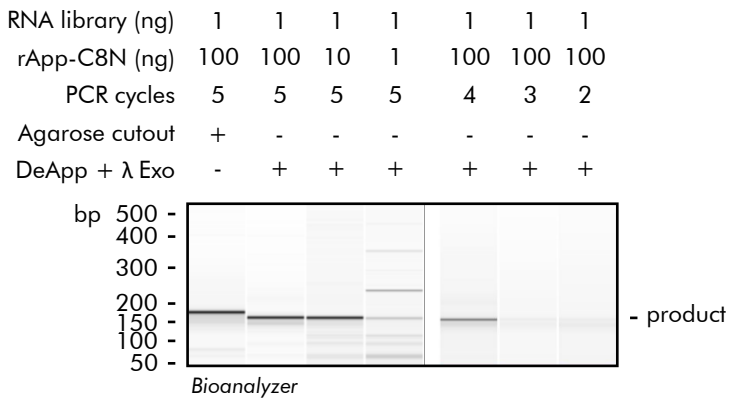


Figure 29. Final optimization of the ligation-based library preparation procedure. Its efficiency could be elevated so far as 10 ng Hopfield fold RNA library, 100 ng of adenylated DNA donor as well as 4 cycles of PCR were sufficient to generate a visible and quantifiable product band in the readout of Bioanalyzer High Sensitivity DNA Analysis. Under the supervision of the author of this thesis, Philip Höflich was involved in the preparation of the experiments that led to these results [127].

At this point, I concluded the development of the ligation-based library preparation approach as the method at hand provided a minimum number of purification steps and used 100% of the input material throughout the entire procedure. The use of minimally labor-intensive purification methods rendered the method suitable for a high throughput of samples. The method proved highly reliable and the only change I made during the course of this thesis was to omit the tedious PCI purification of the ligation reaction after treatment with 5′ Deadenylase and Lambda Exonuclease and to only conduct ethanol precipitation. This change did not seem to influence the procedure in any way except for making it more labor efficient.

APB-PAGE kept being too insensitive to measure the yield of aminoacylation reactions of the Hopfield fold RNA library directly so the only way to assess the yield was through this library preparation procedure. However, due to the multitude of steps involved between input material and output reading on the Bioanalyzer, the results represented aminoacylation yield indirectly and included the variation of the intermediate steps. Nonetheless, the repeated emergence of patterns could be observed in the generated electropherograms. The signal peaks at 35 bp and 10380 bp were to be ignored as these were coming from internal standards introduced by the high sensitivity DNA chip sample preparation. Depending on the used ligation adaptors, the desired dsDNA library produced a signal at around

170 bp. This peak could readily be quantified and served as the main indicator of the quality of a sample (Figure 30).

I established two different positive controls; *type 1* consisted of 990 ng of oxidized Hopfield fold RNA library mixed with 10 ng of native library, *type 2* consisted of only 10 ng library. In both cases, this material was directly used for library preparation starting with ligation. Both of these positive controls were prepared in triplicates and no major differences within each set of triplicates could be observed. Positive controls of type 1 resembled aminoacylation reactions with 1% yield. The main reason to set up the type 2 positive controls was to see if the oxidized RNA caused interferences of any kind. Indeed, aminoacylations conducted with adenylates derived from Lys, Ala, Gly, Val, or Ile resulted in outcomes similar to the type 1 positive control or exceeded in concentration of material at ~170 bp (Figure 30).

As negative control, I treated some L-Lys-AMP, L-Ala-AMP, and Gly-AMP with dilute ammonia to cause their hydrolysis into free amino acids and AMP. The ammonia could be completely removed through lyophilization. I then conducted aminoacylation reactions with this hydrolyzed material followed by library preparation which resulted in no visible material at 170 bp. The signal at around 140 bp was rather inconsistent but in most cases below the one at 170 bp. My intern Suria Moralez Guzmán and I suspected that the signal at 140 bp resulted from RNA degradation but the inclusion of RNAse inhibitor had no effect [98]. The peaks between 35 bp and 140 bp most likely came from residual primers or unspecific side products during PCR; their identity was not further elucidated though. Prior to sequencing, I selected samples with a good ratio between 140 and ~170 bp as well as high amounts of ~170 bp material in general. I used the quantified value of the ~170 bp peak to mix all desired samples in a 1:1:1… ratio. If this multiplexed library exceeded 20 µl in volume, it was lyophilized, dissolved in 10 µl nuclease-free water again and submitted to low-melt agarose gel electrophoresis. I excised the resulting band at ~170 bp and purified the material through digestion with agarase. Thereby, only the desired dsDNA material free from any genetic contamination could be obtained (Figure 30). This purified, multiplexed library was diluted according to Illumina's instructions depending on the chosen sequencing platform and could subsequently be used for sequencing.

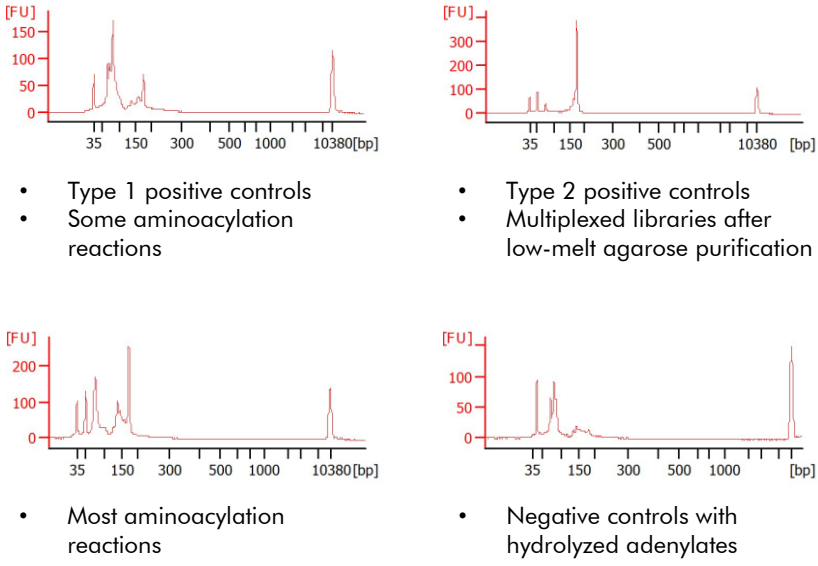


Figure 30. Exemplary electropherograms related to ligation-based library preparation. In all cases, the peaks at 35 bp and 10380 bp resemble internal markers that were added to every sample during sample preparation; all other peaks originated from the analyzed sample. The peaks at around 170 bp correspond to the desired final dsDNA library. The identity of the peaks between 35 and 170 bp could not be finally elucidated but primer dimers and other secondary RT-PCR products seemed likely. Below each electropherogram, corresponding samples were listed.

3.1.8 Sequencing

<u>Technical quality control</u>

In total, I conducted five sequencing runs. The first two were done on an iSeq100 that was available to our research group and the third one was done on the MiSeq of the sequencing core facility of Heidelberg University. They mainly served the purpose of validating the library preparation technique I developed and proved that both sequencing reads as well as both indices could be sequenced as intended. However, these early sequencing runs did not contribute reliable data but primarily served method development. The fourth sequencing run was done on the NextSeq 550 of the core facility and produced the main data of this thesis as 80 samples were sequenced in parallel. During her internship, Mara Behnke assisted me in the preparation of most of these samples [128]. Lastly, another run was

conducted on the iSeq 100 with 4 samples derived from aminoacylation with L-Lys-CMP. During his internship, Niko Jakob assisted me in all steps that were involved in the realization of this last sequencing run [129]. Only in the case of the iSeq 100 runs, I had access to all raw and metadata which was generally not provided by the core facility. I used the Illumina Sequencing Analysis Viewer to access the metadata and to assess the quality of the sequencing run. Clusters passing filter and Q30 values were within acceptable parameters.

Concordant to the sequenced DNA library, the complexity changed over the course of the sequencing run. At the beginning of read1 (R1), the randomized nucleotides of the ligation adaptor were sequenced leading to a relatively constant ratio of each nucleotide (Figure 31). Afterwards, every read resulted from the same sequence, which was in equal proportions frameshifted by 0, 1, 2 or 3 nucleotides. This was the less complex part of the library and resulted in up to 75% of the clusters showing the same base at the same time. At around cycle 35, the ratio shortly normalized as all clusters reached the randomized area of the Hopfield fold's anticodon loop. Even though this area had a length of 7 randomized nucleotides, the ratio was not stable for 7 cycles due to the aforementioned frameshift. Afterwards, the complexity of the library remained low until the read was finished. The indices also consisted of just 4 different sequences leading to low complexity as well. Read2 (R4) started at the randomized region of the Hopfield fold's anticodon loop and went on without frameshift. This was the reason why both the maximum ratio of some bases was much higher compared to read1 (R1) and the complexity increased from cycle 102 to 110 as at this position, the randomized nucleotides of the ligation adaptors were sequenced (Figure 31). This data strongly indicated that the library design and preparation functioned as intended.

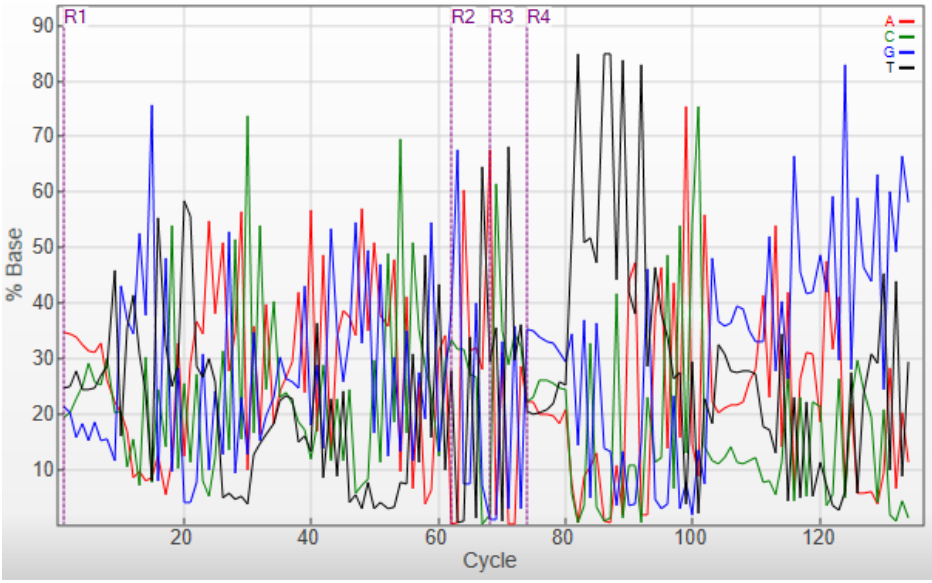


Figure 31. Base distribution per cycle as displayed by the Illumina Sequencing Analysis Viewer. This data corresponded to the first iSeq100 run and depicts the relative ratio of each base during every cycle of the entire sequencing procedure. The iSeq100 started with read1 (R1) followed by sequencing the indices i7 and i5 (R2 or R3, respectively) and eventually ended with read2 (R4). Expected high complexity could be observed in the beginning and middle of read1 and read2.

Data conversion

To further analyze the sequencing results, I first searched for available tools online. A multitude of different software was available to trim and align sequencing data to a reference genome but as the research question of this thesis was different to usual applications of high throughput sequencing, I could not find tools that suited this specific purpose. I therefore solely relied on Microsoft Excel and iteratively developed custom scripts in its programming language Visual Basic for Applications, short *VBA*. I chose this approach because I was already highly familiar with both Excel and VBA but other programming languages would have been equally suited. My efforts resulted in a single Excel file that contained all relevant data from all successful aminoacylation reactions followed by library preparation. Most importantly, I collected the yields of dsDNA library according to bioanalyzer as well as the used ligation adaptors and indices.

The following explains the multi-step process that I developed to convert the data obtained by the sequencer into an analyzable format (Figure 32).

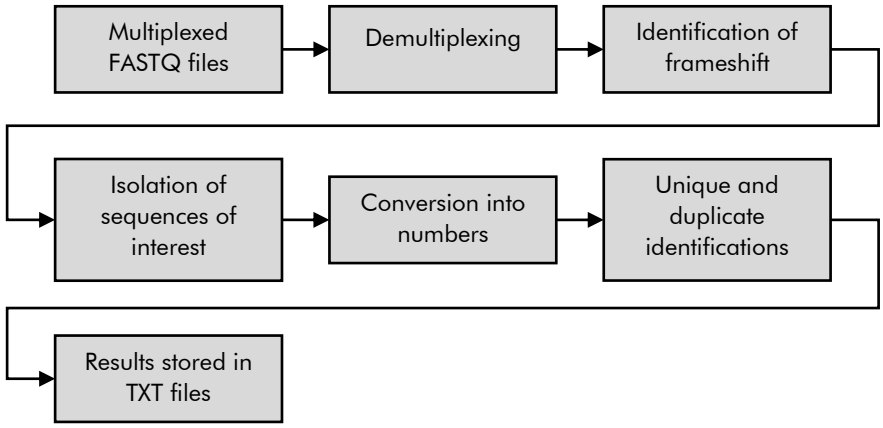


Figure 32. Overview of the data conversion process. All details are given in the main text.

Illumina used the FASTQ format to store the sequencing data. It contained data about the cluster, the sequence, and quality scores of all reads that passed the internal filters. The FASTQ format is completely text based and could be read and modified by standard tools like the MS Editor. I received the sequencing data as compressed .fastq.gz filles which could be decompressed using the freeware WinRAR. On purpose, the results were also in multiplexed form. Demultiplexing could already be performed by the sequencers themselves but as I multiplexed up to 80 samples, I would have had to type in all sample names and indices manually or communicate the core facility to do so. As this would have been comparably cumbersome and error prone, I decided to create a custom script for demultiplexing that could directly access the tabular data that stored all this information anyways. The script functioned by comparing each read with the indices of all samples; when both indices matched, the read was copied into another FASTQ file generating one new file for every sample. During the comparison of the indices, I allowed for the mismatch of one nucleotide for each index. Whenever more than one nucleotide did not match, the read was copied into a FASTQ file collecting all undetermined reads.

As the decompressed FASTQ files exceeded sizes of several gigabytes, they could not be reliably loaded into RAM. To circumvent this limitation, I utilized the method of text streaming in the scripts I developed. This caused the successive copying of singular lines at a time from the file on the hard drive into RAM. Thereby, files of any size could be effectively handled, limited only by the available storage capacity of the hard drive.

I chose to conduct a single-end sequencing run on the NextSeq 550 for cost reasons as a paired-end run was considerably more expensive but would not have yielded data of higher quality. Therefore, I designed the following procedures to only focus on read 1. Upon sequencing the final dsDNA library, the stretch of 8 to 12 random nucleotides coming from the ligation adaptors were the first nucleotides to be sequenced. The following nucleotides were derived from the 3′ region of the Hopfield fold. I recognized that this region was prone to degradation; a considerable fraction of Hopfield folds seemed to be degraded by one or more nucleotides in 3′ to 5′ direction. A small fraction seemed to be longer than expected. This region that was derived from the CCA-end as well as the stem-loop was followed by the randomized anticodon loop sequence followed by the sequencing primer binding site for read 2. As this binding site was constant, it could be used to identify by how many nucleotides the read deviated from what was expected (Figure 33). All reads should show 30 + N, nucleotides until the sequencing primer binding site 2 appeared, with N corresponding to the 8 to 12 nucleotides from the DNA ligation adaptor. I designed the corresponding script to identify at which position this binding site was located and designated its difference from 30 + N as *frameshift*.

I decided to include not all possible frameshifts in the analysis. A frameshift of more than -23 led to losses of the anticodon loop. Therefore, I specified the lower limit of frameshifts to be recognized to -23. As an upper limit I decided to include frameshifts of up to +20 even though I did not expect many hits in this regard. In total, this led to 44 different frameshifts being recognized.

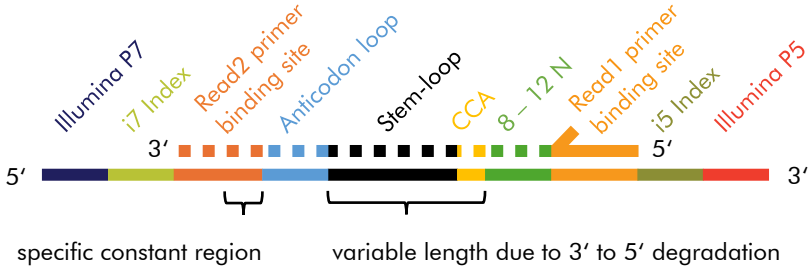


Figure 33. Acquisition of read 1 during sequencing. The 3′ end of RNA Hopfield folds was prone to degradation therefore, losses of material occurred in the CCA (yellow) and stem-loop (black) region. The read length was long enough to reach up to the constant read 2 sequencing primer binding site. In between the two sequencing primer binding sites, 30 + N nucleotides were expected to be with N corresponding to the 8 to 12 nucleotides coming from the ligation adaptor (green). Deviations from these expected 30 + N nucleotides were accordingly treated as frameshifts.

After the frameshift was identified, the precise position of both randomized sequences within the read were known. The randomized stretch coming from the DNA ligation adaptor made up the first 8 to 12 base calls and the sequence coming from the randomized anticodon loop was adjacent to the read 2 sequencing primer binding site (Figure 33). Due to the structure of the dsDNA library, read 1 displayed the reverse complement of these sequences which was compensated by one of the scripts. From here on, these sequences are referred to as the *sequences of interest*.

Storing numbers took up much less storage than to store strings of letters. Therefore, I introduced the conversion of the sequences of interest into numbers based on a quaternary numeral system in which A = 0, C = 1, G = 2, U = 3 (Table 4). Thereby, the sequence AAAAAAA corresponded to 0 and the sequence UUUUUUU to 16,384; sorting the sequences by their corresponding numbers was thereby identical to sorting them alphabetically.

value of position	4^6 = 4096	4^5 = 1024	4^4 = 256	4^3 = 64	4^2 = 16	4^1 = 4	4^0 = 1
letter	C	G	U	A	C	G	U
corresponding multiplicator	1	2	3	0	1	2	3
intermediate result	4096*1 = 4096	1024*2 = 2048	256*3 = 768	64*0 = 0	16*1 = 16	4*2 =8	1*3 = 3
total	6939						
to be stored in line	6939+1 = 6940						

Table 4. Exemplary transformation of the sequence CGUACGU representing a quaternary number into a decimal number with A = 0, C = 1, G = 2, U = 3. To determine the corresponding line of this sequence, +1 had to be added to the decimal number as AAAAAAA = 0 had to be stored in line 1.

Next, I intended to identify PCR duplicates which were combinations of the same anticodon loop sequence and the sequence of the randomized stretch of the ligation adaptor found more than once. I therefore introduced the corresponding designations *unique identifications* and *duplicate identifications*. To process all reads of a sample this way, I introduced an array of the data type *long* which could store numbers ranging up to at least two billion. The array was set to have the dimensions 16,385 * 40,000; for simplicity reasons, the first dimension will from here on be called *lines* and the second dimension *columns*. The number of lines corresponded to the possible numbers, the seven nucleotides of the anticodon loop

could be converted into. I added a +1 to this number to determine which line corresponded to this sequence. The first column stored the number of duplicate identifications which was independent from frameshift and the following 44 columns stored the number of unique identifications depending on their frameshift. The remaining columns to the right of that were supposed to hold the numbers the randomized stretch of the ligation adaptor was converted to. For each read, the corresponding line was determined and checked if in the corresponding columns, the randomized stretch of the ligation adaptor was found before. If so, +1 was added to duplicate identifications; if not, +1 was added to the unique identification with corresponding frameshift and the number corresponding to the adaptor's random stretch was stored in the leftmost free column of the same line.

Previously, I attempted this analysis the same way within an Excel sheet, but the process was slow and prone to crash. Using an array made the process considerably faster and more stable.

After all reads of a sample were processed, the data within the first 45 columns of the array was transferred into corresponding 45 TXT files; all other data was discarded. Thereby, these files had an analogous structure of 16,385 lines with each line corresponding to another possible sequence of the anticodon loop and the displayed number showed either unique or duplicate identifications of this sequence. These files were stored in a folder named by their original sample (Figure 34).

To access this data again, I created a script that initialized an array of the data type long with the dimensions 16384 * 45 * 80 to store each sequence's unique or duplicate identification * all frameshifts * all samples. The script then accessed all corresponding files in the corresponding order and copied the information contained in the TXT files into this array. As variables and arrays were typically stored in RAM, all data was at this point readily available. To browse the data, I created an Excel worksheet that allowed for the selection of the desired sample and frameshifts. A custom script then copied the desired data from the array into the worksheet, which was the overall fastest, most stable and most intuitive way to browse the data. All the following analyses could then be conducted using standard Excel table calculation.

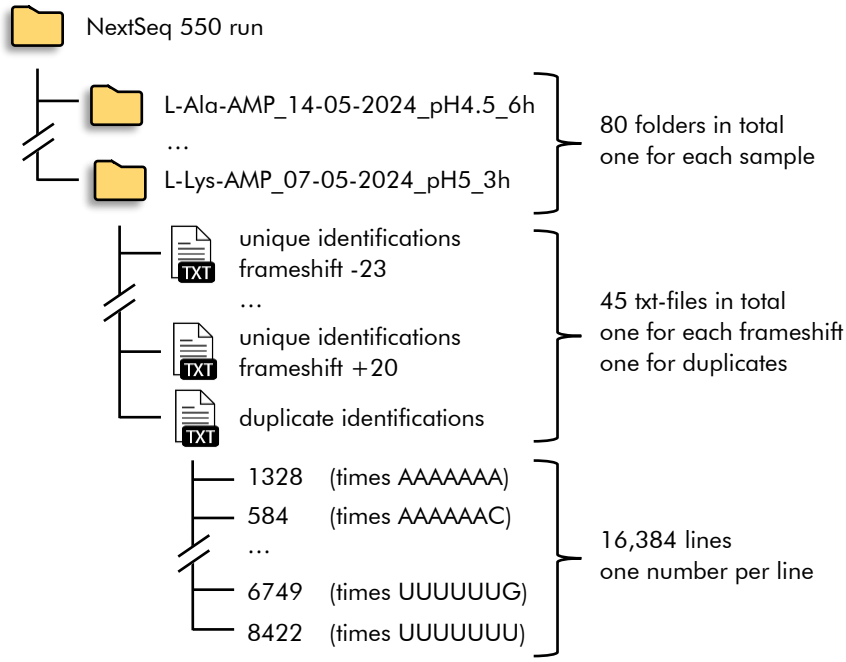


Figure 34. Structure of the sequencing data obtained by the NextSeq 550 run after analysis. As in this sequencing run 80 samples were sequenced, also 80 folders were created each one storing the data corresponding to one sample. Each of these 80 folders contained 45 txt-files; 44 of which stored the unique identifications acquired for each frameshift from -23 to +20. The 45th file stored all duplicate identifications independent of frameshifts. All of these txt-files contained $4^7 + 1 = 16{,}385$ lines and each line corresponded to one possible sequence of NNNNNNN from AAAAAAA to UUUUUUU in alphabetical order. In each line, a number was stored corresponding to the number of unique or duplicate identifications found for the sequence corresponding to each line.

To verify that all the scripts worked as intended, I manually checked for a limited number of reads if they were correctly demultiplexed, the frameshift correctly identified, and the found anticodon sequences stored at the right position. I also manually generated some FASTQ data with known features and verified that it generated the expected results.

The unique identifications of each sample within a multiplexed library was desired to be equally distributed. Therefore, I mixed the DNA libraries of interest in equimolar ratio based on the readings of the Bioanalyzer (Figure 30). After analyzing the sequencing data as described above, I could observe that this strategy worked as intended. Most of the samples sequenced in the NextSeq 550 run

generated roughly the expected number of unique identifications (Figure 35 A left). The most underrepresented samples were mostly derived from experiments under non-ideal conditions or negative controls; it was unsurprising that these samples generated less usable data than the other samples which were conducted mostly under ideal conditions. The most overrepresented samples were positive controls. Similar results were obtained from the iSeq100 run with the four samples derived from aminoacylations with L-Lys-CMP. However, the number of reads passing filter was lower than expected which may have been caused by the fact that the used cartridge which contained all components for sequencing was almost one year past its expiration date or due to overloading. Normally, an iSeq100 run would generate around 6 Mio usable reads but in this case, only 2 Mio unique identifications could be extracted. Nonetheless, all four samples ranged closely around 500,000 unique identifications each (Figure 35 A right).

In summary, the NextSeq 550 run including 80 samples which resulted in 290 Mio reads whose indices matched a sample (100%). In 8.5%, the specific sequence could not be found and were therefore excluded. 3% comprised duplicate identifications. 81% comprised unique identifications within acceptable frameshift and 58% without any frameshift (Figure 35 B).

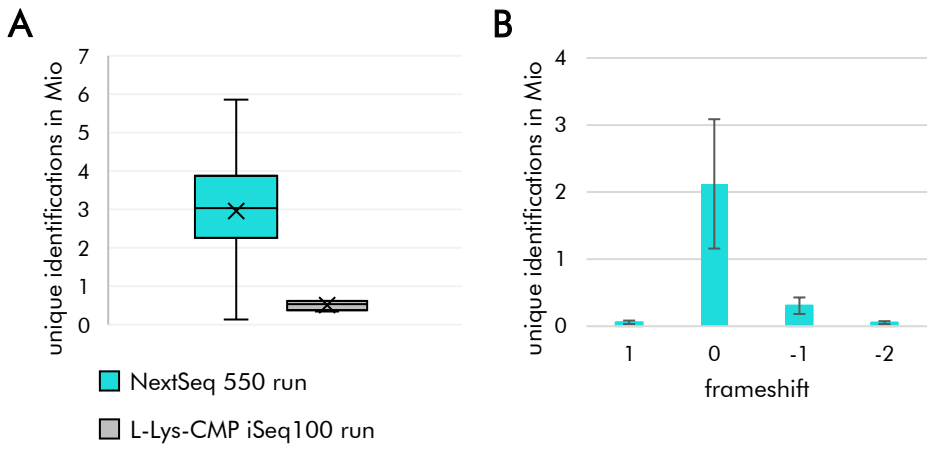


Figure 35. **(A)** Distribution of unique identifications found per sample within the two relevant sequencing runs. The NextSeq 550 run (left) was composed of 80 multiplexed samples and the iSeq 100 run of 4 samples. For both runs, an average of 1.6 Mio reads per sample was expected. **(B)** Distribution of frameshifts within the unique identifications of all 80 samples of the NextSeq 550 run. The error bars indicate standard deviation.

3.2 Testing Hopfield's hypothesis

3.2.1 Comparability of samples

At this point, the tools I created basically counted how many times each of the 16,384 different anticodon loop sequences could be found in each sample. However, due to inaccuracies in quantification or multiplexing, the sum of unique identifications differed between each sample. To account for that, I calculated the *relative abundance* of a sequence within a sample by dividing its unique identifications by the sum of unique identifications of all sequences in the sample (Figure 36, first formula).

The unique identifications and relative abundances were slightly different between the sequences in all triplicates of the type 1 positive control (i.e. 10 ng native HF RNA library + 990 ng oxidized HF RNA library). To account for this uneven pool of starting material, I calculated *enrichment factors* by dividing the relative abundance of a sequence in a sample by the average relative abundance of the same sequence from the triplicates of the type 1 positive controls (Figure 36, second formula).

As the aminoacylation experiments conducted with L-Lys-AMP, L-Ala-AMP, and Gly-AMP were the most comparable ones due to their similar reaction conditions and incubation times of around 3 to 6 h, I continued focusing solely on samples derived from these three adenylates and neglected the samples derived from L-Val-AMP and L-Ile-AMP as their incubation times of around 24 to 48 h respectively were much longer. To calculate the *percentual specificity* of a sequence for one of these first three adenylates, I divided its relative abundance in the corresponding sample by the sum of relative abundances found in all three samples. Average relative abundances of replicate samples could also be used (Figure 36, third formula). Using enrichment factors instead of relative abundances would not have made a difference as the division of all mentioned data points by the same relative abundances of the type 1 positive controls would have canceled itself.

To advance data analysis, I created a script which enabled fast comparisons between an arbitrary selection of possible anticodon loop sequences in an independently chosen set of samples. I therefore named this script *comparator* and I used it to check if replicate samples generated similar data (Figure 37). Most of them did but some differences came from using different batches of adenylate, followed by conducting similar experiments on different days, followed by apparently random fluctuations. Thereafter, I included only samples derived from

the same batch of adenylate and excluded the most striking outliers compared to the majority of samples.

$$A_{k,p} = \frac{R_{k,p}}{\sum_{k=1}^{16384} R_{k,p}}$$

$$E_{k,p} = \frac{A_{k,p}}{\frac{1}{3}\sum_{c=1}^{3} A_{k,c}}$$

$$S_{k,p} = 100 * \frac{A_{k,p}}{A_{k,Lys}+A_{k,Ala}+A_{k,Gly}}$$

R = read count

A = relative abundance

E = enrichment

S = specificity (in percent)

k = sequence (AAAAAAA to UUUUUUU)

p = sample (e.g. L-Lys-AMP_pH5_3h)

c = type 1 positive control

Figure 36. Formulas to calculate central parameters of sequencing data analysis. In the case of (percentual) specificity, exclusively the data generated by aminoacylations with L-Lys-AMP, L-Ala-AMP, and Gly-AMP or were considered.

3.2.2 Identification of top performing sequences

After selecting the most reliable samples derived from aminoacylations with L-Lys-AMP, L-Ala-AMP, and Gly-AMP, I used the comparator alongside regular Excel table calculation to identify the anticodon loop sequences that were both the most specific and most enriched ones for each adenylate. I named them the *L-Lys-specific*, *L-Ala-specific*, and *Gly-specific sequence*, respectively. It is to be especially highlighted that their specificity was related to this subset of three adenylates. Their behavior towards any other activated amino acid was not known up until this point. E.g. it could later be shown that the so-called L-Ala-specific sequence actually reacted more strongly with D-Lys-AMP than with L-Ala-AMP (Figure 43). Besides these three specific sequences, I could also identify another sequence that was among the most strongly enriched sequences in all samples. Due to its high activity and low specificity, I called this sequence the *universal acceptor*. From here on, these four sequences may be simply referred to as the *selected sequences* (Figure 37).

		L-Lys-AMP		L-Ala-AMP		Gly-AMP	
		3 h	6 h	3 h	6 h	3 h	6 h
Universal acceptor	UCAUGAG	79.8	74.1	30.0	32.1	49.3	51.2
L-Lys-AMP specific	UCAUGCA	36.0	33.9	4.5	5.0	7.4	8.4
L-Ala-AMP specific	GCACCCA	1.2	1.3	20.2	19.6	1.7	2.1
Gly-AMP specific	CAUCGCG	27.3	27.7	46.3	41.8	87.7	79.4

Figure 37. Exemplary results generated by the automated comparator script. As input, it accepted any number of possible anticodon loop sequences (y-axis) and samples (x-axis). As output, it displayed enrichment factors based on the triplicate type 1 positive control. This example illustrates the differences between various adenylates and incubation times used during aminoacylation for the four specifically selected sequences.

All these sequences were highly and significantly enriched upon reaction with their corresponding amino acid compared to the triplicates of the type 1 positive controls. It could also be shown that type 1 and type 2 positive controls differed in some regards, yet not as much as type 1 positive controls compared to any aminoacylation sample (Figure 38).

The sequencing data from aminoacylation experiments using L-Lys-CMP were excluded from the selection of specific sequences because these experiments were conducted at a later stage of this thesis. Retrospectively, the inclusion of this data would not have influenced the selection, as I demonstrated that none of the selected sequences exhibited significant reactivity with the cytidylate (Figure 43).

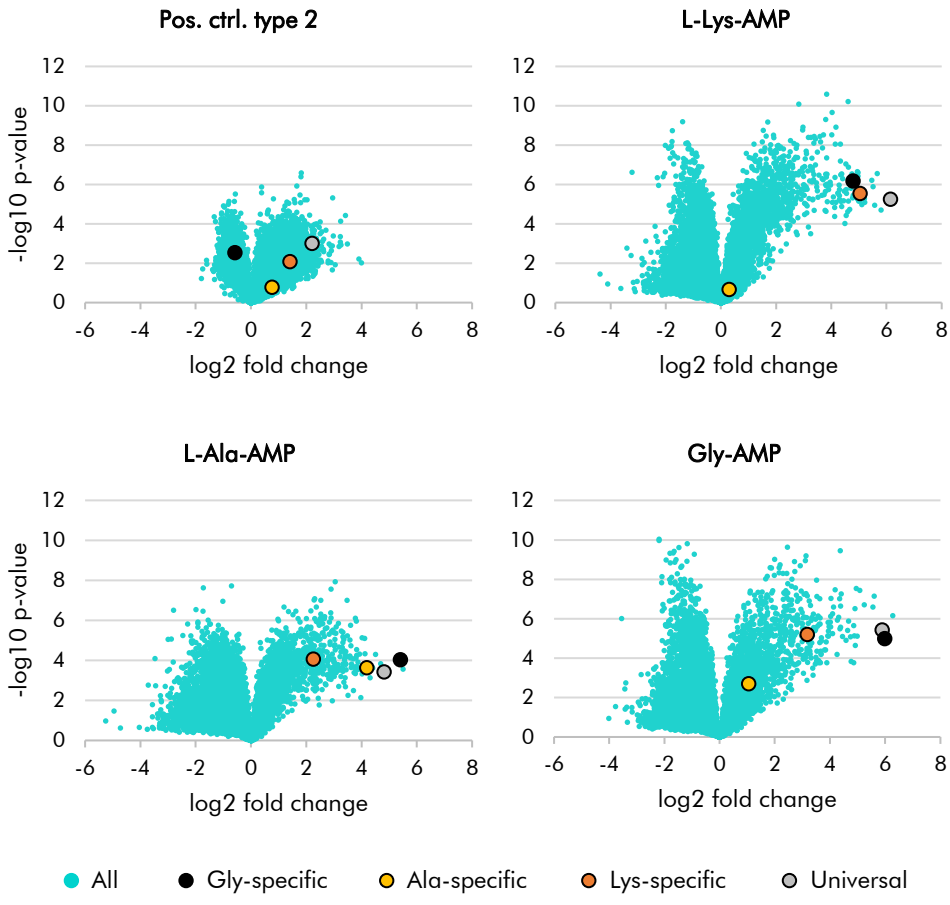


Figure 38. Volcano plots depicting the logarithm on base 2 of the fold change of the replicates of the aminoacylation experiments or type 2 positive controls in comparison to the triplicates of the type 1 positive controls (x-axis). The negative logarithm on base 10 of the p-value (y-axis) was calculated through two-sided heteroskedastic t-tests. Four selected anticodon loop sequences were highlighted, see main text for details.

3.2.3 Elucidating the binding mechanism

<u>Compilation of single mutants</u>

To elucidate the binding mechanism of a sequence of interest, I compiled the enrichment factor of this sequence alongside the enrichment factors of all its possible single mutants in a grid like manner. The composition of hydrogen bond

donors and acceptors on the Watson-Crick edge of each nucleotide was also included. Thereby, the influence of each nucleotide at each position could be assessed. This was done with the four selected top performing sequences as well as the most enriched sequence derived from experiments with L-Lys-CMP (Figure 39 A).

Through this approach, it could be observed that each position of the anticodon loop was differently involved in the realization of self-aminoacylation and exhibited various tolerances towards mutation. If a sequence harbored a G at its fifth position (N5), no other nucleotide could be situated there without reducing the sequence's activity to a minimum. Only the L-Ala-specific sequence harbored a C at this position, but this sequence was by far the least reactive of the four selected sequences. Sequencing data suggested that changing this C to a G would increase reactivity but would equally reduce the specificity of this sequence. In the case of the Universal Acceptor and the L-Lys-AMP specific sequence, the nucleotide in the first position (N1) of the anticodon's sequence had almost no impact on its reactivity and exhibited only a slight preference for U. Regarding aminoacylations with L-Lys-AMP and L-Lys-CMP, the sequences that harbored a purine base at N7 were more active than those harboring a pyrimidine. In the case of the Gly-specific sequence, a nucleobase that harbored a hydrogen bond donor on the first position of its Watson-Crick edge and a hydrogen bond acceptor at the second site seemed to be necessary to properly function; the third site appeared to be irrelevant as well as if the nucleobase was a purine or pyrimidine base (Figure 39 A).

In general, it appeared that the further a randomized nucleotide was positioned away from the 3′ end, the less influence it exerted on the aminoacylation behavior. According to Hopfield's original postulation, the anticodon would be represented by N2, N3, and N4, with N4 decoding the first letter of a codon, N3 the middle one, and N2 the last one. In these cases, the observation that increasing distance from the 3′ end diminished the influence of a nucleotide on aminoacylation behavior aligned with the broader finding that the first two nucleotides of a codon are most decisive in determining which amino acid is encoded. Hopfield's hypothesis did not originally account for insertions or deletions of the Hopfield fold during its evolution into extant tRNA. However, a minor shift of one or two nucleotides would have been sufficient to relocate the anticodon to the positions most critical for the activity of the Hopfield fold, as tested herein.

Originally, Hopfield did not consider the coding potential of the acceptor stem. His hypothesis was focused on the single stranded overhang near the 3′ end. However, these findings revealed a significant influence of the double stranded region near the 3′ end on aminoacylation activity as well.

The found sequences did not contain extant anticodons neither at the expected positions N2, N3, and N4 nor consistently in other locations. The extant code uses CUU and UUU as anticodons for L-Lysine, NGC for L-Alanine, and NCC for Glycine. Maybe in the context of specificity in a broader set of different adenylates than utilized here, extant anticodons may emerge but in the most reactive and specific subset identified here, no resemblance could be observed.

In how far the topology of each nucleotide was involved in the realization of self-aminoacylation could not be elucidated with certainty. Patterns may be interpreted into the sequencing data (Figure 39 A) but to reliably investigate the involvement of each hydrogen donor and acceptor as well as purine and pyrimidine bases at every position, non-canonical nucleotides that differ from the canonical ones in singular sites would have to be included in the design of future RNA libraries. Thereby, the difference of each singular hydrogen bond donor and acceptor could be directly experimentally assessed.

Sequence logos

As sequence logos were a common tool to analyze the presence of conserved motifs within aligned sequences, I opted to use them for my data as well. In simplified terms, common sequence logos depicted the information content of each position within an alignment by the size of a stack of letters corresponding to nucleotides or amino acids. If the distribution of nucleotides in a position equaled the general distribution of nucleotides within the entire dataset, the stack height was zero. The stack height and therefore the information content was typically at the maximum if just one nucleotide was being found within the alignment at a location. I tried to use available online resources but they all were designed to analyze datasets of aligned sequencing data derived from biological samples. The data generated during this thesis was much different and so I could not generate trustable results with these online tools. Skylign was one of these online tools utilizing established mathematical formulas which they also published in detail together with exemplary data. I adopted their formulas in another custom script which generated the expected results from their exemplary data [130]. I extended the script to generate bar charts in Excel which it then modified to display the corresponding letters. This way, reliable and customizable sequence logos could be generated with any desired input data (Figure 39 B, Figure 40).

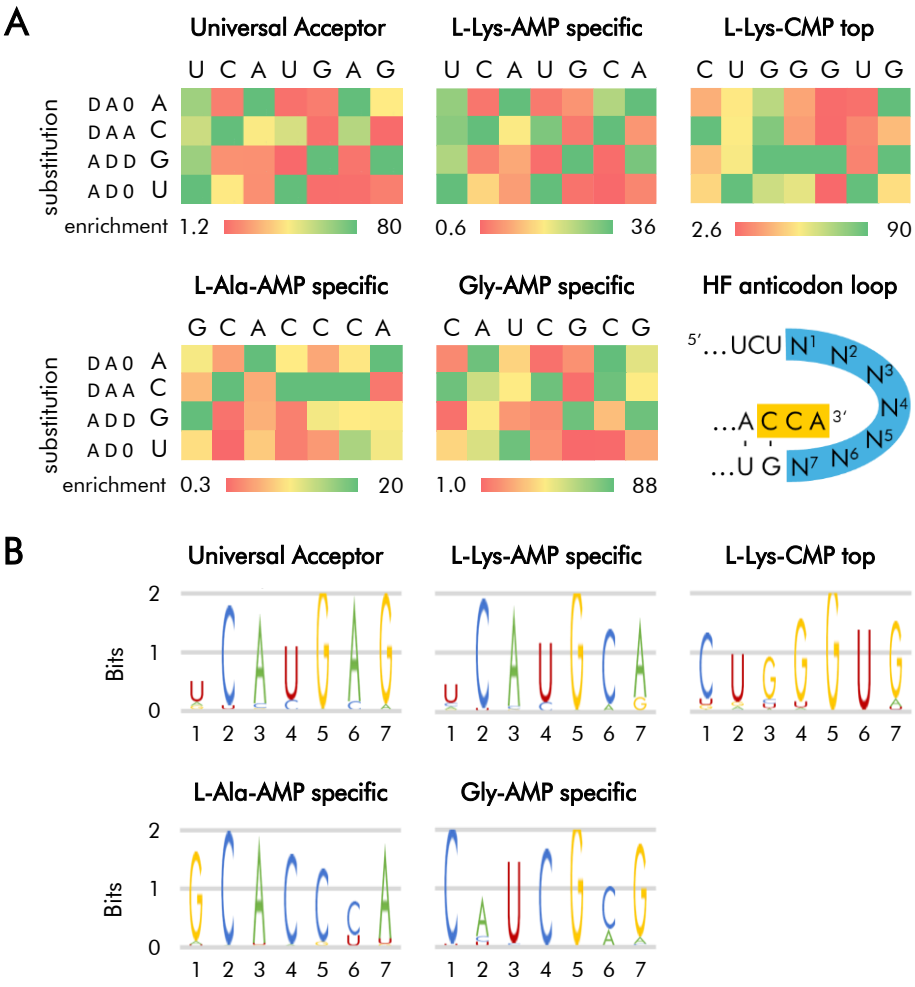


Figure 39. (A) Influence of single nucleotide mutations on the enrichment factor and thereby presumed activity of the depicted sequences towards their corresponding activated amino acid; in the case of the Universal Acceptor, its activity towards L-Lys-AMP is displayed. The results were mostly identical if another adenylate was chosen for the selected sequences. The x axis depicts the original sequence, and the y-axis changes to any other ribonucleotide alongside the distribution of hydrogen bond donors (D) and acceptors (A) at the Watson-Crick edge of the corresponding nucleobase. The color coding depicts the range of determined enrichment factors for each depicted sample. An illustration of the Hopfield fold's anticodon loop was included for positional reference. **(B)** The same data conveyed through sequence logos; each possible single mutation was weighed according to its enrichment factor. The average base distribution of the triplicate type 1 positive controls at each position was taken as reference.

Sequence logos of the 100 most enriched sequences derived from aminoacylation experiments were strongly influenced by the sequences of the Universal Acceptor in the case of adenylates or the top performing sequence in the case of the cytidylate. Therefore, these sequence logos did not differ much if the top 10, 100, or 200 most enriched sequences were considered. If all sequences were considered, the sequence logos tended to display minimal information (Figure 40).

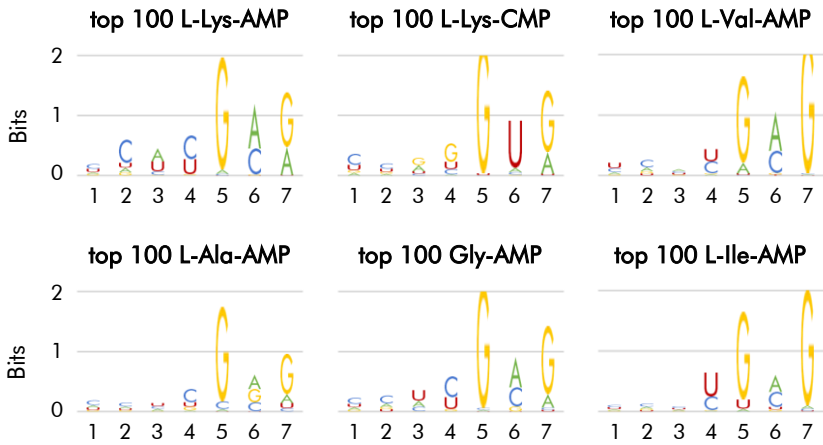


Figure 40. Sequence logos derived from the top 100 most enriched sequences coming from aminoacylation reactions with the indicated activated amino acid. The average base distribution of the triplicate type 1 positive controls at each position was taken as reference.

In a general sense, this data revealed that the adenylate and cytidylate moieties markedly influenced which sequences were enriched in a sample. All data derived from adenylates share similar sequence logos of their most enriched sequences. This indicated that the side chain of the amino acids played only a minor role in the mechanism of self-aminoacylation. A Hopfield fold's anticodon loop seemed to primarily interact with the adenosine or cytosine to facilitate aminoacylation. No amino acid side chain could alter the outcome of the sequence logos notably. However, the use of a cytidylate instead of an adenylate could do so. In the case of the cytidylate it could clearly be observed that N6 tended to be a U instead of an A or C in the case of the adenylates. No other striking changes could be observed though. N6 was the nucleotide that could hybridize with the 3' terminal nucleotide which in the case of the used Hopfield fold RNA library was an A. It therefore seemed that aminoacylation with a cytidylate preferred this terminal A to be more

tightly bound as A and U could form a canonical base pair. In the case of adenylates, the preferred bases at this position were A and C which do not pair with A; adenylates therefore seemed to benefit from a higher degree of freedom at the acceptor site. This may have compensated for the spatial requirements necessary to bring the two purine bases involved in the conserved G of the fifth position and in the adenylate in close proximity and correct orientation to catalyze aminoacylation.

In all cases, a G was preferred in the seventh position. This facilitated base pairing with the C in the complementary strand, so a more rigid conformation appeared to be preferred. An A at this position was also tolerated; maybe any purine base could meet the structural requirements needed at this position. Only in the case of L-Val-AMP and L-Ile-AMP, the conservation of G at position 7 was almost absolute. Maybe the lower reaction velocity made a more rigid structure necessary, but the precise implications were not further investigated.

Interestingly, the conserved G at position 5 could facilitate direct base pairing with the cytidylate which was not possible in the case of adenylates. If this base pairing constituted an additional advantage in the positioning of cytidylates to realize aminoacylation reactions was not investigated during this thesis. If so, future kinetic studies may reveal elevated reactivity of cytidylates compared to adenylates.

In extant biology, primarily adenosine and guanosine are involved in biochemical reactions beyond the production of genetic material like the formation of aminoacyl-5'-adenylates. Aminoacyl-5'-cytidylates, -uridylates and -thymidylates may not be relevant for extant biology, but they may contribute to the elucidation of the mechanism of self-aminoacylation of Hopfield folds and other RNA structures of interest.

With the Excel sheets and scripts I created, a plethora of analyses like the previously described ones could be conducted relatively easily for all desired sequences and samples. However, all such analyses about the potential mechanism of the interaction of the sequences with the activated amino acids remained speculative. During the course of this thesis, no further attempts were undertaken to experimentally elucidate the mechanisms of binding and catalysis involved herein.

3.2.4 Kinetic sequencing

The sequencing data obtained by the NextSeq 550 run could effectively be used to determine the relative enrichment of every sequence in every sample. But neither the sequencing data nor the Bioanalyzer data could be used to obtain absolute

yields of the aminoacylation reactions. Therefore, kinetics could not be derived from the sequencing data directly.

In the case of the iSeq 100 run with the four aminoacylation samples derived from L-Lys-CMP, I undertook my first and only attempt of realizing kinetic sequencing during this project. Inspired by the approach of Chen and coworkers [131], I added a DNA strand during the initial setup of the aminoacylation experiments at the same time the Hopfield fold RNA library was added. The amount of this DNA strand was known and through the use of a master mix, the ratio between DNA and RNA was kept identical amongst all samples. The DNA strand was blocked at its 3′ end with a C3 spacer to prevent the formation of undesired products during ligation and it was modified at its 5′ end with biotin to protect it from digestion through Lambda Exonuclease after ligation. Its sequence and structure mimicked a regular ligation product but with two altered nucleotides for identification (Figure 16, Figure 22). This DNA was supposed to undergo all steps from aminoacylation to Bioanalyzer readout alongside the RNA except for oxidation and ligation. It was also thought to fully pass through the ligation-based library preparation in contrast to the RNA of which only the non-oxidized fraction could do so. By comparing the number of reads obtained by DNA and RNA, the aminoacylation yield of the RNA could have been determined.

The final outcome of this attempted k-Seq did not yield any data that could resemble kinetics, though. I tried several mathematical normalization approaches, but none produced acceptable results.

3.3 Wet-lab validation

3.3.1 APB-PAGE densitometry

Aminoacylation efficiency of the Hopfield fold RNA library could only be estimated indirectly through the bioanalyzer readout after library preparation due to its low yield. But as even the values of replicates showed high degrees of variation, this data was not reliable to derive kinetic insights. To assess kinetics directly, densitometry of APB-PAGE gels was to be employed on aminoacylation experiments with the four previously selected sequences.

To produce these RNA sequences, I first ordered DNA templates for in-vitro transcription. Even though clear product bands of RNA of about the right size were forming, no aminoacylation activity could be observed. The most likely explanation for this was that IVT is prone to produce heterogeneous 3′ and 5′ ends. In addition to that, sequencing generated some data derived from Hopfield folds that were truncated by one nucleotide at their 3′ end. The enriched sequences in this case seemed different to the ones derived from full-length Hopfield folds, further pronouncing the importance of a well-defined 3′ end. In extant life, the integrity of tRNA 3′ ends is as well ensured by a dedicated enzymatic machinery [132]. As missing or additional nucleotides at the 3′ end strongly alter the structure of the active center, it was unsurprising that a method that was prone to such alterations was unable to produce reactive RNA strands in a reasonable manner. Alterations at the 5′ end were supposed to influence aminoacylation activity to a much lower degree due to its distance to the active center.

IVT would have been advantageous to chemical synthesis due to much lower cost per RNA sequence to be produced. To result in RNA strands with homogeneous 3′ and 5′ ends, likely the use of hammerhead or other self-cleaving ribozymes would have had to be included. Such self-cleaving ribozymes could be added to both ends of the RNA to be produced and would excise themselves from the desired RNA product at a predefined position resulting in precisely the desired ends. But as the number of RNAs of interest was limited, the optimization of such sequences was deemed too time-consuming, so the four selected sequences were instead commercially ordered to be produced by chemical synthesis.

The chemically synthesized sequences exhibited the expected aminoacylation behavior but unlike in all previous experiments with the Hopfield fold RNA library, I could directly observe hybridization bands after aminoacylation and APB-PAGE. As these hybridization bands made densitometric readouts less reliable, I tested a limited combination of denaturing conditions but as urea was standardly part of

the gel loading dye, an additional heating step sufficed to denature the sample. The addition of formaldehyde led to no beneficial effects.

In addition to the newly added denaturation step, other standards, like thinner gels and lower amounts of RNA per lane were established alongside. All these conditions ensured that all samples were optimally stained to produce reliable results. Also, consistent resolution and laser settings were applied on the Typhoon biomolecular scanner during imaging. Densitometric analyses were conducted through the Software ImageQuant which was developed by the same company that supplied the Typhoon scanner. I always conducted the densitometric analyses in manual mode even though the program provided some automatic approaches. I always defined the bands of interest manually and chose the *rolling ball* method for the subtraction of the background signal (Figure 41). The output data could be stored in tabular form for further analysis.

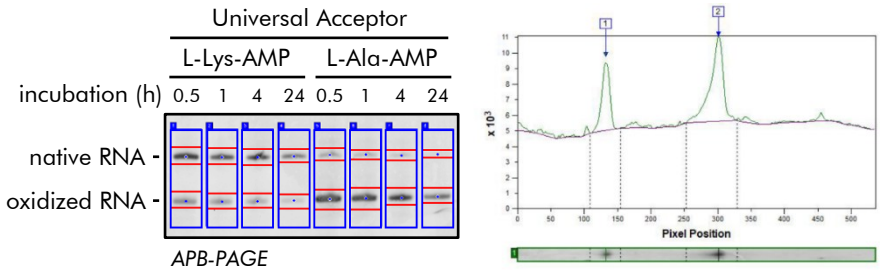


Figure 41. Exemplary densitometric analysis. **(Left)** Original image after APB-PAGE of two aminoacylation reactions with technical markup showing the analyzed area of each lane and the manually selected areas of interest around each band. **(Right)** Original readout as generated by ImageQuant showing a typical densitogram with rolling ball background subtraction and manual marking of bands displayed.

The quantified data obtained from the densitometric analyses could be used to calculate the percentual yield of the reaction (Figure 42). As this approach was ratiometric, its outcome was independent from the amount of input material per lane as long as it was within linear quantification range. Therefore, minor fluctuations in the input amount were standardly normalized. As the absolute amount of input RNA per aminoacylation was known, it could be used to calculate the absolute amount of aminoacylated RNA or its concentration within the reaction volume.

$$percentual\ yield = 100 * \frac{c(native)}{c(native) + c(oxidized)}$$

Figure 42. Formula by which the percentual yield of an aminoacylation reaction was calculated from the densitometric readout. c() depicted the integral of the signal derived from either the oxidized or non-oxidized (native) RNA in a sample.

Most likely, the aforementioned hybridization bands originated from dimers of Hopfield folds as the sequence of the stem loop enabled not only the formation of a stem loop but could also facilitate dimerization leading to an equal conformation of anticodon loops and 3′ ends but with the anticodon loop of one molecule being in close proximity to the 3′ end of another one.

On the one hand, no hybridization bands in the case of the Hopfield fold RNA library with randomized anticodon loops could be observed but if during the aminoacylation reaction, Hopfield folds could couple and decouple, this could facilitate terminal aminoacylation also of unreactive sequences. This in turn would lead to a more pronounced background and top performing sequences were potentially less strongly enriched. On the other hand, clear hybridization bands could be observed in the case of the selected sequences with defined anticodon loops and whether or not they acted in trans could not be determined.

This could add another facet to the evolution of tRNA from Hopfield folds. In an early system, Hopfield folds may have also acted in cis and trans to realize aminoacylation and function as tRNA precursor in a single kind of molecule. Over time, they might have evolved into two lines with one becoming a more specialized tRNA and the other line becoming a more specialized trans-acting aminoacylating ribozyme comparable to the flexizyme system. Eventually, the latter would have been displaced by the emergence of more capable protein-based aaRSs.

3.3.2 Comparison with sequencing data

To verify the sequencing data, the percentual yield obtained by densitometry had to be compared with the enrichment factors obtained by sequencing. It is to be noted that the densitometry data was mostly obtained by aminoacylation incubated for 4 h as this led to the maximum yield. For sequencing I have employed 3 h or 6 h of incubation time, though. As I could not observe much of a difference in the sequencing data between these two incubation times, I decided to use data from both incubation times equally (Figure 37). Further on, I figured to keep the densitometry data completely unaltered and to only adapt the sequencing data. To

compare densitometric with sequencing data from similar reactions, I treated the enrichment factors as if they were percentual yields, too. I divided all of them by the same correction factor so that their sum was equal to the sum of percentual yields derived from densitometry. These sequencing-derived values on their own had no practical meaning but their ratio amongst each other was fully maintained and the ratio of the sums of sequencing-derived datapoints and densitometry-derived datapoints was 1:1.

By doing so, the densitometric data seemed quite correlative with the sequencing data with the only exception of the Gly-specific sequence reacting with L-Lys-AMP or Gly-AMP (Figure 43). Except for these two latter cases, all other data points were shared notable similarity and the overall spread of replicate data appeared rather low. This indicated that the enrichment factors obtained by sequencing could indeed well be considered as a proxy for reactivity. Even though it remained unclear why the Gly-specific sequence behaved so differently in the sequencing approach compared to the individual testing, the general findings strongly support the validity of also the data derived from sequencing that was not further confirmed.

All sequencing data was derived from L-amino acids if applicable but to further explore the capabilities of the selected sequences, D-Lys-AMP and D-Ala-AMP were prepared to conduct aminoacylation experiments with. In all cases, the use of D-Lys-AMP led to lower aminoacylation yields compared to L-Lys-AMP expect for the so-called L-Ala specific sequence. Its specificity was determined from experiments with L-Lys-AMP, L-Ala-AMP, and Gly-AMP. If paired with D-Lys-AMP though, this sequence exhibited the highest aminoacylation yield exceeding what could be observed in experiments using L-Ala-AMP. In any other case, the use of D-Lys-AMP or D-Ala-AMP resulted in lower or equal yields compared to L-Lys-AMP or L-Ala-AMP, respectively (Figure 43). This indicated that the alpha amino group seemed to be involved in aminoacylation catalysis, but the significance of its conformation was sequence-dependent. The subset of sequences tested here was too small to infer general effects of the L- or D-conformation of aminoacyl-5'-adenylates on the aminoacylation activity of Hopfield folds.

Aminoacylation experiments with the Hopfield fold RNA library and activated D-amino acids were not conducted solely due to time constraints. If such experiments were to be conducted in the future, they could further elucidate the role of the alpha amino group in aminoacylation. In combination with kinetic sequencing, it could be revealed if L- or D-amino acids led to generally higher aminoacylation yields or more selective sequences. But it may also be revealed that no general effects exist, and all differences remain sequence-dependent.

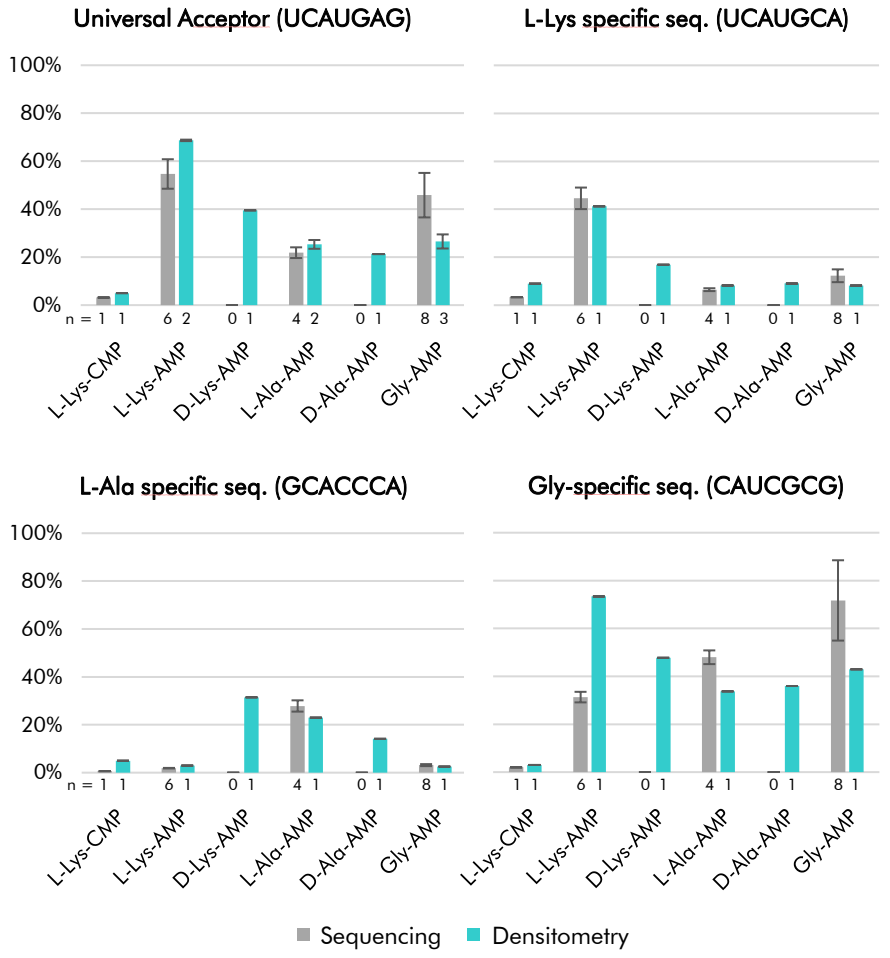


Figure 43. Comparison of data obtained through sequencing and densitometry. See main text on how sequencing data was normalized. Sample size is indicated by n below each bar. Error bars indicate standard deviation.

3.3.3 Further characterization of the Universal Acceptor sequence

Aminoacylation and densitometric analysis were conducted with the Universal Acceptor sequence and the aminoacyl-5'-adenylates of L-Tyrosine, L-Proline, L-Phenylalanine, L-Leucine, and L-Tryptophan even though no corresponding prior sequencing data was generated. It could be observed that the sequence reacted to a limited degree with these adenylates reaching 30% yield in the case of L-Tyrosine,

around 5 to 10% in the case of the other adenylates except for L-Tryptophan which resulted in no detectable yield (Figure 44).

These findings further pronounced the universality of the sequence. However, it would constitute an evolutionary challenge, if for Hopfield folds universality was easier to achieve than specificity. If a primordial translation machinery was to evolve from Hopfield folds and most of the mutations that occurred in the Hopfield folds led to a loss of activity or specificity, it would be hard for such a system to evolve towards the minimization of errors during translation.

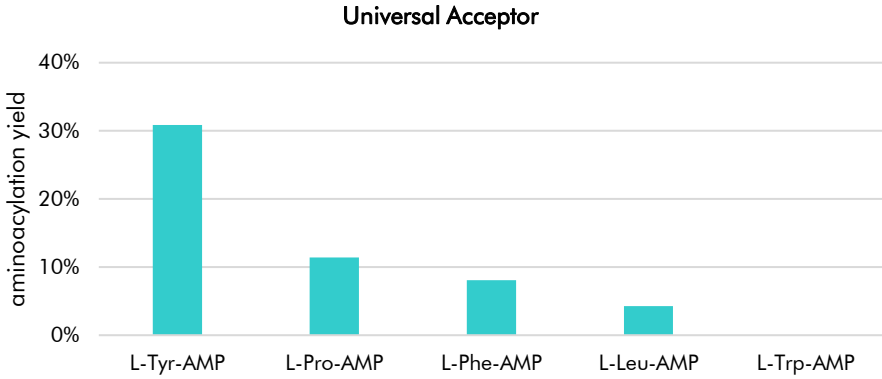


Figure 44. Aminoacylation yield of the Universal Acceptor paired with the indicated aminoacyl-5′-adenylates. The reaction was incubated for 4 h and analyzed through strandard densitometry. No yield could be detected in the case of L-Trp-AMP.

As ribozymes and aptamers usually exhibit complex interactions between distant parts of their sequence like kissing loops or pseudoknots, the activity of two 5′-truncated versions of a Hopfield fold were tested. Again, the Universal Acceptor sequence was chosen for this experiment as it exhibited consistently high yields with most adenylates tested. I ordered two such truncated sequences to be chemically synthesized; in the case of the *short* version, half of the nucleotides from the 5′ end to the anticodon loop were deleted. In the case of the *minimal* version, all these nucleotides were deleted so that one nucleotide of the anticodon loop directly formed the 5′ end. Employing the usual densitometric assays, it could be show that the short version exhibited markedly decreased aminoacylation yield compared to the full-length version while maintaining the ratio of aminoacylation yields amongst the used adenylates. The minimal version showed no detectable activity (Figure 45).

This indicated that the 5′ region was essential for efficient aminoacylation even though it harbored sequences that were supposed to be only relevant for Illumina sequencing. Either these sequences were by chance suited to form auxiliary structures with the rest of the sequence by specific interactions like hydrogen-bonding or a much larger set of sequences would have been able to exhibit similar effects by non-specific interactions like pi-stacking or by simply acting as ballast, restricting the anticodon loop in its flexibility. The precise involvement of the 5′ region in the aminoacylation reaction was not further investigated though. Smaller steps of truncation or similar experiments with other sequences were not conducted due to the high cost of chemically synthesized RNA.

If similar experiments with other sequences led to similar results, it would suggest that a Hopfield fold must possess a minimal length to exhibit self-aminoacylation and that both a defined stem loop structure at the 3′ area and sufficient material at the 5′ area must have been present before evolution towards self-aminoacylation could have occurred.

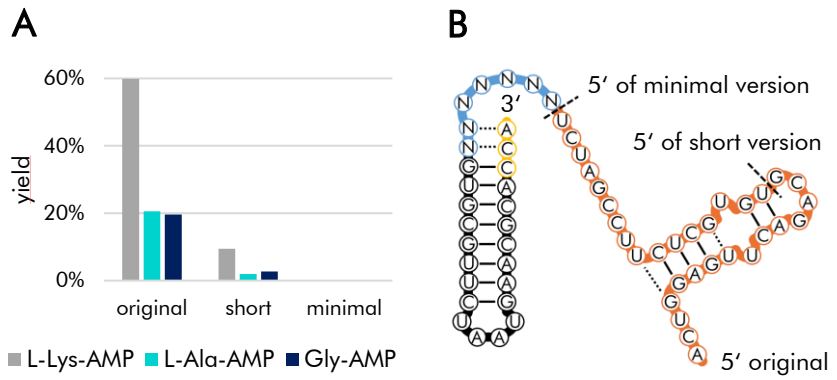


Figure 45. Influence of 5′ truncation on the aminoacylation activity of the Universal Acceptor sequence. **(A)** Bar chart depicts yields after 4 h upon aminoacylation with the indicated adenylates. **(B)** Indication of which parts were truncated in the *short* and *minimal* version of the Universal Acceptor.

As described in Chapter 3.2.3 and seen in Figure 43, the adenylate or cytidylate moiety of activated amino acids and not their side chains seemed to mainly influence the selection of reactive sequences. However, both the conformation of the alpha amino group as previously described, and the amino acid side chain seemed to influence aminoacylation activity in a sequence-dependent manner (Table 2, Figure 44).

I therefore assumed that the side chains of amino acids primarily exerted steric hindrance and rather little attractive interaction. This could explain why amino acids with no or small side chains like Glycine and Alanine resulted in high aminoacylation yield but specificity could be achieved much harder as the Gly-specific sequence was the least specific one compared to the L-Lys-specific sequence and the L-Ala-specific sequence and the latter being the least reactive in comparison. Also, in the case of the Universal Acceptor paired with L-Pro-AMP, L-Phe-AMP, L-Leu-AMP, and L-Trp-AMP, the molecular weight of the amino acids' side chains vaguely correlated with their associated decreasing aminoacylation yield (Figure 44).

The only exceptions to this general trend seemed to be Tyrosine and Lysine. Tyrosine was the only used amino acid that harbored a hydroxy group as part of its side chain. In the cases of Serine and Threonine, adenylate synthesis was unsuccessful. This hydroxy group could have enabled the formation of attractive hydrogen bends during aminoacylation with the Universal Acceptor sequence. Thereby, the steric demands of the side chain could in part be compensated as through the lack of this hydroxy group, aminoacylation yield was only about a third as seen in the case of Phenylalanine (Figure 44).

Lysine is the only proteinogenic amino acid that contains a second amino group and is next to arginine one of two proteinogenic amino acids that can regularly exhibit a second positive charge. Especially at pH 5, both amino groups of lysine or a lysyl-moiety are likely to be protonated, exhibiting positive charges. Analogous to cations, this positive charge may interact with the negative charge of the phosphate backbone of the RNA Hopfield folds giving it a sequence-independent advantage in binding affinity compared to other amino acids. If employed in as high concentrations as in the case of regular aminoacylation reactions, these additional positive charges may even exhibit a stabilizing effect on the secondary structures of the Hopfield folds. Even though lysine was the most reliable amino acid both in terms of adenylate synthesis and aminoacylation, its additional positive charge may not be the sole explanation for this behavior. Lysyl-5'-adenylate did not react unspecifically with the Hopfield fold RNA library. There was a limited selection of high performing sequences whereas most other sequences did not show increased reactivity to L-Lys-AMP. This indicated that the potentially enhanced unspecific binding of lysine to RNA played only a minor role in aminoacylation efficiency and that still a complex and well defined active center was necessary to enable efficient terminal aminoacylation.

Amino acids that result in low aminoacylation yields may cause the enrichment of more specific sequences as due to their increased size, a larger set of Hopfield fold sequences would be sterically hindered to realize self-aminoacylation. If these assumptions were true, the hypothetical evolution of primordial Hopfield folds

would face the opposing tendencies of simple amino acids which impede specificity would be available earlier in evolutionary terms than more complex amino acids which may be more specific yet harder to bind at all.

To further investigate these mechanisms, sequencing experiments with more aminoacyl-5′-adenylates may be conducted which may also include amino acids with non-canonical side chains. For example, analogous to the experiment with Phenylalanine and Tyrosine, the comparison of Alanine and Serine may elucidate the influence of an additional hydroxy moiety in another context. The comparison of Lysine and Norleucine may elucidate the effect of the epsilon amino group and so on.

3.3.4 Determination of rate constants

To gain further insight into the kinetics of the given aminoacylation reactions, Prof. Jäschke recommended me to use the software DynaFit [133] to extract reaction rate constants from the obtained data. DynaFit uses nonlinear least-squares regression to analyze chemical kinetics and equilibria. We hypothesized that aminoacylation reactions should mainly be dictated by three separate reactions and their corresponding reaction rate constants. The first rate constant corresponded to the hydrolysis of the activated amino acid during the aminoacylation reaction into free amino acids and AMP; we defined the corresponding rate constant as k_1. Alternative to hydrolysis, the activated amino acid and RNA could react to form 2′ or 3′ terminally aminoacylated RNA and AMP; potential internal 2′-aminoacylation was neglected. We defined the rate constant of this terminal aminoacylation as k_2. Eventually, the terminally aminoacylated RNA hydrolyzed into RNA and free amino acids whose rate constant we defined as k_3 (Figure 46). We developed the necessary scripts to properly run the software, but the actual analyses were solely conducted by the author of this thesis.

k_1: aa-AMP → aa + AMP

k_2: RNA + aa-AMP → RNA-aa + AMP

k_3: RNA-aa → RNA + aa

Figure 46. Definition of rate constants (k). Hydrolysis of aminoacyl-5′-adenylates is represented by k_1. Terminal aminoacylation of RNA through an adenylate is represented by k_2. Hydrolysis of aminoacylated RNA is represented by k_3.

The rate constants of the hydrolysis of L-Lys-AMP, L-Ala-AMP, and Gly-AMP (k_1) were extracted from NMR data. As described before (Chapter 3.1.2), I always measured the integrity of the used adenylate alongside every aminoacylation experiment. In early studies, NMR measurements were conducted within the same time intervals alongside ongoing aminoacylation experiments. Through this data, k_1 could be obtained in replicates of the three aforementioned adenylates at pH 5. In the case of L-Lys-AMP, hydrolysis was also investigated at pH 6 and 7. It could clearly be shown that the rate of hydrolysis (k_1) was dependent on pH; the higher the pH, the faster hydrolysis progressed and the higher k_1 was. A first order rate equation including k_1 could well be used to model the hydrolysis of all tested adenylates and pH values as exemplarily shown for L-Lys-AMP in Figure 47.

If buffers with a pH above 5 were used, the measured initial concentration of intact adenylate was considerably lower compared to measurements conducted in pure D_2O or buffers at pH 5 and lower. Even though the hydrolysis rate of adenylates was markedly increased at above pH 5, due to handling, only a few minutes lay between the dissolution of the dry adenylate and its ^{31}P-NMR measurement, which did not appear sufficient to explain the observed degree of loss. This initial loss effect could be observed highly consistently, but no profound explanation could be found for this phenomenon (Figure 47).

For the modeling of aminoacylation reactions and the determination of k_2 and k_3, DynaFit required the initial concentration of adenylate at the start of the reaction alongside the previously determined rate constant k_1 to calculate the decay of the adenylate during the reaction. In the case of all aminoacylation experiments, this initial concentration was measured by ^{31}P-NMR. Through densitometry as previously described, the yield of aminoacylated RNA was measured throughout the ongoing reaction at different time intervals.

Most aminoacylations were conducted at pH 5 but in the case of the Universal Acceptor paired with L-Lys-AMP, reactions were also conducted at pH 6 and 7. It could be observed that with increasing pH, maximum yield decreased and the hydrolysis rate of aminoacylated RNA increased. At pH 7, values obtained by densitometry were barely above the limit of detection. It could also be observed that rate equations based on the assumed model were unable to model aminoacylation reactions that used L-Lys-AMP. It could be observed that independent from the used sequence, if L-Lys-AMP was used, measurements from early timepoints lay below the modeled curve and measurements from later timepoints lay above (Figure 47). However, aminoacylation reactions that involved L-Ala-AMP or Gly-AMP could be represented well by the assumed the model (Figure 47).

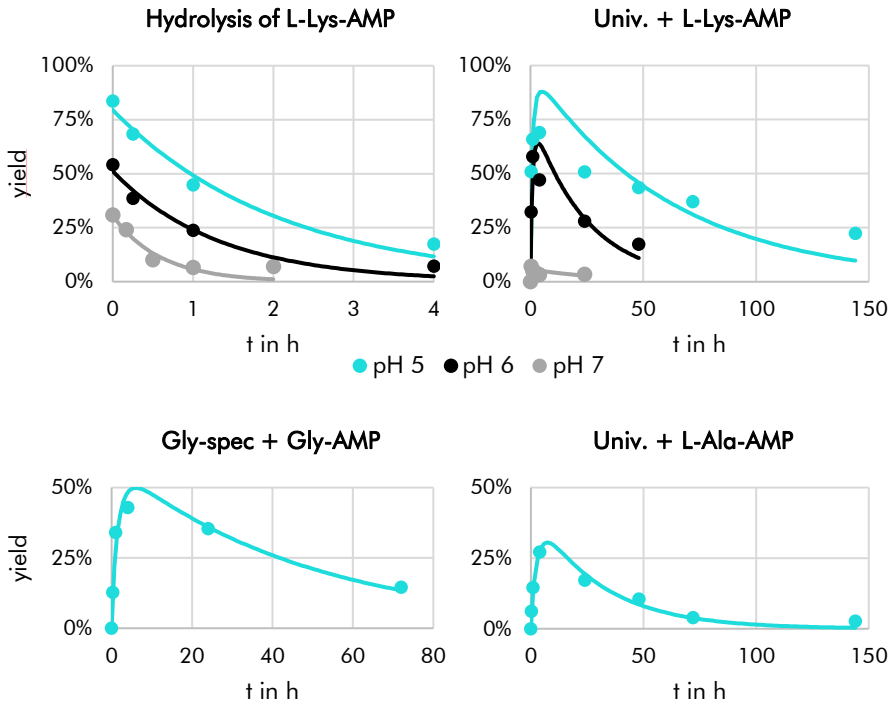


Figure 47. Exemplary hydrolysis and aminoacylation data including modeled regression curves. No aminoacylation reaction including L-Lys-AMP could be modeled properly with the proposed model whereas all other ones could be fitted well. Also, pH 5 was employed in most reactions; all reliable data generated at other pH values are displayed here.

This indicated that the reaction mechanism of L-Lys-AMP with Hopfield folds was more complex than initially assumed and different to L-Ala-AMP and Gly-AMP which seemed to correspond to the assumed model. Potentially, only L-Lys-AMP may cause terminal 2′ and 3′ aminoacylation with one of them being less hydrolytically stable than the other. Looking at the data points, it appeared as if after the maximum aminoacylation yield was reached, the hydrolysis rate was increased for a short duration but was stable at the latest after 24 hours of incubation.

Depending on the pH applied, the incubation time until maximum yield was achieved as well as the half-life of aminoacylated RNA significantly differed. At pH 5, maximum yield was achieved after ~4 h of incubation and the half-life was ~45 h, maximum yield at pH 6 after ~2 h and half-life of ~20 h, maximum yield at pH 7 after ~1 h and half-life of ~2 h. According to these findings, the tested

Hopfield folds performed best at elevated pH but showed almost no activity at neutral pH which is closer to extant cytosolic pH. However, most charged extant tRNAs also exhibit a half-life of ~2 h at cytosolic pH so hydrolytic instability does not seem to hinder translation in general [134].

To gain a more complete picture on the kinetics of each of the four selected sequences paired with different adenylates, their resulting rate constants were collected in tabular form (Figure 48 A & B). It could be observed that different pairings of sequences and adenylates resulted in various degrees of variation. It could be observed that the values obtained for k_2 showed the highest variation in the column of L-Lys-AMP and in the row of the Gly-specific sequence. In the case of the Gly-specific sequence paired with Gly-AMP, k_2 even appeared to correlate with the initial concentration of adenylate in such a way that it decreased with rising concentration (Figure 48 C). A possible explanation for this behavior could be inhibition by a surplus of substrate. This could also explain the high degree of variation observed in the sequencing data and initial validations (Figure 43). In any case, this suggested that the kinetic model of aminoacylation requires further refinement.

In the case of k_3, increased hydrolytic stability of the L-Lys-specific sequence and L-Ala-specific sequence compared to the other two sequences could be observed (Figure 48 B). This suggested that hydrolysis could be generally stabilized in a sequence-dependent manner. However, this would contrast with one of the key points of Hopfield's hypothesis which was that hydrolysis would be minimal only for correctly charged Hopfield folds. At least the four selected sequences did not exhibit a significantly lowered hydrolysis rate for one amino acid.

The low activity at neutral pH, low specificity, and the predominance of aminoacylation over stabilizing hydrolysis seemed to disprove Hopfield's hypothesis. But the design of the aminoacylation experiments conducted herein introduced a large amount of adenylate at once and sequences with highest aminoacylation yield after a relatively short incubation time were favored. In a prebiotic setting it appears more likely that an activated amino acid was supplied constantly yet in much lower concentrations. If a selection of sequences that achieve maximum yield and specificity was to be conducted under such conditions, the results may be significantly different from the ones obtained here. Under such conditions, sequences that exhibit minimal hydrolysis over long periods of time may be the most enriched and specificity may be realized indeed through different hydrolysis rates rather than aminoacylation rates.

A

[RNA + aa-AMP → RNA-aa + AMP]:
k_2 in $pM^{-1}s^{-1}$

		L-Lys	D-Lys	L-Ala	D-Ala	Gly
UCAUGAG	Universal acceptor	527 [6] ±367	14,3 [1]	5,6 [2] ±0,34	4,66 [1]	6,57 [3] ±0,32
UCAUGCA	L-Lys-specific	111 [5] ±54,7	5,39 [1]	1,41 [1]	1,61 [1]	2,17 [1]
GCACCCA	L-Ala-specific	0,56 [1]	10,7 [1]	3,28 [2] ±1,3	2,79 [1]	0,65 [1]
CAUCGCG	Gly-specific	210 [2] ±170	18,2 [1]	23,2 [2] ±15,3	8,36 [1]	92,1 [6] ±57,1

k2 [n] ±s.d.

reactive

unreactive

B

[RNA-aa → RNA + aa]:
k_3 in ms^{-1}

		L-Lys	D-Lys	L-Ala	D-Ala	Gly
UCAUGAG	Universal acceptor	19,8 [6] ±9,63	22,8 [1]	35,9 [2] ±1,95	28,1 [1]	14,1 [3] ±3,39
UCAUGCA	L-Lys-specific	9,2 [5] ±2,43	2,5 [1]	7,6 [1]	6,59 [1]	11 [1]
GCACCCA	L-Ala-specific	18,9 [1]	7,6 [1]	8,2 [2] ±0,9	7,8 [1]	7 [1]
CAUCGCG	Gly-specific	21,3 [2] ±3,65	20,2 [1]	28 [2] ±9,31	28,3 [1]	31,3 [6] ±20,3

k3 [n] ±s.d.

stable

unstable

C

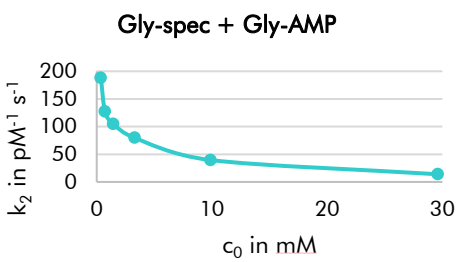


Figure 48. Reaction rate constants (k) alongside sample size (n) and standard deviation (±s.d.). **(A)** Rate constants corresponding to the aminoacylation reaction of the indicated combinations of sequences and aminoacyl-5′-adenylates. Color coding for most reactive (green) and least reactive (red) combinations. **(B)** Rate constants corresponding to the hydrolysis of the indicated combinations of aminoacylated RNA. Color coding for hydrolytically most stable (green) and least stable (red) combinations. **(C)** Exclusively in the case of the Gly-specific sequence combined with Gly-AMP, k_2 seemed to correlate with the initial concentration of adenylate (c_0).

3.3.5 Determination and implications of Km values

The Km values of each selected sequence towards its specific adenylate and that of the Universal Acceptor sequence towards L-Lys-AMP were to be determined. Therefore, aminoacylation experiments were conducted at continuously lowered initial concentrations (c_0) of adenylate but otherwise identical conditions. In combination with the resulting initial reaction velocities (v(ini)) as provided by DynaFit, the Km values were determined through extrapolation on Lineweaver-Burk plots (Figure 49). In the indicated cases, the resulting data points exhibited almost no deviation from the linear regression, rendering the extrapolated Km values reliable. Km for the Ala-specific sequence paired with L-Ala-AMP could not be determined because the corresponding densitometric data was close to or below the quantification limit and therefore not usable.

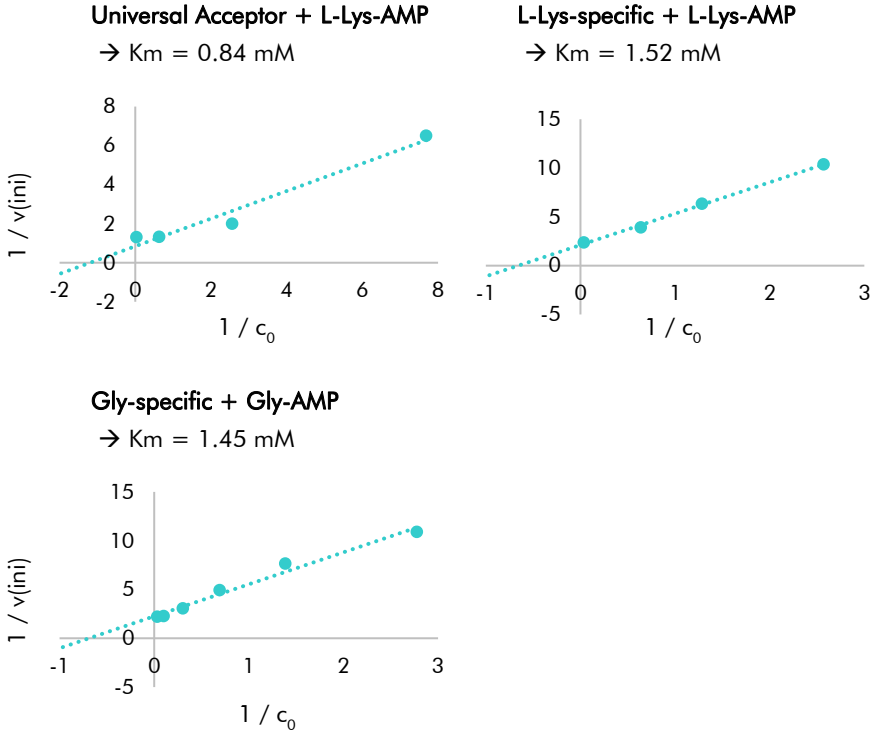


Figure 49. Lineweaver-Burk plots of the indicated combinations of sequences and adenylates from which Km was extrapolated in each case. The corresponding data of the L-Ala-AMP specific sequence paired with L-Ala-AMP was of too low quality due to low yield to be suitable for this analysis.

So far, initial adenylate concentrations of ~30 mM have been consistently used but after Km values were determined, further aminoacylation experiments with starting concentration near the Km values of 0.5 to 1.5 mM were conducted to investigate effects on yield and specificity (Figure 50).

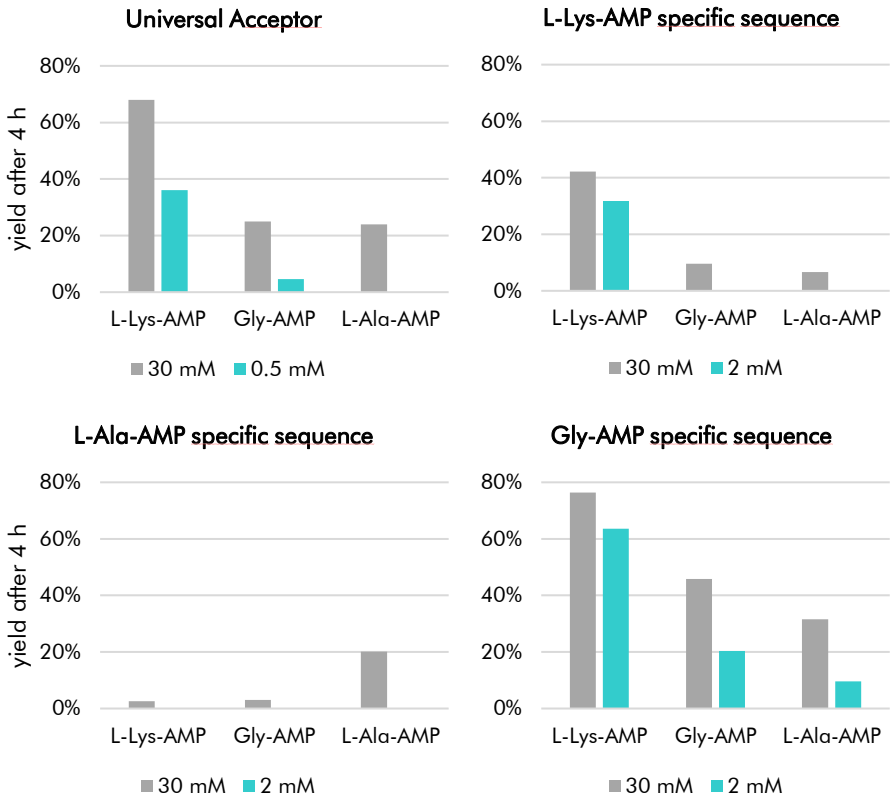


Figure 50. Aminoacylation reactions with the regular starting concentration of 30 mM compared to reactions with starting concentrations near Km values of 0.5 to 2 mM. Percentual yields of aminoacylated RNA after 4 h of incubation as determined by densitometry are displayed.

Even though the yield resulted to be generally lower, the specificity for L-Lys-AMP was visibly higher in the case of the Universal Acceptor and absolute in the case of the L-Lys-AMP specific sequence. The L-Ala-AMP specific sequence showed no

detectable activity at lowered initial concentrations and the specificity of the supposed Gly-AMP specific sequence was generally unaltered.

As previously mentioned, a consistent supply of activated amino acids at low concentrations appears to be prebiotically more plausible than singular events of high concentrations of an activated amino acid. In evolutionary terms, it would be highly beneficial if Hopfield folds in general reacted more specifically at low concentrations of activated amino acids. Again, this would in turn decrease the error rate of the associated translation machinery which would be critical for the successful production of functional proteins.

3.4 Further coding potential of the acceptor stem

During this thesis, a cooperation with the research group of Prof. Dr. John Sutherland (MRC Laboratory of Molecular Biology) was undertaken. They were working on an own approach to the origin of the genetic code that happened to be compatible with the sequencing techniques developed herein. Their approach focused on two short RNA strands with one being aminoacylated at its 5′ end (donor) and able to hybridize with another RNA strand which formed a single stranded overhang with its 3′ region (acceptor) (Figure 51 A). The aim was to investigate the efficiency of the transfer reaction of the bound amino acid from the 5′ end of the donor to the 3′ end of the acceptor. The group could already show that the sequence of the overhang had significant influence on this reaction. They already found highly reactive 3′ overhang sequences by manually screening individual sequences one at a time.

To further investigate the influence of the double stranded region directly adjacent to the single stranded overhang, they randomized both the three nucleotides in the 5′ terminal region of the donor strand as well as the corresponding nucleotides in the acceptor strand. Thereby, a three base pairs long double stranded region was formed (Figure 51 A). Through the sequencing technique I developed, the influence of this randomized stretch on the aminoacylation activity should be investigated.

Meng Su was conducting the laboratory work and sent me RNA samples after they underwent the described aminoacyl transfer reaction and oxidative fixation which I shared my procedures for. Only the donor strand was compatible to the ligation-based library preparation I developed. The donor strand was highly limited in length due to the chemistry involved in its 5′ aminoacylation. To simplify library preparation, we added parts of the Illumina P7 sequence to the 5′ end of the RNA acceptor to act as a primer binding site for PCR. To avoid this additional single stranded region to interfere with the aminoacyl transfer reaction, a DNA blocking strand was introduced that could hybridize with this P7 overhang (Figure 51).

In this setup, only the sequences of one of the strands forming the randomized double stranded region could be sequenced. Preferences for canonical or non-canonical base pairing could therefore not be elucidated. The obtained sequencing results were published in a joint publication [135].

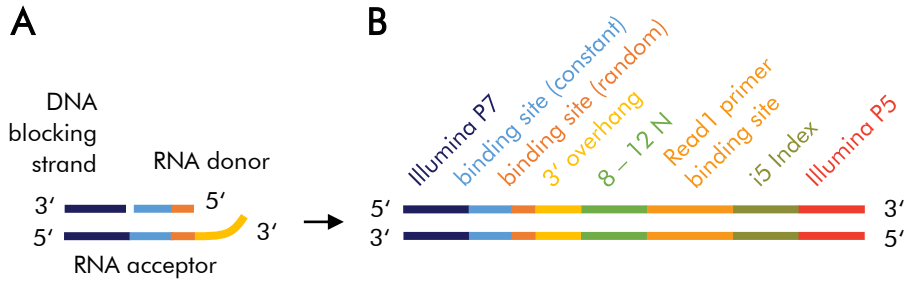


Figure 51. (A) An aminoacyl transfer reaction was realized through the use of a regular RNA strand called *acceptor*, a 5′ aminoacylated RNA strand called *donor*, and a *DNA blocking strand* that hybridized with the 5′ overhang of the RNA acceptor to avoid it interfering with the reaction. **(B)** Only 3′ aminoacylated RNA acceptors were compatible with the ligation-based library preparation procedure described in this thesis (Figure 16) giving rise to the indicated dsDNA library. Read 2 sequencing primer binding site and i7 index were omitted to keep the RNA acceptor as short as possible.

It was demonstrated that different sequences were enriched compared to an untreated, non-oxidized positive control, depending on the amino acid used for 5′ aminoacylation of the donor. The use of L-Ala, D-Ala, and Gly resulted in enriched sequences that were markedly distinct from each other and from those enriched using other amino acids. Among these, Gly showed the strongest preference for a single sequence, while L-Val, L-Leu, and, to a lesser extent, L-Pro led to the enrichment of similar sequences [135].

The sequencing data could not provide absolute values or kinetic information, so the effect of the randomized region on the reactivity of the system could not be quantified. However, qualitatively, the randomized region appeared to strongly influence both specificity and reactivity. It is important to note that the sequence space in this experiment was highly restricted, as three randomized nucleotides allowed for only $4^3 = 64$ different sequences. Moreover, as previously mentioned, only one strand of this supposedly double-stranded randomized structure was sequenced.

Eventually, the data obtained did not resemble any extant codons or anticodons associated with the used amino acids. Nonetheless, this system roughly resembled the acceptor stem at the 3′ region of tRNA and parallel to the results of this experiment, extant acceptor stems influence the aminoacylation behavior of their respective tRNA, too. Albeit the aminoacylation of extant tRNA is mediated by interactions with their cognate aaRS (“second genetic code”) including the acceptor stem and not by autocatalysis.

In the case of the Hopfield fold RNA library used herein, significant influence of a double stranded region on aminoacylation could be demonstrated, too (Chapter 3.2.3). Hopfield folds form a double stranded 3′ end with a single stranded 5′ region at its proximity. In contrast, the system employed by Sutherland and coworkers formed a double stranded 5′ end with a single stranded 3′ overhang. In general, the fact that both setups could readily realize aminoacylation reactions in addition to the manifold systems developed by Yarus, Chen, and coworkers could again demonstrate the pronounced ability of RNA to achieve aminoacylation in different ways. More specifically, if double stranded regions significantly influenced aminoacylation behavior, loop-forming RNA sequences could be employed that harbor double stranded regions of randomized base pairs. Thereby, the effects of non-canonical base pairing could be thoroughly investigated as unlike the system employed by Sutherland and coworkers, any loop-forming sequence could be completely sequenced and would be compatible with the ligation-based library design as long as it harbored a constant sequence at its 5′ end.

If structures other than the anticodon loop were involved in the realization of self-aminoacylation of the 3′ end, these could be defined as a "third genetic code". The "first genetic code" refers to the encoding of amino acids and the "second genetic code" generally refers to the elements involved in the recognition of tRNAs by their cognate aaRSs. A "third genetic code" that regulated the self-aminoacylation of primordial tRNA precursors might have left molecular vestiges across the entirety of extant tRNAs sequences. This hypothesis was also affirmed by the conducted truncation experiments which revealed that distant, 5′ terminal sequences were critical for the Universal Acceptor sequence to realize self-aminoacylation (Figure 45).

4 Conclusion and outlook

4.1 Overview

The objective of this thesis was to develop a sequencing-based method to investigate Hopfield's hypothesis on the stereochemical origin of the genetic code.

A suitable method was successfully developed which involved the optimization of a multitude of singular processes and their effective interconnection (Figure 52). The method was optimized to minimize the required manual labor and losses of genetic material throughout the entire procedure. Replicate samples showed little variance and the data generated through sequencing could be validated by densitometric assays rendering the final results highly reliable.

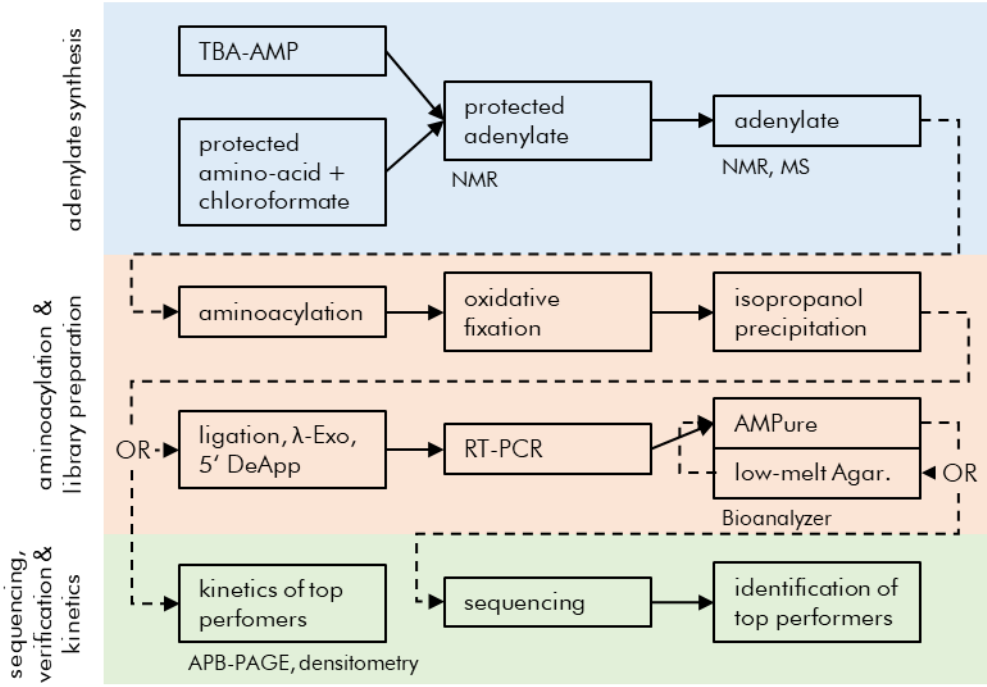


Figure 52. Overview of all experimental steps and their interconnections optimized during this thesis. Where applicable, the corresponding analysis technique was given below boxes.

The self-aminoacylation behavior of an RNA library with structural resemblance to Hopfield folds towards aminoacyl-5′-adenylates derived from different amino acids and L-lysyl-cytidylate could thereby reliably be elucidated. Furthermore, the method is compatible with different inputs of ssRNA or ssDNA as long as they harbor a known constant region at their 5′ terminal region and do not exceed around 125 nt in length. Instead of adenylates and cytidylates, other forms of activated amino acids could also be used.

The employed Hopfield fold RNA library did not fully resemble originally proposed Hopfield folds derived from the sequences of extant tRNAs. It exhibited a comparable structure, but its sequence was shorter, highly different and contained no modified nucleotides.

Through the use of this model system, it could be shown that the structure of a Hopfield fold could indeed exhibit self-aminoacylation. From an initial pool of 16,384 different Hopfield folds with randomized anticodon loops, only a small fraction showed self-aminoacylation activity. The most significant influence on which sequences performed best had the adenylate- or cytidylate-moiety, respectively. The sequences enriched through the use of different aminoacyl-5′-adenylates all showed fundamental similarities whereas the sequences enriched through L-lysyl-cytidylate showed notable differences.

Experiments with aminoacyl-5′-adenylates derived either from D- or L-amino acids revealed a sequence-dependent influence on aminoacylation activity that ranged from being not detectable to several orders of magnitude.

The side chains of the employed L-amino acids generally appeared to exert inhibitory effects most likely through steric hindrance that correlated with their mass; the heavier, the lower its aminoacylation rate and potentially the lower the number of compatible Hopfield fold sequences. Exceptions to this tendency could be observed by amino acids whose side chains carried moieties that facilitated attractive interaction with RNA. Lysine harbored an additional amino group that could interact with the phosphate backbone of RNA and form hydrogen bonds. It was the overall best performing amino acid that resulted in both the highest yields and specificity. The most similar amino acids that were tested were leucine and isoleucine which did not harbor additional amino groups and resulted in significantly lower yields and reaction velocities. Tyrosine harbored a hydroxy group that could also form hydrogen bonds. In a single experiment done with one specific Hopfield fold sequence, it resulted in around three-fold higher aminoacylation yield than its most similar amino acid tested phenylalanine which in comparison lacks only this hydroxy group.

One Hopfield fold sequence could be identified that could realize self-aminoacylation with aminoacyl-5′-adenylates mostly independent from the amino

acid they were generated from. This sequence was the least specific and overall most reactive one. Within L-Lys-AMP, L-Ala-AMP, and Gly-AMP, a sequence could be found that reacted specifically with L-Lys-AMP and another one that reacted specifically with L-Ala-AMP albeit with markedly decreased velocity. None of the identified sequences with outstanding reactivity and specificity harbored anticodons in their anticodon loops that matched with extant anticodons of the amino acid they reacted most readily with. In addition, it could be shown that not only the single stranded anticodon loop but also the double stranded region at the 3′ end as well as the distant 5′ region of the tested Hopfield folds critically influenced aminoacylation activity. This led to the theory of a "third genetic code" that summarizes all structural elements of a potential tRNA precursor that were necessary to realize specific self-aminoacylation.

Kinetic analyses revealed that for some pairs of sequences and aminoacyl-5′-adenylates, the reaction mechanism resembled a second order reaction. These reactions as well as the hydrolysis of adenylates could successfully be modeled and reaction constants be extracted. Other combinations appeared to result in more complex reaction mechanisms that could not be fully elucidated. It could also be shown that Hopfield folds could effectively exert self-aminoacylation at sub millimolar concentrations of aminoacyl-5′-adenylates and that the specificity of Hopfield fold sequences towards one amino acid appeared to increase with decreasing concentrations of activated amino acid.

The employed pH exerted major influence on the stability of the adenylates, aminoacylation yields, and stability of aminoacylated RNA. An optimum was found at precisely pH 5. At higher pH values of up to 7, stabilities and yields markedly declined and the maximum yield was reached faster. In general, the stabilities of aminoacyl-5′-adenylates, aminoacylated Hopfield folds, and aminoacylated extant tRNAs correlated with each other and were dependent from the amino acid side chain.

In principle, this study could validate Hopfield's hypothesis as the Hopfield folds tested herein exhibited the main postulated feature of sequence-dependent self-aminoacylation. However, this study left space for optimization in several regards as discussed in the following. These issues had to be addressed before Hopfield's hypothesis could be further approved or rejected.

4.2 Amino acid activation

As activated amino acids, primarily aminoacyl-5′-adenylates and L-lysyl-5′-cytidylate were used. I was able to improve the original procedure from around

50% yield to near-quantitative yield, primarily through the thorough use of anhydrous conditions. The most substantial drawback of this technique was the deprotection step involving trifluoroacetic acid. Up to this point, most products were obtained with near-quantitative yields, but they hydrolyzed to varying degrees during deprotection. For Lys, Ala, and Gly, these losses were acceptable; however, in the case of Met and Asn, all of the adenylate consistently hydrolyzed. Therefore, it appeared most promising to improve this deprotection step. Strategies beyond the experiments conducted with HCl in dioxane and Reagent H may be explored.

In addition, more amino acids, including some beyond the proteinogenic set with appropriate protecting groups, could be investigated for the synthesis of adenylates, cytidylates, guanylates, thymidylates, or uridylates. Different stereoisomers, such as L- and D-amino acids or L- and D-ribose or deoxyribose, could also be used in future experiments. Regardless of their prebiotic or biotic relevance, such a collection of activated amino acids could contribute to a deeper understanding of the interactions between activated amino acids and RNA.

Beyond adenylates and their described analogs, other activation strategies, such as oxazolones as employed by Chen and coworkers [90], aminoacyl ethyl phosphates [136,137], aminoacyl-benzotriazoles [138], imidazolides [139], phosphoramidate-linked hybrids [140], or in-situ activation using carbodiimides, could also be further explored.

4.3 Aminoacylation

During the course of this thesis, I explored only a limited set of aminoacylation conditions and found the pH to be most critical. Precisely pH 5.0 repeatedly resulted in highest aminoacylation yield even in comparison with as little deviation as pH 4.5 and 5.5. At pH 6, the yield was very low and at any other pH beyond this, no aminoacylation yield was detectable. In contrast to that, the aminoacylation reactions conducted by both Yarus and coworkers [78] as well as Chen and coworkers [90], were adjusted to pH 7. In addition to that, their approaches also contained a larger variety of ions. The reaction buffer that was developed herein only contained sodium acetate and magnesium chloride but theirs also included potassium, calcium, manganese, copper, zink as well as sulfate. These and possibly more ions may therefore be investigated in future experiments. The addition of insoluble materials may also be considered.

All aminoacylation reactions conducted herein employed the singular addition of a relatively large amount of activated amino acid and sequences were selected that could realize the highest yield after a relatively short incubation time. In a prebiotic

setting, it appears more plausible that activated amino acids were supplied at a relatively constant rate over longer periods of time yet at lower concentrations. Another selection of Hopfield folds may be conducted under these circumstances and possibly at near neutral pH. This may lead to the enrichment of more selective Hopfield folds with minimized hydrolysis rates which would fit better to Hopfield's hypothesis unlike the selected Hopfield folds herein that appeared to exhibit maximized aminoacylation rates.

Also, all aminoacylation experiments were conducted at ambient temperature. Further studies may explore the influence of higher or lower temperatures on the stability of activated amino acids as well as on the aminoacylation reaction and hydrolytic stability of aminoacylated RNA sequences. Such variations solely in reaction temperature may eventually be used for van't Hoff and Eyring analyses to separate enthalpic and entropic effects [141].

The experiments conducted in this thesis combined either a multitude of sequences with a single activated amino acid or a single sequence with a single activated amino acid. In a future experiment, aminoacylation reactions may be conducted with a single sequence like the L-Lys-AMP specific sequence and a multitude of activate amino acids like L-Lys-AMP, L-Ala-AMP, and Gly-AMP. After the aminoacylation reaction, unreacted adenylates and unbound amino acids may be separated from the RNA. Afterwards, a fraction of the RNA may undergo oxidative fixation and APB-PAGE densitometry to reveal the overall turnover but the main part of the aminoacylation reaction may be deacylated and the released amino acids may be identified and quantified. The results of such experiments may be able to further validate the specificity of certain sequences as in prebiotic scenarios it also appears likely that several different activated amino acids were present simultaneously. This apporach may also reveal the extent of internal 2′ aminoacylation if stoichiometries other than 1:1 between RNA and amino acid were revealed after aminoacylation.

4.4 Sequencing

The entire library preparation process and subsequent data processing appeared highly optimized in my opinion. I conducted the entire procedure with about 100 samples and lost only very few of them. Yet, the most critical step may be ligation. If future aminoacylation experiments reach considerably higher yields than I could using the Hopfield fold RNA library, the ligation may have to be rescaled to be able to ligate all molecules of interest as it is currently limited to a maximum of 14% aminoacylation yield. The following reverse transcription should also work reliably as long as the primer binding site on the DNA ligation adaptor does not exhibit

stable secondary structures. In future experiments, the thermocycles conducted in the final PCR may be increased. This may be beneficial for very low yielding aminoacylation reactions like the ones conducted with the Hopfield fold RNA library and adenylates derived from amino acids other than Lys, Ala, Gly, Val, or Ile.

Even though it was planned from the beginning, functional kinetic sequencing could not be implemented. The techniques developed were able to produce reliable relative data on how reactive every sequence was but establishing a way of normalization failed which would have allowed to obtain absolute values directly from the sequencing data. In the future, different RNA and DNA sequences may be spiked in between certain steps of the entire procedure. My attempt to use a DNA strand from the beginning which mimicked a ligation product seemed not to lead to the desired results. Maybe an RNA strand spiked in after oxidative fixation will perform better in future experiments. In any case, more experimental work would have to be conducted to realize reliable kinetic sequencing with the procedure presented in this thesis.

4.5 Alternatives to APB-PAGE densitometry

Densitometry as I conducted it in combination with APB-PAGE always seemed to produce reliable results (Figure 41, Figure 43). Whenever the yield was reasonably high like above 10%, the resulting data points could be fitted well with the proposed kinetic model using DynaFit (Chapter 3.3.4). Extreme outliers were seldom and usually the result of particles in or on the gel that caused strong false positive fluorescence. The main disadvantages of this technique were its high labor intensity and unreliability at lower yields at around <10%. A contributing factor to this may have been the higher background that SYBR Gold produced on APB-PAGE compared to PAGE. Using higher amounts of material per lane was no option as this decreased reproducibility of the staining and thereby quantification through densitometry. In extreme cases, too much input can cause the material to result in a smear over the length of the lane and not in distinguishable bands. Therefore, the low yields of around 2% resulting from aminoacylation reactions with the Hopfield fold RNA library could never be directly visualized.

To overcome these issues, I devised an approach using an HPLC method inspired by the work of Yarus and coworkers who used NHS esters to attach a hydrophobic tag to their aminoacylated RNAs [78]. This NHS derivatization was though conducted at elevated pH which enhanced hydrolysis. I therefore thought of an approach which facilitated a column material that carried immobilized boronate groups analogous to APB-PAGE. Combined with an automatic sample injector, a large number of samples could be measured automatically, a higher amount of material

was supposed to be injectable without causing smear, and I expected the detection to be more sensitive as no gel matrix or similar material was part of HPLC detectors. If boronate affinity chromatography turned out to be unfeasible, derivatization techniques similar to the aforementioned NHS coupling may be developed that worked at acidic pH. In the case of oxazolones, hydrophobic or other tags may directly be introduced through aminoacylation as in oxazolones, the amino group of the amino acid had to be derivatized anyways to facilitate synthesis.

4.6 Towards coded peptide synthesis

Realizing self-aminoacylation of a tRNA precursor actually constituted just half of the overarching question about the stereochemical origin of the genetic code. The other half was that such a precursor also had to be able to contribute to coded peptide synthesis. This issue was not addressed in this thesis, but future projects may do so by developing RNA sequences that could aminoacylate themselves through the use of their anticodon as chemical sensor and by changing their structure were also able to enter a proto-ribosome and contribute to coded peptide synthesis.

In this regard, several other aspects may be of interest to be investigated. As it could be shown that the double stranded region near the 3' of a Hopfield fold could significantly affect its aminoacylation behavior. The highly conserved CCA end of tRNAs may therefore have a stereochemical origin just as well as the genetic code itself might have.

Modified, non-canonical or completely artificial nucleotides may be incorporated into the structure of potential tRNA precursors as modified nucleotides commonly contribute to the structure and interactions of extant tRNAs.

So far, the Hopfield fold as tRNA precursor contained just one structural element, namely a stem-loop. To incorporate other elements may have insightful effects as RNA could form a multitude of different structures (Figure 53).

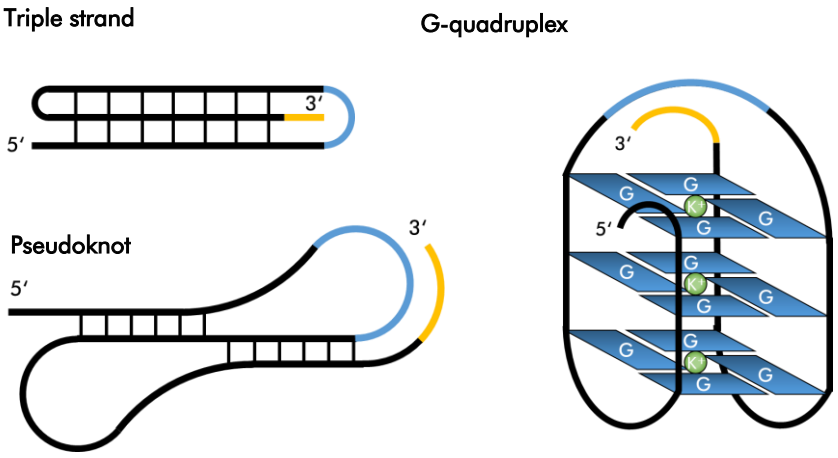


Figure 53. Three structural elements that RNA can regularly form. If these structures were to represent Hopfield folds as utilized in this thesis, the CCA 3′ end (yellow) could be located in close proximity to a single stranded stretch representing an anticodon loop (blue).

Such structures or other interactions may also be used in multi molecule constructs similar to the optimized aminoacylating ribozyme by Yarus and coworkers that could work in trans [83] or like flexizymes which were of widespread use to aminoacylate tRNAs with non-canonical amino acids [142,143].

Multi molecule interactions could also be used to screen for peptidyl transfer centers that could act as proto-ribosomes. In this regard, SELEX and RNA origami may be combined to screen for an entire system of different RNAs that were able to realize coded peptide synthesis without involvement of proteins.

Lastly, short polypeptides could form under prebiotically plausible conditions. Even though such events resulted in random sequences, already short polypeptides could possess catalytic activities. Including such polypeptides in the selection process of a proto translation machinery might have beneficial effects [144].

5 Materials and methods

5.1 Materials

5.1.1 Chemicals and solvents

Chemical	Supplier
Acryloylaminophenyl boronic acid (APB)	in-house preparation
Ammonium persulfate (APS)	Merck KGaA
Boric acid	Thermo Fisher Scientific
Dithiothreitol (DTT)	Merck KGaA
Ethylenediaminetetraacetic aicd (EDTA)	Merck KGaA
Isobutylchloroformate	Merck KGaA
Magnesium chloride	Merck KGaA; Thermo Fisher Scientific
Na-HEPES (N-2-hydroxyethylpiperazine-N'-2-ethanesulfonic acid, sodium)	Merck KGaA
Sodium periodate	Merck KGaA
Tetrabutylammonium hydroxide	Merck KGaA
Tetramethylethylenediamine (TEMED)	Merck KGaA
Tributylamine	Merck KGaA
Trizma-base	Merck KGaA
Urea	Merck KGaA

Solvent	Supplier
1,4-Dioxane (anhydrous with AcrosSeal)	Acros Organics
Acetic acid	Honeywell Riedel-De Haen
Deuterated solvent D_2O	Eurisotop
Deuterated solvent DMSO-d6	Eurisotop
Dichloromethane	Thermo Fisher Scientific
Diethyl ether	Honeywell Riedel-De Haen
Dimethylformamide (DMF, anhydrous with AcrosSeal)	Acros Organics

Ethanol	Zentralbereich Neuenheimer Feld
Ethanol (absolute)	Sigma-Aldrich by Merck
Ethylene glycol	Fluka Chemicals
Formamide	Sigma-Aldrich by Merck
Isopropanol	Sigma-Aldrich by Merck
Trifluoroacetic acid	Sigma-Aldrich by Merck

Boc-protected amino acid	Supplier
Boc-L-Lys(Boc)-OH	Carbolution
Boc-D-Lys(Boc)-OH	Carbolution
Boc-L-Ala-OH	Carbolution
Boc-D-Ala-OH	Carbolution
Boc-Gly-OH	Carbolution
Boc-Ile-OH	Carbolution
Boc-Val-OH	Carbolution
Boc-Asn(Xan)-OH	Carbolution
Boc-Met-OH	Carbolution
Boc-Phe-OH	Carbolution
Boc-Tyr-OH	Carbolution
Boc-Ile-OH	Carbolution
Boc-Pro-OH	Carbolution
Boc-His-OH	Carbolution
Boc-Trp-OH	Carbolution
Boc-Thr-OH	Carbolution
Boc-Ser-OH	Carbolution

5.1.2 Bioreagents and enzymes

Bioreagent	Supplier
Adenosine 5'-monophosphate monohydrate	Merck KGaA
Adenosine triphosphate (ATP)	Merck KGaA
Deoxyadenosine triphosphate (dATP)	Rapidozym
Deoxycytidine triphosphate (dCTP)	Rapidozym

Deoxyguanosine triphosphate (dGTP)	Rapidozym
Deoxythymidine triphosphate (dTTP)	Rapidozym
SYBR™ Gold Nucleic Acid Gel Stain (10,000X Concentrate in DMSO)	Invitrogen by Thermo Fisher Scientific
Cytidine 5'-monophosphate monohydrate	Thermo Fisher Scientific
Glycogen, RNA grade	Thermo Fisher Scientific
5-DBCO-PEG4-dUTP	Jena Bioscience GmbH
Ethidium bromide solution 0.025%	Carl Roth
Bromophenol blue	Merck KGaA
Xylene cyanol	Merck KGaA
Agarose	Sigma-Aldrich by Merck
UltraPure™ Low Melting Point Agarose	Thermo Fisher Scientific
50% PEG 8000	New England Biolabs

Enzyme	Supplier
Q5 Hot-Start DNA Polymerase	New England Biolabs
SuperScript IV Reverse Transcriptase	Invitrogen by Thermo Fisher Scientific
T4 RNA Ligase 2, truncated KQ	New England Biolabs
5'-Deadenylase	New England Biolabs
Lambda Exonuclease	New England Biolabs
RecJf	New England Biolabs
T7 RNA Polymerase	in-house preparation
GoTaq® DNA Polymerase	Promega
Phusion® High-Fidelity DNA Polymerase	New England Biolabs
Taq DNA Polymerase	in-house preparation

5.1.3 Buffers and kits

Commercial buffers	Supplier
Q5 reaction buffer (5x)	New England Biolabs
ROTI® Aqua-P/C/I (for RNA extraction, phenol/chloroform/isoamyl alcohol 25:24:1, pH 4.5–5.0)	Carl Roth
ROTIPHORESE® Sequencing gel concentrate (25%, 19:1)	Carl Roth

SSIV buffer (5x)	Thermo Fisher Scientific
T4 RNA Ligase Reaction Buffer	New England Biolabs
TriTrack DNA Loading Dye (6X)	Thermo Fisher Scientific
NEBuffer™ 2	New England Biolabs
5X Green GoTaq® Reaction Buffer	Promega
5X Colorless GoTaq® Reaction Buffer	Promega
MgCl2 solution	New England Biolabs
DMSO	New England Biolabs
Phusion® HF Buffer Pack	New England Biolabs
Phusion® GC Buffer Pack	New England Biolabs
Lambda Exonuclease Reaction Buffer	New England Biolabs

Self-prepared buffers	Composition
Acetate buffer 1 M	Sodium acetate/acetic acid (1 M, pH 4.5, 5 or 5.5)
APB-PAGE loading buffer	2x HAE buffer, urea (8 M), bromophenol blue and xylene cyanol dye
Sodium acetate solution for precipitation	Sodium acetate/acetic acid (3 M, pH 5.2)
PA gel diluter	H2O with 50% (w/v) urea (8.3 M)
TBE buffer (10x)	Tris-borate (1 M, pH 8.3), EDTA (20 mM)
Phosphate buffer 1 M	Sodium phosphate/phosphoric acid (1 M, pH 1, 2 or 3)
MES buffer 1 M	MES/NaOH (1 M, pH 6 or 7)
Deacylation buffer	Tris-HCl (1 M, pH 9), EDTA (50 mM)

Commercial Kit	Supplier
Bioanalyzer High Sensitivity DNA Analysis Kit	Agilent Technologies
G-CAPSULE	Merck KGaA
QIAquick Gel Extraction Kit	QIAGEN
AMPure XP Bead-Based Reagent	Beckman Coulter
Wizard® SV Gel and PCR Clean-Up System	Promega
Size Exclusion Chromatography Kit	Bio-Rad

5.1.4 Oligonucleotides

If not mentioned otherwise, the following oligonucleotides depict DNA strands. All oligonucleotides were purchased from Integrated DNA Technologies, Inc.

All i7 indices were adapted from NEBNext Adaptors and Primers Set 1 for Illumina [115].

Oligonucleotide	Sequence 5' to 3'
Hopfield fold RNA library	ACUGGAGUUCAGACGUGUGCUCUUCCGAUC UNNNNNNNGUGCGUUCUAAUGAACGCACCA
8N Ligation Adaptor	/5Phos/NNNNNNNNAGATCGGAAGAGCGTCGT GTAGGGAAAGAGTGT/3SpC3/
C8N Ligation Adaptor	/5Phos/CNNNNNNNNAGATCGGAAGAGCGTCG TGTAGGGAAAGAGTGT/3SpC3/
10N Ligation Adaptor	/5Phos/NNNNNNNNNNAGATCGGAAGAGCGTC GTGTAGGGAAAGAGTGT/3SpC3/
11N Ligation Adaptor	/5Phos/NNNNNNNNNNNAGATCGGAAGAGCG TCGTGTAGGGAAAGAGTGT/3SpC3/
12N Ligation Adaptor	/5Phos/NNNNNNNNNNNNAGATCGGAAGAGC GTCGTGTAGGGAAAGAGTGT/3SpC3/
P5 primer, no i5 index	AATGATACGGCGACCACCGAGATCTACACTCTTT CCCTACACGACG
P5 primer, i5 index #1	AATGATACGGCGACCACCGAGATCTACACTAGT GAACACTCTTTCCCTACACGACG
P5 primer, i5 index #2	AATGATACGGCGACCACCGAGATCTACACACA GATACACTCTTTCCCTACACGACG
P5 primer, i5 index #3	AATGATACGGCGACCACCGAGATCTACACCGC TTGACACTCTTTCCCTACACGACG
P5 primer, i5 index #4	AATGATACGGCGACCACCGAGATCTACACGTTA TAACACTCTTTCCCTACACGACG
P5 primer, i5 index #5	AATGATACGGCGACCACCGAGATCTACACGGA CGTACACTCTTTCCCTACACGACG
P5 primer, i5 index #6	AATGATACGGCGACCACCGAGATCTACACTTCG AGACACTCTTTCCCTACACGACG
P5 primer, i5 index #7	AATGATACGGCGACCACCGAGATCTACACTACG TCACACTCTTTCCCTACACGACG
P5 primer, i5 index #8	AATGATACGGCGACCACCGAGATCTACACCTG CATACACTCTTTCCCTACACGACG
P5 primer, i5 index #9	AATGATACGGCGACCACCGAGATCTACACTATA CGACACTCTTTCCCTACACGACG

P5 primer, i5 index #10	AATGATACGGCGACCACCGAGATCTACACGTC GACACACTCTTTCCCTACACGACG
P5 primer, i5 index #11	AATGATACGGCGACCACCGAGATCTACACAGA CTAACACTCTTTCCCTACACGACG
P5 primer, i5 index #12	AATGATACGGCGACCACCGAGATCTACACGCG TCTACACTCTTTCCCTACACGACG
P7 primer, i7 index #1	CAAGCAGAAGACGGCATACGAGATCGTGATGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #2	CAAGCAGAAGACGGCATACGAGATACATCGGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #3	CAAGCAGAAGACGGCATACGAGATGCCTAAGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #4	CAAGCAGAAGACGGCATACGAGATTGGTCAGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #5	CAAGCAGAAGACGGCATACGAGATCACTGTGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #6	CAAGCAGAAGACGGCATACGAGATATTGGCGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #7	CAAGCAGAAGACGGCATACGAGATGATCTGGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #8	CAAGCAGAAGACGGCATACGAGATTCAAGTGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #9	CAAGCAGAAGACGGCATACGAGATCTGATCGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #10	CAAGCAGAAGACGGCATACGAGATAAGCTAGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #11	CAAGCAGAAGACGGCATACGAGATGTAGCCGT GACTGGAGTTCAGACGTGTGC
P7 primer, i7 index #12	CAAGCAGAAGACGGCATACGAGATTACAAGGT GACTGGAGTTCAGACGTGTGC
Sutherland RNA acceptor	CAGAAGACGGCAUACGAGAUUCNNNUUCCA
Sutherland DNA blocking strand	ATCTCGTATGCCGTCTT/3spc3/
Sutherland RNA donor	NNNga
Sutherland P7 primer	CAAGCAGAAGACGGCATACGAGAT

5.1.5 Devices and other equipment

Device	Supplier
Agilent 2100 Bioanalyzer System	Agilent Technologies
ALPHA 1-4 LD plus Lyophilizer	Christ
AlphaImager 2200	Alpha Innotech
Concentrator 5301	Eppendorf
Concentrator Plus	Eppendorf
Electrophoresis Gel Chamber Mini	LTF Labortechnik
Eletrophoresis Power Supply – EPS 301/3501 XL	Amersham Pharmacia Biotech
Magnetic stirrer MR 3001 K	Heidolph
Mercury Plus 500 MHz Spectrometer	Varian
micrOTOF QII ESI Mass Spectrometry System	Bruker
MIKRO 120 Centrifuge	Hettich
MIKRO 185 Centrifuge	Hettich
MIKRO 200R Centrifuge	Hettich
MilliQ QGard 2 System	Merck Millipore
Mini-PROTEAN Tetra System	Bio-Rad Laboratories
Multifuge 4KR	Heraeus
NanoDrop One Spectrophotometer	Thermo Fisher Scientific
NextSeq 550 System	Illumina
T-100 Thermal Cycler	Bio-Rad Laboratories
Thermomixer Comfort	Eppendorf
Typhoon FLA 9500 Biomolecular Imager	GE Healthcare
ABS 80-4 Analytical Balance	KERN & SOHN GmbH
Puriflash 420	Interchim

Other equipment	Supplier
DNA LoBind® Tubes, 1.5 mL	Eppendorf
Illustra NAP-5 columns	GE Healthcare
Magnetic Separation Rack, 12-Tube	New England Biolabs
Magnetic Separator – PCR Strip, SMARTer-Seq	TaKaRa
PF-30C18AQ-F004 chromatography column	Interchim

5.1.6 Software and web applications

Software	Supplier
Agilent 2100 Bioanalyzer 2100 Expert Software	Agilent Technologies
ChemDraw Professional	Perkin Elmer Informatics
EndNote 20	Clarivate
ImageQuant	GE Healthcare
MestReNova v9.0.1-13254	Mestrelab Research S.L.
Microsoft Office 2016	Microsoft
Typhoon FLA 9500 Control Software	GE Healthcare

Web applications	Supplier
NEBioCalculator	New England Biolabs
NEB Tm Calculator	New England Biolabs
Thermo Tm calculator	Thermo Fisher Scientific
Thermo multiple primer analyzer	Thermo Fisher Scientific

5.2 Methods

5.2.1 Synthesis of aminoacyl-5′-adenylates & -cytidylates

Preparation of tetrabutylammonium 5′ adenosine monophosphate (TBA-AMP) and 5′ cytosine monophosphate (TBA-CMP)

TBA-AMP and TBA-CMP were prepared in batches; the following procedures describe the preparation of 5 aliquots which was enough for 5 to 10 reactions of regular size. The procedure could be upscaled arbitrarily.

First, 0.434 g (0.00125 mol) of 5′ adenosine monophosphate monohydrate (AMP) was weighed into a 15 ml falcon tube. Afterwards, 1 ml of nuclease-free water and 1.3 ml of a 40% aqueous tributyl ammonium hydroxide (TBA) solution were added. The solution usually became clear after ~3 min upon occasional agitation. At this point, the pH should be neutral; if it was not, it could be adjusted through the addition of additional TBA solution. The solution was frozen at -80 °C or in liquid nitrogen and was subjected to lyophilization for no less than a week. Once it was completely dry, the material became crystalline. It was eventually dissolved in 2.5 ml of dimethylformamide (DMF) which regularly took more than 30 min and heating to at least 37 °C to finish; using an ultrasonic bath could sometimes speed up the dissolution. The solution was distributed into aliquots of ~550 µl into 1.5 ml reaction tubes and stored at -20 °C to -80 °C.

Preparation of N-tert-butoxycarbonyl amino acid aliquotes in dioxane

Per reaction, 0.0005 mol of N-t-Boc protected amino acid were used; the table below shows the according masses for each used amino acid. Usually, at least enough material for 3 reactions was prepared at once. The material was given into a 15 ml falcon tube and was submitted to lyophilization for at least overnight. Afterwards, the material was dissolved in 500 µl dioxane per reaction; in some cases, it was necessary to heat the sample to 37 °C to facilitate complete dissolution. The solution was eventually distributed into aliquots of ~550 µl into 2 ml reaction tubes and stored at -20 °C.

In the case of Boc-Asn(Xan)-OH and cytidylate synthesis, the protected amino acid was instead dissolved in DMF as otherwise precipitation occurred.

Preparation of N-tert-butoxycarbonyl amino acid adenylate (N-t-Boc amino acid-5'-AMP)

Different conditions were found to be ideal to maximize the yield of this reaction depending on which amino acid is being used. Refer to Table 5 for further details. Most reactions were conducted using one aliquot of the desired N-t-boc protected amino acid described above but this reaction could also be successfully conducted 5-fold upscaled. First, the protected amino acid in dioxane had to be thawed and brought to the indicated temperature for the 1st incubation step (Table 5). Afterwards, 0.001 mol (237 µl) of tri-n-butylamine and 0.0006 mol (78 µl) of isobutyl chloroformate were added in this order. The reaction was mixed thoroughly and incubated at the indicated temperature and duration for the 1st incubation (Table 5). In the next step, the TBA-AMP was added; originally, 0.5 mmol of amino acid were combined with 0.25 mmol of AMP (i.e. one aliquot of each) but using 0.5 mmol of amino acid and 0.125 mmol of AMP (i.e. half an aliquot and 225 µl DMF) markedly improved the purity which the adenylate could be obtained with. The latter was therefore regularly employed during later stages of this thesis. After the addition of the TBA-AMP, the reaction entered the 2nd incubation; the according temperature and duration is listed in Table 5. Some amino acids required a 3rd stage of incubation which was solely a change in incubation temperature according to Table 5.

Amino acid	Molecular weight	Mass per reaction	1st incubation	2nd incubation	3rd incubation
Boc-Lys(Boc)-OH	346.42 g/mol	173.2 mg/rxn	RT, 30 min	RT, 4 h	-
Boc-Ala-OH	189.21 g/mol	94.6 mg/rxn	RT, 30 min	RT, 4 h	-
Boc-Gly-OH	175.18 g/mol	87.6 mg/rxn	RT, 30 min	RT, 4 h	-
Boc-Val-OH	217.26 g/mol	108.6 mg/rxn	RT, 1 h	RT, 6 h	-
Boc-Ile-OH	240.30 g/mol	120.2 mg/rxn	RT, 1 h	RT, 6 h	-
Boc-Tyr-OH	281.30 g/mol	140.7 mg/rxn	RT, 30 min	RT, 4 h	-
Boc-Pro-OH	215.25 g/mol	107.6 mg/rxn	RT, 30 min	RT, 4 h	-
Boc-Phe-OH	265.30 g/mol	132.7 mg/rxn	RT, 30 min	RT, 4 h	-
Boc-Leu-OH	231.29 g/mol	115.6 mg/rxn	RT, 30 min	RT, 4 h	-
Boc-Trp(Boc)-OH	404.46 g/mol	202.2 mg/rxn	4 °C, 30 min	RT, 4 h	-
Boc-Met-OH	249.33 g/mol	124.7 mg/rxn	4 °C, 30 min	4 °C, 2 h	RT, 2 h
Boc-Asn(Xan)-OH	412.44 g/mol	206.2 mg/rxn	4 °C, 30 min	4 °C, 2 h	RT, 2 h
Boc-His-OH	497.58 g/mol	248.8 mg/rxn	4 °C, 30 min	4 °C, 2 h	RT, 2 h
Boc-Ser-OH	223.22 g/mol	111.6 mg/rxn	N/A	N/A	N/A
Boc-Thr-OH	219.23 g/mol	109.6 mg/rxn	N/A	N/A	N/A

Table 5. Ideal conditions of adenylate synthesis for each amino acid used. In the case of Lysine and Alanine, both enantiomeres behaved identically; other enantiomeres were not tested. Per reaction (rxn), 0.0005 mol of protected amino acid were used. In most cases, no third incubation step was employed; RT = room temperature = ~25 °C. For the last 3 amino acids, no ideal conditions could be found yet. Refer to Table 2 for expectable yields.

After all incubations were finished, the reaction was poured into a 15 ml falcon tube containing 12 ml of diethyl ether which caused the reaction to precipitate. The precipitated material was collected at the bottom of the tube through centrifugation

at room temperature and ~1500 g for 1 min. The supernatant was decanted, replaced by 7 ml fresh diethyl ether and the material was resuspended as thoroughly as possible. Centrifugation was repeated as previously, and the supernatant was discarded again. The tube was then inserted into a vacuum concentrator with the lid loosely attached; a program for solvents with high vapor pressure was applied for 5 to 10 min. Afterwards, the caps were removed completely, and the same program was let run for another 5 to 10 min. Eventually, the tube including the material was subjected to lyophilization for at least overnight. After lyophilization was finished, a sample of 1 to 5 mg was usually taken to conduct ^{31}P-NMR measurement in deuterated DMSO to monitor the yield of the reaction.

Deprotection – preparation of pure amino acid adenylate

The desired amount of N-t-boc protected aminoacyl-5'-adenylate or –cytidylate was dissolved in 1 ml trifluoroacetic acid and incubated at room temperature for no longer than 15 min. Analogously to the previous procedure, 12 ml of diethyl ether were added after the reaction was finished to precipitate the material. It was then centrifuged at room temperature and ~1500 g for 1 min and the supernatant was discarded. This washing step was repeated two times with 7 ml of dichloromethane and one last time with 7 ml of diethyl ether. The sample was again subjected vacuum concentration with the cap loosely attached for 5 to 10 min using a program suited for solvents with vapor pressure followed by another 5 to 10 min without cap under the same conditions. The material was again subjected to lyophilization for at least overnight which regardless of which amino acid was used, resulted in a white crystalline powder that was transferred into 1.5 or 2 ml reaction tubes. The powder was homogenized by grinding with a spatula. At this point, a 1 to 2 mg sample was taken of the material for ^{31}P-NMR measurement in D_2O to monitor the yield of the reaction.

5.2.2 Chromatographic purification

A PF-30C18AQ-F004 column was used in a reverse phase chromatography machine which was first purged with water at a flow rate of 50 ml/min for ~1 min. The column was inserted at a flow rate of 5 ml/min to prevent trapping air; this flow rate was maintained for all steps as long as the column was attached. The column was then purged with a mixture of 90% acetonitrile and 10% water for 5 min followed by an equilibration step using an aqueous 0.1% dilution of trifluoroacetic acid (TFA) for 5 min. Up to 100 mg of Ile-AMP could be dissolved in

0.5 to 1 ml of aqueous 0.1% TFA and injected into the column. For purification, a flow of aqueous 0.1% TFA was maintained for 5 min followed by a gradient towards 60% water and 40% acetonitrile both containing 0.1% TFA over the course of 25 min. Fractions were collected automatically that surpassed a threshold in their absorbance at 260 nm which resulted in two peaks with the first one being derived from free AMP and the second one from aminoacyl-5′-adenylates. After the purification was finished, the column was purged with a mixture of 50% water and 50% acetonitrile (without TFA) for ~1 min. The column was taken out of the machine at this point and stored at 4 °C until further use. Lastly, the machine was purged with 100% ethanol at a flow rate of 50 ml/min for 1 min.

5.2.3 Aminoacylation & NMR measurements

All aminoacylation experiments were conducted with 1 µg RNA and 300 µg of aminoacyl-5′-adenylates or L-lysyl-5′-cytidylate in 20 µl. The concentration of the buffer reagent was 500 mM and in the case of pH 1 to 3, phosphoric acid was used; for pH 4.5 to 5.5, acetic acid; and for pH 6 and 7, 2-(N-morpholino)ethanesulfonic acid (MES). All reactions also contained 5 mM $MgCl_2$.

All reactions were accompanied by at least one ^{31}P-NMR measurement to monitor the quality of the used adenylate or cytidylate. In some experiments, additional measurements were conducted in parallel with the ongoing aminoacylation reaction after defined time intervals to determine hydrolysis rates. All such NMR measurements were prepared in the same buffer as the corresponding aminoacylation experiments but differed in the use of 1 mg of adenylate and no RNA per 20 µl. Just before the measurement, 20 µl of this setup were diluted in 480 µl of D_2O, mixed and transferred into an NMR-tube. Measurements were conducted by the author of this thesis.

5.2.4 Oxidative fixation

Oxidative fixation was regularly conducted using 20 µl of an RNA containing acidic buffer as starting material; usually 500 mM acetic acid buffer pH 4.5 to 5.5 containing 1 µg of RNA. To this material, 20 µl of an aqueous solution of 200 mM sodium periodate was given, mixed, and incubated at room temperature for 3 minutes. Subsequently, 4 µl of ethylene glycol were given to the reaction, mixed thoroughly, and incubated for at least 3 minutes; after the addition of ethylene glycol, the sample could be left at room temperature for up to a few hours without

compromising RNA integrity which was occasionally done during extended aminoacylation experiments. Next, the sample could either be stored at -20 °C for later continuation of the procedure or be submitted to alkaline deacylation directly. For that, 44 µl of 1 M Tris 50 mM EDTA pH 9 were given to the reaction and incubated at 37 °C for 30 min. Afterwards, 1 µl of glycogen (RNA grade, 20 mg/ml), 178 µl nuclease-free water, 30 µl 3 M sodium acetate pH 5.2, and 297 µl isopropanol were given to reaction, mixed thoroughly, and incubated either at room temperature for 30 min or at -20 °C overnight or for several days. The glycogen, water, and sodium acetate solution were regularly mixed in advance; this mix was stored at -20 °C, thawed when needed and 209 µl of this mix were added to the reaction. If the reaction including isopropanol was incubated at -20 °C, a lot of precipitate formed which dissolved again upon heating to room temperature or slightly above. Next, the sample was centrifuged at ~16.000 g at room temperature for 30 minutes which normally resulted in a barely visible pellet. The supernatant was discarded and replaced with 500 µl of 75% (v/v) ethanol diluted with water followed by centrifugation at ~16.000 g at room temperature or 4 °C for 1 minute. The pellet became much better visible at this point and the supernatant was discarded as thoroughly as possible. The pellet was air-dried for 10 minutes and was eventually dissolved in either 12.5 µl nuclease-free water if the sample was to be used for ligation and further sequencing library preparation or 20 µl of nuclease-free water if the sample was to be analyzed via APB-PAGE. In both cases, the sample could be stored at -20 °C for an extended period of time.

5.2.1 Ethanol/isopropanol precipitation

Precipitation reactions were regularly conducted by adding 1 µl of glycogen (RNA grade, 20 mg/ml), 1/10th of the sample's volume of 3 M sodium acetate solution pH 5.2 and either 1 sample volume of isopropanol or 3 sample volumes of ethanol. If during the course of an experiment, glycogen was added once, it was not added in subsequent precipitation reactions. The sample was incubated at -20 °C for at least 30 minutes and subsequently centrifuged at ~16.000 g for 30 minutes at 4 °C. The supernatant was removed and replaced by 500 µl of 75% (v/v) ethanol diluted with water. The supernatant was thoroughly removed again, and the sample was air-dried for 10 min. Eventually, the pellet was dissolved in 10 to 20 µl nuclease-free water (e.g. if reverse transcription was the subsequent step, the pellet was dissolved in 11 µl as this was the maximum input for the employed reverse transcription reaction).

5.2.1 5′ adenylation of DNA ligation adaptors

Method adapted from Hafner et al. [125].

DNA ligation adaptors were ordered harboring a phosphate group at their 5′-terminus and a C3-linker at their 3′-terminus. They were dissolved in as much nuclease-free water to result in a 1 mM stock solution. For the 5′-adenylation reaction, 7.1 mg of adenosine-5′-phosphoimidazolide (ImpA) was dissolved in 90 µl 0.1 M $MgCl_2$ (it was normal for the solution to remain cloudy). 60 µl of the adaptor stock solution was mixed with 60 µl of the ImpA solution and incubated at 50 °C for 1.5 h. Next, 30 µl nuclease-free water and the remaining 30 µl of ImpA solution were given to the reaction and incubated for another 1.5 h at 50 °C. Eventually, 820 µl nuclease-free water were given to the reaction to bring it to 1 ml final volume. The reactions had to be diluted like that to make them compatible with NAP-5 columns which were subsequently used to purify the DNA from side products like adenosine and imidazole through size exclusion chromatography. The NAP-5 columns were operated by gravity flow and fractions of two droplets were collected in Eppendorf tubes. Spectroscopic readout revealed two peaks at 260 nm absorbance over the course of the elution with the first one corresponding to the DNA oligomer and the second one mostly to adenosine. The fractions around the first peak were pooled together. APB-PAGE analysis regularly revealed a yield of 50%; to improve complete digestion of excess ligation adaptor after ligation, 5 µg of the partially 5′-adenylated DNA adaptor were mixed with 5 µl 10X reaction buffer for lambda exonuclease, 1 µl lambda exonuclease ad 50 µl with nuclease-free water. The reaction was incubated at 37 °C for 30 min and afterwards heat inactivated at 75 °C for 10 min. Adenylated DNA ligation adaptors were protected from digestion so this treatment left behind nearly quantitatively 5′-adenylated DNA ligation adaptors which were eventually purified by PCI extraction and ethanol precipitation.

5.2.2 Ligation

The standard ligation procedure for the generation of sequencing libraries was regularly conducted with samples after oxidative fixation or in the case of positive controls, with untreated RNA Hopfield library or a mixture of oxidized and untreated RNA Hopfield library. In either case, the volume of input material was 12.5 µl which 2 µl of 10X T4 RNA ligase reaction buffer, 4 µl of 50% PEG8000, 1 µl of pre-digested DNA ligation adaptor (100 ng/µl) and 0.5 µl of T4 RNA ligase 2 truncated KQ were given to and mixed thoroughly. The reaction was incubated at room temperature for 2 h and could afterwards be stored at -20 °C until further

proceeding. Next, 0.5 µl of 5'-deadenylase were given to the reaction, mixed thoroughly, and incubated at 30 °C for 30 minutes followed by the addition of 0.5 µl lambda exonuclease and incubation at 37 °C for another 30 minutes. Early during the work of this thesis, the sample was submitted to PCI extraction and subsequent ethanol precipitation at this point. As PCI extraction was later found unnecessary during the project, most sample were directly processed by ethanol precipitation.

5.2.1 Lambda Exonuclease

Adenylated DNA ligation adaptors were pre-digested through the use of Lambda Exonuclease by New England Biolabs according to the manufacturer's instructions using the maximum input amount of 5 µg. After he reaction was finished, the ligation adaptors were purified through PCI extraction followed by ethanol precipitation.

5.2.2 PCI extraction

Extraction with phenol (50%), chloroform (48%), and isoamyl alcohol (2%) were solely conducted with 5'-preadenylated DNA ligation adaptors or samples after ligation. As all of these samples contained DNA of interest, exclusively PCI at neutral pH was used. In any case, the sample was mixed with an equal volume of PCI and mixed thoroughly on a vortex shaker. Afterwards, the sample was centrifuged at ~16.000 g at room temperature for 3 minutes. The upper phase was extracted as completely as possible whilst avoiding aspiration of the interphase or lower phase. The upper aqueous phase was subsequently subjected to ethanol precipitation.

5.2.3 Reverse transcription

All reverse transcription reactions were conducted using Super Script IV by Thermo Fisher Scientific. The instructions by the manufacturer were adopted with minor changes: The material to be reverse transcribed was dissolved in 11 µl nuclease-free water to which 1 µl of a dNTP mix (dATP, dCTP, dGTP, dTTP 10 mM each) and 1 µl of 2 µM primer were added and incubated at 65 °C for 5 min followed by 4 °C for at least 1 min. Meanwhile a mix of 4 µl 5X SSIV reaction buffer, 1 µl 0.1 M DTT, 1.5 µl nuclease-free water and 0.5 µl SuperScript IV (200 U/µl) was prepared and

given to sample once the incubations were finished. Directly afterwards, the sample was incubated at 55 °C for 10 min followed by 80 °C for another 10 min. The reaction was collected by brief centrifugation and was regularly used entirely as a template in subsequent PCR.

5.2.4 Polymerase chain reaction

Polymerase chain reactions (PCRs) were regularly conducted using the Q5 DNA polymerase supplied by New England Biolabs combined with the according instructions given by the manufacturer. In most cases, the entire 20 µl of a reverse transcription reaction as described above served as template to which 13.5 µl nuclease-free water, 10 µl 10X Q5 reaction buffer, 2.5 µl forward primer, 2.5 µl reverse primer, 1 µl dNTP mix (dATP, dCTP, dGTP, dTTP 10 mM each), and 0.5 µl Q5 DNA polymerase were given (Table 6).

The initial denaturation was set to 98 °C for 30 sec followed by 4 to 35 cycles of denaturation at 98 °C for 10 sec, alignment at 67 °C for 15 sec and elongation at 72 °C for 10 sec. The final extension step was set to 72 °C for 45 sec with eventual cooling at 4 °C until the sample was withdrawn from the thermocycler. During all PCRs involved in the final form of the ligation-based library preparation, four of the above-mentioned cycles were employed (Table 7).

Component	50 µl rxn	Stock conc.	Final conc.
Nuclease-free H_2O	13.5 µl		
5X buffer	10 µl	5X	1X
Template cDNA	20 µl	e.g. 5 ng/µl	5 ng to 1 µg
For. primer	2.5 µl	10 µM	500 nM
Rev. primer	2.5 µl	10 µM	500 nM
dNTPs	1 µl	10 mM	200 µM
Q5 DNA Polymerase	0.5 µl	2 U/µl	0.02 U/µl

Table 6. Composition of a typical PCR involved in the final form of the ligation-based library preparation.

Step	Temperature	Time
Initial denaturation	98 °C	30 sec
3 – 35 cycles (usually 4)	98 °C 67 °C 72 °C	10 sec 15 sec 10 sec
Final extension	72 °C	45 sec
Cooling	4 °C	hold

Table 7. Thermocycle program as it was regularly used for PCRs involved in the final form of ligation-based library preparation. The annealing temperature of 67 °C was identical for all primers employed.

5.2.1 NaOH digestion

Hydrolysis of RNA using sodium hydroxide was conducted at a final concentration of 0.3 M sodium hydroxide and incubation at 55 °C for 25 min. Equimolar amounts of diluted hydrochloric acid were used to roughly neutralize the reaction followed by ethanol precipitation.

5.2.2 APB-PAGE

In the majority of cases, polyacrylamide gelelectrophoresis (PAGE) was conducted in combination with acryloylphenylboronic acid which was a compound that copolymerized with the acrylamide and interacted with vicinal diols, slowing them down during electrophoresis. In this thesis, the only difference between PAGE and APB-PAGE was the omission or addition of APB to the gel mixture. In any case, all gels were cast using the Mini-PROTEAN system by Bio-Rad which was assembled according to manufacturer's instructions. If densitometry was desired to be conducted with the gels, the 0.75 mm spacer glass plate was chosen; for any other purposes, the 1 mm spacer glass plate was used.

A single gel was prepared according to the following table:

Ingredient/thickness	0.75 mm	1.00 mm
8 M aqueous urea	2.5 ml	3.25 ml
10X HAE	0.5 ml	0.65 ml
25 % AA/BA	2.0 ml	2.6 ml
APB	15 mg	19.5 mg
10 % APS	50 µl	65 µl
TEMED	5 µl	6.5 µl

Table 8. Recipe for the preparation of a single polyacrylamide gel using the Mini-PROTEAN system by Bio-Rad. This recipe resulted in a final concentration of AA/BA of 10% (w/v) and of APB of 0.3% (w/v). APS and TEMED were given to the solution just prior to casting the gel.

If APB was included, the mixture had to be heated to 50 °C for ~10 min to completely dissolve the APB. Before APS and TEMED were given to the mixture to cast the gel, it was cooled back to room temperature again. For densitometric assays, always 50 ng of material was applied per lane and the sample were heated to 80 °C for 2 min beforehand; for other purposes, various amounts were used, and the samples were not denatured. The gels were regularly run at 200 V, max A, max W which resulted in ~5 W per 0.75 mm gel and ~7 W per 1 mm gel. Each run took ~1 h until the bromophenol blue reached the lower edge of the gel which was taken as an indicator for the run to be finished. For any purpose other than densitometry, the gel was stained in 1X SYBR Gold solution in water for ~1 min. In the case of densitometry, the gel was stained in 1X SYBR Gold solution in 1X HAE buffer for 30 to 40 min; this ensured optimal reproducibility and avoided thick bands to not show fluorescence in their central area. No washing steps were necessary before or after staining. Eventually, the gels were imaged using the Typhoon FLA 9500 biomolecular imager. For densitometry, it was always ensured that measurements were conducted with 500 Volt laser intensity and 100 µm pixel size and no more than 50 ng of material were loaded onto each lane.

5.2.3 Electroelution

This technique was employed to effectively purify genetic material from polyacrylamide-based gel snippets using 2% agarose gels in TBE buffer. Analogously to the comb inserted into agarose gels to form the wells for sample

loading, I created another comb-like object made of Styrofoam and cuvettes that could be placed on top of an agarose gel slide while the gel was solidifying to create notches with a square base area of 1 cm by 1 cm. Their height depended on the amount of agarose gel being cast but usually reached ~1 cm, too which resulted in a volume of ~1 cm^3 = ~1 ml. The notches were placed in such a way that an acrylamide gel snippet could be placed inside a regular well and upon electrophoresis, the genetic material traveled into the running buffer within the notch. This process took ~10 min and the genetic material could be obtained in high purity and yield upon ethanol or isopropanol precipitation from the buffer within the notch. Care had to be taken not to fill the electrophoresis chamber too high with running buffer. Ideally, the running buffer touched only the upper edge of the gel leaving the notches disconnected from the running buffer and from one another. This ensured that no genetic material could diffuse out of the notches.

The electrophoresis could be monitored if ethidium bromide was given to the gel by carefully transferring the gel from the electrophoresis chamber into the imager and optionally back again.

5.2.4 Agarose gel electrophoresis

Agarose gel electrophoresis was always conducted using 2% gels in TBE. 1.6 g of either low-melt or regular agarose were given into 80 ml of 1X TBE buffer and the entire vessel was weighed. The solution was heated until it boiled using a microwave oven until the agarose was completely dissolved. After letting the solution cool down a little, the vessel was set to its initial weight using distilled water to compensate for losses during boiling and one drop of 0.025% ethidium bromide solution was added. After mixing thoroughly, the molten gel was poured into a prepared slide with casting aid and a suitable comb was inserted. The samples were mixed with 6X TriTrack agarose loading dye prior to loading onto the gel. Electrophoresis was standardly conducted at a constant 120 V for ~45 min. Eventually, the gel was imaged using the AlphaImager 2200.

5.2.5 Freeze & Squeeze

The standard method I employed for extracting genetic material from agarose gel snippets was the so called *Freeze & Squeeze* method [117,118]. For that, nylon-based spin filters were used which the gel snippet was put into and frozen at -20 °C. While still frozen, the assembly was put into an unrefrigerated centrifuge and centrifuged

at ~16.000 g for 5 min. This caused the agarose to be retained as a rather dry pellet on top of the filter and contained buffer and genetic material being collected below. The genetic material was purified from the buffer by ethanol precipitation. This method was fast, cheap, and simple and gave consistently reasonable results in terms of yield and purity. Smaller samples (<100 ng), especially those that were directly afterwards to be used in sequencing, were purified through agarase digestion as it was superior in yield and purity than this method.

5.2.6 Agarase digestion

Samples of <100 ng that were supposed to undergo sequencing (i.e. mostly final multiplexed DNA libraries) were purified using this technique. For agarose gel electrophoresis, agarose with a lower melting point had to be used; besides that, handling and electrophoresis was identical to regular agarose gel electrophoresis. After electrophoresis, the part of the agarose gel that contained the material of interest was excised and weighed. It was then incubated at 70 °C for ~10 min or until the entire piece was molten followed by incubation at 42 °C for 5 min. While kept at 42 °C, 1 U of agarase per 100 mg of 1% agarose was given to the gel; higher amounts or percentages scaled linearly. The digestion was incubated at 42 °C for 30 min. Ammonium acetate was given to the reaction to reach a final concentration of 2.5 M; the mixture was incubated at 4 °C for 5 min. The reaction was centrifuged at ~16.000 g for 10 min to pellet undigested carbohydrates; the supernatant was transferred into a fresh tube. 2.5 of the sample's volume of ethanol was added, thoroughly mixed and incubated at -20 °C for at least 30 min. Analogous to any ethanol precipitation, the sample was centrifuged at 16.000 g for ~30 min, the supernatant was discarded, 500 µl of 75% ethanol were added, and the vessel was centrifuged at 16.000 g for 1 min. After discarding the supernatant, the pellet was air-dried for 10 min and eventually dissolved in an appropriate volume of nuclease-free water.

5.2.1 AMPure XP beads cleanup

For library preparation, the final DNA library was purified after PCR using AMPure XP magnetic beads. The regular 50 µl of an entire PCR reaction was mixed with 90 µl of well resuspended bead suspension, mixed well and incubated at room temperature for 5 min. Next, the PCR-tube containing the sample was placed on a magnetic rack for around 5 min until all or most of the beads accumulated on the inner wall of the tube facing the magnet. The supernatant was carefully discarded

and replaced by 180 µl 75% (v/v) ethanol diluted with water without extensive mixing and whilst keeping the tube in the rack. The supernatant was removed again, and the entire washing was repeated once. Afterwards, the tube was briefly centrifuged to collect the remaining ethanol at the bottom. The ethanol was removed as thoroughly as possible and the tube was air-dried for 10 min. Finally, the dried beads were resuspended in 15 µl nuclease-free water, put back into the rack, and after separation, the clear supernatant containing the material of interest was transferred into a fresh tube.

5.2.2 In vitro transcription

Template preparation

Two overlapping DNA oligonucleotides were ordered which could hybridize at their 3'-termini. A Q5 PCR approach was used to transform the oligos into fully double stranded DNA that could serve as template for the subsequent in vitro transcription reaction. For that, the recipe described in this thesis was upscaled to 100 µl and the two oligos were treated as primers but with 10-fold higher concentration. The reaction underwent 5 thermal cylces in total before the product was purified using a PCR purification kit provided by Qiagen.

In vitro transcription

Regular in-vitro transcription reactions were prepared according to the following recipe (Table 9):

Reagent	Volume in µl	Stock concentration
Nuclease-free water	335	
PCR template	50	variable
NTPs	10	10 mM each
DTT	5	1 M
10X IVT Buffer	50	10X
T7 RNA polymerase	20	25X

Table 9. Composition of an in-vitro transcription reaction.

All reagents were added in the indicated order and eventually thoroughly mixed by pipetting. The reaction was incubated at 37 °C for 4 h. Afterwards, 2 µl DNAse A were added, mixed, and incubated at 37 °C for another 30 min. The desired RNA was purified through denaturing PAGE as previously described.

5.2.1 VBA scripts

Demultiplexing

The following script handled demultiplexing as described in Chapter 3.1.8

```
Option Explicit
Option Base 1

Sub Demultiplex()

'Application.ScreenUpdating = False
'Application.Calculation = xlCalculationManual

'Specify how many lanes the sequencer posseses
'E.g. iSeq100 = 1; NextSeq 550 = 4
Dim lnLanes As Long: lnLanes = 1

'Specify if i5 index is being read forward or as reverse complement
'E.g. iSeq100, NextSeq550 = reverse complement
Dim bli5RevCom As Boolean: bli5RevCom = True

Dim wsMain As Worksheet: Set wsMain = ThisWorkbook.Worksheets("Command Center")
Dim lnLastRow As Long: lnLastRow = wsMain.Range("B1048576").End(xlUp).Row
Dim lnForCounter As Long
Dim lnForCounter2 As Long

'Counts how many samples are set to "Active" in Column "B"
Dim lnHowMany As Long: lnHowMany = 0
For lnForCounter = 3 To lnLastRow
   If wsMain.Range("B" & lnForCounter).Value = "Active" Then _
      lnHowMany = lnHowMany + 1
Next

If lnHowMany < 2 Then
   MsgBox "No or just one sample is stated as >Active<", _
      vbExclamation + vbOKOnly, "Macro aborted"
```

```
    Exit Sub
End If

'Creates the needed directories
MkDir (ThisWorkbook.Path & "\Demultiplexed fastq data\")
For lnForCounter = 3 To lnLastRow Step 1
    If wsMain.Range("B" & lnForCounter).Value = "Active" Then
        MkDir (ThisWorkbook.Path & "\Demultiplexed fastq data\" & _
            ThisWorkbook.Worksheets("Command Center").Range("C" & lnForCounter) & _
            "\")
    End If
Next
MkDir (ThisWorkbook.Path & "\Demultiplexed fastq data\Still undetermined\")

'Initializing the pointers for .fastq data streaming
Dim objWriteRead1() As Object
ReDim objWriteRead1(lnHowMany)
Dim objWriteIndex1() As Object
ReDim objWriteIndex1(lnHowMany)
Dim objWriteIndex2() As Object
ReDim objWriteIndex2(lnHowMany)

'Like above but only relevant if spike-ins are included
Dim objWriteRead1SPIKE() As Object
ReDim objWriteRead1SPIKE(lnHowMany)
Dim objWriteIndex1SPIKE() As Object
ReDim objWriteIndex1SPIKE(lnHowMany)
Dim objWriteIndex2SPIKE() As Object
ReDim objWriteIndex2SPIKE(lnHowMany)

'Storing the indices of each active sample
Dim strArray() As String
ReDim strArray(lnHowMany, 2)

lnForCounter2 = 1
For lnForCounter = 3 To lnLastRow Step 1
    If wsMain.Range("B" & lnForCounter).Value = "Active" Then
        'i7 index
        strArray(lnForCounter2, 1) = wsMain.Range("W" & lnForCounter).Value
        If bli5RevCom Then
            'i5 index
            strArray(lnForCounter2, 2) = wsMain.Range("AA" & lnForCounter).Value
        Else
            'i5 index
            strArray(lnForCounter2, 2) = wsMain.Range("Z" & lnForCounter).Value
        End If
        lnForCounter2 = lnForCounter2 + 1
```

```
    End If
Next

'Initializing all other needed variables
Dim strName As String
Dim byLane As Byte: byLane = 0
Dim strTrimIndex1 As String
Dim strTrimIndex2 As String
Dim strCurrentChar As String
Dim byNsDetected As Byte
Dim strCheckSeq As String: strCheckSeq = "AGATCGGAAGAG"
Dim lnTotalReads As Long: lnTotalReads = 0

'Arrays of (4) because .fastq uses 4 lines per read
Dim strReadRead1(4) As String
Dim strReadIndex1(4) As String
Dim strReadIndex2(4) As String

Dim strDataSourcePath As String: strDataSourcePath = ThisWorkbook.Path & _
    "\Unpacked fastq data\"
Dim strDataSinkPath As String: strDataSinkPath = ThisWorkbook.Path & _
    "\Demultiplexed fastq data\"

'Needed for text streaming
Dim objFSO As Scripting.FileSystemObject
Set objFSO = New Scripting.FileSystemObject

Dim objReadRead1 As Object
Dim objReadIndex1 As Object
Dim objReadIndex2 As Object

Dim objWriteStillUndeterminedRead1 As Object
Dim objWriteStillUndeterminedIndex1 As Object
Dim objWriteStillUndeterminedIndex2 As Object

NextLane:

byLane = byLane + 1
If byLane > lnLanes Then
    wsMain.Range("C" & lnLastRow + 9).Value = _
        "MayDemultiplex was successfully completed at:"
    wsMain.Range("C" & lnLastRow + 10) = Now
    wsMain.Range("C" & lnLastRow + 12).Value = _
        "And went through a total of reads:"
    wsMain.Range("C" & lnLastRow + 13).Value = lnTotalReads
    Application.ScreenUpdating = True
    Application.Calculation = xlCalculationAutomatic
```

```
    Exit Sub
End If

'Opening the .fastq files of read 1 and index reads 1 & 2
Set objReadRead1 = objFSO.OpenTextFile(strDataSourcePath & _
    "Undetermined_S0_L00" & byLane & "_R1_001.fastq", ForReading)
Set objReadIndex1 = objFSO.OpenTextFile(strDataSourcePath & _
    "Undetermined_S0_L00" & byLane & "_I1_001.fastq", ForReading)
Set objReadIndex2 = objFSO.OpenTextFile(strDataSourcePath & _
    "Undetermined_S0_L00" & byLane & "_I2_001.fastq", ForReading)

'Creating the output files and corresponding data pointers
lnForCounter2 = 1
For lnForCounter = 1 To lnLastRow - 2 Step 1
    If wsMain.Range("B" & lnForCounter + 2).Value = "Active" Then
        strName = wsMain.Range("C" & lnForCounter + 2).Value
        Set objWriteRead1(lnForCounter2) = objFSO.OpenTextFile(strDataSinkPath & _
            strName & "\" & strName & "_L00" & byLane & _
            "_R1.fastq", ForAppending, True)
        Set objWriteIndex1(lnForCounter2) = objFSO.OpenTextFile(strDataSinkPath & _
            strName & "\" & strName & "_L00" & byLane & _
            "_I1.fastq", ForAppending, True)
        Set objWriteIndex2(lnForCounter2) = objFSO.OpenTextFile(strDataSinkPath & _
            strName & "\" & strName & "_L00" & byLane & _
            "_I2.fastq", ForAppending, True)

        'Only relevant if spike-in sequences were used
        Set objWriteRead1SPIKE(lnForCounter2) = _
            objFSO.OpenTextFile(strDataSinkPath & strName & "\" & _
            strName & "_L00" & byLane & "_R1_SPIKE.fastq", ForAppending, True)
        Set objWriteIndex1SPIKE(lnForCounter2) = _
            objFSO.OpenTextFile(strDataSinkPath & strName & "\" & _
            strName & "_L00" & byLane & "_I1_SPIKE.fastq", ForAppending, True)
        Set objWriteIndex2SPIKE(lnForCounter2) = _
            objFSO.OpenTextFile(strDataSinkPath & strName & "\" & _
            strName & "_L00" & byLane & "_I2_SPIKE.fastq", ForAppending, True)

        lnForCounter2 = lnForCounter2 + 1
    End If
Next

'Creating output files and data pointers for reads that match to no searched combination
of indices
Set objWriteStillUndeterminedRead1 = objFSO.OpenTextFile(strDataSinkPath & _
    "Still undetermined\StillUndetermined_L00" & byLane & _
    "_R1.fastq", ForAppending, True)
Set objWriteStillUndeterminedIndex1 = objFSO.OpenTextFile(strDataSinkPath & _
```

```
    "Still undetermined\StillUndetermined_L00" & byLane & _
    "_I1.fastq", ForAppending, True)
Set objWriteStillUndeterminedIndex2 = objFSO.OpenTextFile(strDataSinkPath & _
    "Still undetermined\StillUndetermined_L00" & byLane & _
    "_I2.fastq", ForAppending, True)

NextRead:

lnTotalReads = lnTotalReads + 1

'In a fastq file, the second and from then on every
'forth row contains the actual base calls
If objReadRead1.AtEndOfStream Then GoTo NextLane
For lnForCounter = 1 To 4 Step 1
    strReadRead1(lnForCounter) = objReadRead1.ReadLine
    strReadIndex1(lnForCounter) = objReadIndex1.ReadLine
    strReadIndex2(lnForCounter) = objReadIndex2.ReadLine
    If objReadRead1.AtEndOfStream Then GoTo NextLane
Next

'This block creates strTrimIndex1 (i7) which
'replaces up to one "N" in the sequence read by a "?"
byNsDetected = 0
strTrimIndex1 = vbNullString
For lnForCounter = 1 To 6 Step 1
    strCurrentChar = Mid(strReadIndex1(2), lnForCounter, 1)
    If strCurrentChar = "N" Then
        byNsDetected = byNsDetected + 1

'Here, the tolerance for undetermined bases in the index sequence can be adjusted
        If byNsDetected >= 2 Then GoTo WriteStillUndetermined

        strTrimIndex1 = strTrimIndex1 & "?"
    Else
        strTrimIndex1 = strTrimIndex1 & strCurrentChar
    End If
Next

'This block creates strTrimIndex2 (i5) which recplaces
'up to one "N" in the sequence read by a "?"
byNsDetected = 0
strTrimIndex2 = vbNullString
For lnForCounter = 1 To 6 Step 1
    strCurrentChar = Mid(strReadIndex2(2), lnForCounter, 1)
    If strCurrentChar = "N" Then
        byNsDetected = byNsDetected + 1
```

```
'Here, the tolerance for undetermined bases in the index sequence can be adjusted
        If byNsDetected >= 2 Then GoTo WriteStillUndetermined

        strTrimIndex2 = strTrimIndex2 & "?"
    Else
        strTrimIndex2 = strTrimIndex2 & strCurrentChar
    End If
Next

'The following compares the index reads with
'the expected sequences of each active sample
For lnForCounter = 1 To lnHowMany Step 1
    If strArray(lnForCounter, 1) Like strTrimIndex1 Then        'Compares i7
        If strArray(lnForCounter, 2) Like strTrimIndex2 Then    'Compares i5

            'Identifies sequences derived from DNA spike-in
            If lnStr(1, strReadRead1(2), "AGATCGGAAGAG", vbTextCompare) = _
                40 And Mid(strReadRead1(2), 11, 2) = "CA" Then

                'Writes the currently stored four lines into
                'the corresponding demultiplexed .fasq files
                For lnForCounter2 = 1 To 4 Step 1
                    objWriteRead1SPIKE(lnForCounter).WriteLine _
                        strReadRead1(lnForCounter2)
                    objWriteIndex1SPIKE(lnForCounter).WriteLine _
                        strReadIndex1(lnForCounter2)
                    objWriteIndex2SPIKE(lnForCounter).WriteLine _
                        strReadIndex2(lnForCounter2)
                Next

            Else

                'Writes the currently stored four lines into
                'the corresponding demultiplexed .fasq files
                For lnForCounter2 = 1 To 4 Step 1
                    objWriteRead1(lnForCounter).WriteLine _
                        strReadRead1(lnForCounter2)
                    objWriteIndex1(lnForCounter).WriteLine _
                        strReadIndex1(lnForCounter2)
                    objWriteIndex2(lnForCounter).WriteLine _
                        strReadIndex2(lnForCounter2)
                Next

            End If
            GoTo NextRead
        End If
    End If
```

```
Next

WriteStillUndetermined:
'Writes the currently stored four lines into the corresponding residual .fasq files
For lnForCounter2 = 1 To 4 Step 1
   objWriteStillUndeterminedRead1.WriteLine strReadRead1(lnForCounter2)
   objWriteStillUndeterminedIndex1.WriteLine strReadIndex1(lnForCounter2)
   objWriteStillUndeterminedIndex2.WriteLine strReadIndex2(lnForCounter2)
Next

GoTo NextRead

MsgBox "All done."

End Sub
```

Data conversion

In the following, the two most critical VBA scripts were displayed as described in Chapter 3.1.8. These scripts handled the conversion of the input demultiplexed FASTQ data into an analyzable format saved as TXT files. The second script handled the conversion of sequences into numbers (Table 4) and the first script every other task involved.

```
Option Explicit
Option Base 1

Sub DataConversion()

Dim byLanes As Byte: byLanes = 1      'INPUT: Specify number of lanes.
'For each sample, one lane after the other has to be streamed without clearing the array in
between.

Dim wsMain As Worksheet: Set wsMain = ThisWorkbook.Worksheets("Command Center")

Dim lnArray() As Long
ReDim lnArray(16384, 40000)

'The columns in lnArray are ment to contain the following data:
'1. column = Duplicate Identifications
'2. to 45. column = Unique Identifications per frameshift
'Column 46 to x = space for the adaptorNs to be listed (as long)

Dim lnForCounter        As Long
Dim lnForCounter2       As Long
Dim byCommandRow        As Byte
Dim strName             As String
Dim strRead             As String
Dim byLane              As Byte
Dim strCheckSeq         As String   ': strCheckSeq = "AGATCGGAAGAG"
Dim lnFrameshift        As Long
Dim lnFrameReference    As Long     'Expected position of constant sequence
Dim lnStartOfROI        As Long     'Start of ROI as listed in worksheet
Dim lnLenAdaptorN       As Long     'Length of AdaptorN as listed in worksheet
Dim strROI              As String
Dim strAdaptorN         As String
Dim lnROI               As Long
Dim lnAdaptorN          As Long
Dim strPath             As String
strPath = ThisWorkbook.Path & "\Demultiplexed fastq data\"
```

```
Dim objFSO                  As Scripting.FileSystemObject
Set objFSO = New Scripting.FileSystemObject

Dim objFSCounts(44)         As Object
Dim objFSReads(20)          As Object   'Reads from FrameShift +1 to +20
Dim objDuplicateCounts      As Object

Dim objReadRead1            As Object
Dim objMetrics As Object
Dim lnMetrics(7) As Long
'lnMetrics(1) = Reads with matching indices:
'lnMetrics(2) = Prüfsequenz not found:
'lnMetrics(3) = Uncertain bases in ROI:
'lnMetrics(4) = Uncertain bases in AdaptorN:
'lnMetrics(5) = Total duplicate identifications:
'lnMetrics(6) = Total unique identifications:

Dim lnLastRow As Long: lnLastRow = wsMain.Range("C1048576").End(xlUp).Row

wsMain.Range("C" & lnLastRow + 4).Value = "Analysis started at:"
wsMain.Range("D" & lnLastRow + 4).Value = Now



For byCommandRow = 3 To lnLastRow Step 1

   If wsMain.Range("B" & byCommandRow).Value = "Active" Then

      'Making everything ready for the next sample.
      wsMain.Range("B" & byCommandRow).Value = "Processing"
      strName = wsMain.Range("C" & byCommandRow).Value
      lnStartOfROI = wsMain.Range("Q" & byCommandRow).Value
      lnFrameReference = wsMain.Range("T" & byCommandRow).Value
      lnLenAdaptorN = wsMain.Range("U" & byCommandRow).Value
      lnMetrics(1) = 1
      strCheckSeq = wsMain.Range("S" & byCommandRow).Value

      If Not objFSO.FolderExists(strPath & strName & "\Analysis\") Then
         MkDir (strPath & strName & "\Analysis\")
      End If
      If Not objFSO.FolderExists(strPath & strName & "\Positive Frameshift Reads\") Then
         MkDir (strPath & strName & "\Positive Frameshift Reads\")
      End If

      Set objDuplicateCounts = objFSO.OpenTextFile(strPath & strName & _
          "\Analysis\" & strName & "_DuplicateCounts.txt", ForAppending, True)
      For lnForCounter = -23 To 20 Step 1
```

```
        Set objFSCounts(lnForCounter + 24) = objFSO.OpenTextFile(strPath & _
            strName & "\Analysis\" & strName & "_FS" & lnForCounter & _
            "_Counts.txt", ForAppending, True)
    Next
    For lnForCounter = 1 To 20 Step 1
        Set objFSReads(lnForCounter) = objFSO.OpenTextFile(strPath & _
            strName & "\Positive Frameshift Reads\" & strName & "_FS" & _
            lnForCounter & "_Reads.txt", ForAppending, True)
    Next

    byLane = 0
NextLane:
    byLane = byLane + 1
    If byLane > byLanes Then GoTo NextSample

    'In a FASTQ file, the second and from then on every forth row
    'contains the actual sequence.

    Set objReadRead1 = objFSO.OpenTextFile(strPath & strName & "\" & _
        strName & "_L00" & byLane & "_R1.fastq", ForReading)
    If objReadRead1.AtEndOfStream Then GoTo NextLane
    objReadRead1.SkipLine
    If objReadRead1.AtEndOfStream Then GoTo NextLane
    strRead = objReadRead1.ReadLine

    Do

        'This block determines the frameshift of each individual read
        lnFrameshift = lnStr(1, strRead, strCheckSeq, vbTextCompare)
        If lnFrameshift = 0 Then    'Here: "= 0" means "not found"; NOT "no frameshift"!
            lnMetrics(2) = lnMetrics(2) + 1
            GoTo NextRead
        End If
        lnFrameshift = lnFrameshift - lnFrameReference    'Calculates the actual frameshift
        If lnFrameshift < -23 Then GoTo NextRead
        If lnFrameshift > 20 Then GoTo NextRead

        'This block determines the ROI and if it contains "N"
        strROI = Mid(strRead, lnStartOfROI + lnFrameshift, 7)
        If lnStr(1, strROI, "N", vbTextCompare) <> 0 Then
            lnMetrics(3) = lnMetrics(3) + 1
            GoTo NextRead
        End If

        'This block determines the sequence of adaptorN and if it contains "N"
        strAdaptorN = Mid(strRead, 1, lnLenAdaptorN)
        If lnStr(1, strAdaptorN, "N", vbTextCompare) <> 0 Then
```

```
        lnMetrics(4) = lnMetrics(4) + 1
        GoTo NextRead
    End If

    'At this point, the sequences for ROI and AdaptorN were isolated and the
    'frameshift was calculated.
    'Next, these data have to be written into the main Array.

    lnROI = fnSeqToRevComLong(strROI) + 1
    If lnROI = 1000000000 Then
        GoTo NextRead
        lnMetrics(7) = lnMetrics(7) + 1
    End If
    lnAdaptorN = fnSeqToRevComLong(strAdaptorN) + 1
    '+1 is added to prevent confusion between empty spaces and AAAAAAA = 0
    If lnAdaptorN = 1000000000 Then
        GoTo NextRead
        lnMetrics(7) = lnMetrics(7) + 1
    End If

    lnForCounter = 46
    Do While lnArray(lnROI, lnForCounter) <> 0
        If lnArray(lnROI, lnForCounter) = lnAdaptorN Then
            lnArray(lnROI, 1) = lnArray(lnROI, 1) + 1
            lnMetrics(5) = lnMetrics(5) + 1
            GoTo NextRead
        End If
        lnForCounter = lnForCounter + 1
    Loop
    lnArray(lnROI, lnForCounter) = lnAdaptorN
    lnArray(lnROI, lnFrameshift + 25) = lnArray(lnROI, lnFrameshift + 25) + 1
    lnMetrics(6) = lnMetrics(6) + 1
    If lnFrameshift > 0 Then objFSReads(lnFrameshift).WriteLine strRead

NextRead:

    'In a FASTQ file, the second and from then on every forth row contains the
    'actual sequences.
    If objReadRead1.AtEndOfStream Then GoTo NextLane
    objReadRead1.SkipLine
    If objReadRead1.AtEndOfStream Then GoTo NextLane
    objReadRead1.SkipLine
    If objReadRead1.AtEndOfStream Then GoTo NextLane
    objReadRead1.SkipLine
    If objReadRead1.AtEndOfStream Then GoTo NextLane
    strRead = objReadRead1.ReadLine
    lnMetrics(1) = lnMetrics(1) + 1
```

```
    Loop

NextSample:

    'Writes all relevant data from the main array into the previously initialized TXT files.
    For lnForCounter = 1 To 16384 Step 1
       objDuplicateCounts.WriteLine lnArray(lnForCounter, 1)
    Next
    For lnForCounter = 2 To 45 Step 1
       For lnForCounter2 = 1 To 16384 Step 1
          objFSCounts(lnForCounter - 1).WriteLine lnArray(lnForCounter2, lnForCounter)
       Next
    Next

    'The main array is being reset (all values should be "0" now):
    ReDim lnArray(16384, 40000)

    'Finally for this sample, the Metrics.txt file is being created.
    Set objMetrics = objFSO.OpenTextFile(strPath & strName & "\Analysis\" & _
        strName & "_Metrics.txt", ForAppending, True)

    objMetrics.WriteLine "Reads with matching indices:"
    objMetrics.WriteLine lnMetrics(1)
    lnMetrics(1) = 0

    objMetrics.WriteLine "Prüfsequenz not found:"
    objMetrics.WriteLine lnMetrics(2)
    lnMetrics(2) = 0

    objMetrics.WriteLine "Uncertain bases in ROI:"
    objMetrics.WriteLine lnMetrics(3)
    lnMetrics(3) = 0

    objMetrics.WriteLine "Uncertain bases in AdaptorN:"
    objMetrics.WriteLine lnMetrics(4)
    lnMetrics(4) = 0

    objMetrics.WriteLine "Total duplicate identifications:"
    objMetrics.WriteLine lnMetrics(5)
    lnMetrics(5) = 0

    objMetrics.WriteLine "Total unique identifications:"
    objMetrics.WriteLine lnMetrics(6)
    lnMetrics(6) = 0

    objMetrics.WriteLine "Errors occured in fnSeqToRevComLong:"
    objMetrics.WriteLine lnMetrics(7)
```

```
        lnMetrics(7) = 0

        wsMain.Range("B" & byCommandRow).Value = "Done"

    End If  'wsMain.Range("B" & byCommandRow).Value <> "Active"

Next    'Next sample as in Worksheet until all active samples were analyzed.

wsMain.Range("C" & lnLastRow + 5).Value = "Analysis finished at:"
wsMain.Range("D" & lnLastRow + 5).Value = Now

Application.Calculation = xlCalculationAutomatic
Application.ScreenUpdating = True

ThisWorkbook.Save

MsgBox "All done."

End Sub
```

```
Function fnSeqToRevComLong(strInput As String) As Long
'This function converts any ACGT-input into a number
'corresponding to its reverse complement.
'E.g. TTTTTTT -> AAAAAAA -> 0000000 = 0
'E.g. GTTTTTT -> AAAAAAC -> 0000001 = 1
'This can both be used for AdaptorN management and
'determining the line in which to write a specific ROI.

Dim lnFor As Long

fnSeqToRevComLong = 0
For lnFor = Len(strInput) To 1 Step -1
   Select Case Mid(strInput, lnFor, 1)
      Case "A"
         '=> "U"
         fnSeqToRevComLong = fnSeqToRevComLong + 3 * (4 ^ (lnFor - 1))
      Case "C"
         '=> "G"
         fnSeqToRevComLong = fnSeqToRevComLong + 2 * (4 ^ (lnFor - 1))
      Case "G"
         '=> "C"
         fnSeqToRevComLong = fnSeqToRevComLong + 1 * (4 ^ (lnFor - 1))
      Case "T"
         '=> "A"
         'fnSeqToRevComLong = fnSeqToRevComLong + 0 * (4 ^ (lnFor - 1))
      Case Else
         'Most likely if it contains "N" but there should be a filter beforehand.
         fnSeqToRevComLong = 999999999
         '9 times "9"; "0" cannot be used as it belongs to AAAAAAA.
         Exit Function
   End Select
Next

End Function
```

5.2.2 DynaFit scripts

Within DynaFit scripts, semicolons ";" indicate comments and question marks "?" indicate, that a parameter was to be determined alongside an initially estimated value. The lack of a question mark set a parameter as constant.

Adenylate hydrolysis

The following shows the DynaFit script that was used to model the hydrolysis rate of adenylates according to the data obtained by ^{31}P-NMR.

```
[task]
   data = progress
   task = fit

[mechanism]
   AA-Ade --> AA + Ade : k

[constants]
   k = 0.50 ? ; 0.5 h-1

[concentrations]
   AA-Ade = [starting concentration of adenylate according to NMR] ?

[responses]
   AA-Ade = 1

[data]
   delay 0
   file [path to input data file].txt

[output]
   directory [path to store output data]

[end]
```

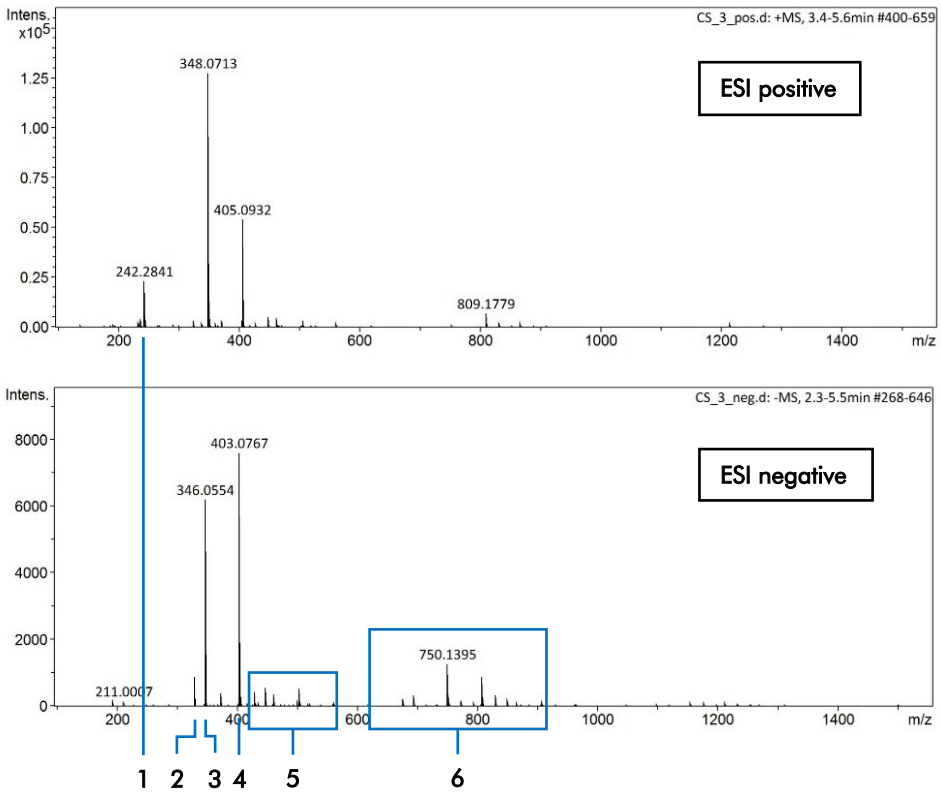
Formation and hydrolysis of aminoacylated RNA

The following shows the DynaFit script that was used to model the formation and hydrolysis of aminoacylated RNA. As input, it required the initial concentration of adenylate which in every case was obtained by a ^{31}P-NMR measurement, k_1 as provided by the previous script, as well as the data obtained by densitometry.

```
[task]
   data  = progress
   task  = fit

[mechanism]
   AA-Ade --> AA + Ade  : k1
   AA-Ade + RNA --> AA-RNA + Ade : k2
   AA-RNA --> AA + RNA : k3

[constants]
   k1  =  [provided by the previous script] ; h-1
   k2  =  0.00001 ? ; μM-1 h-1
   k3  =  0.02 ? ; h-1

[concentrations]
   AA-Ade = [starting concentration of adenylate according to NMR] ; υM
   RNA = 2.61 ; υM

[responses]
   AA-RNA = 1

[data]
   delay 0
   file [path to input data file].txt

[output]
   directory [path to store output data]

[end]
```

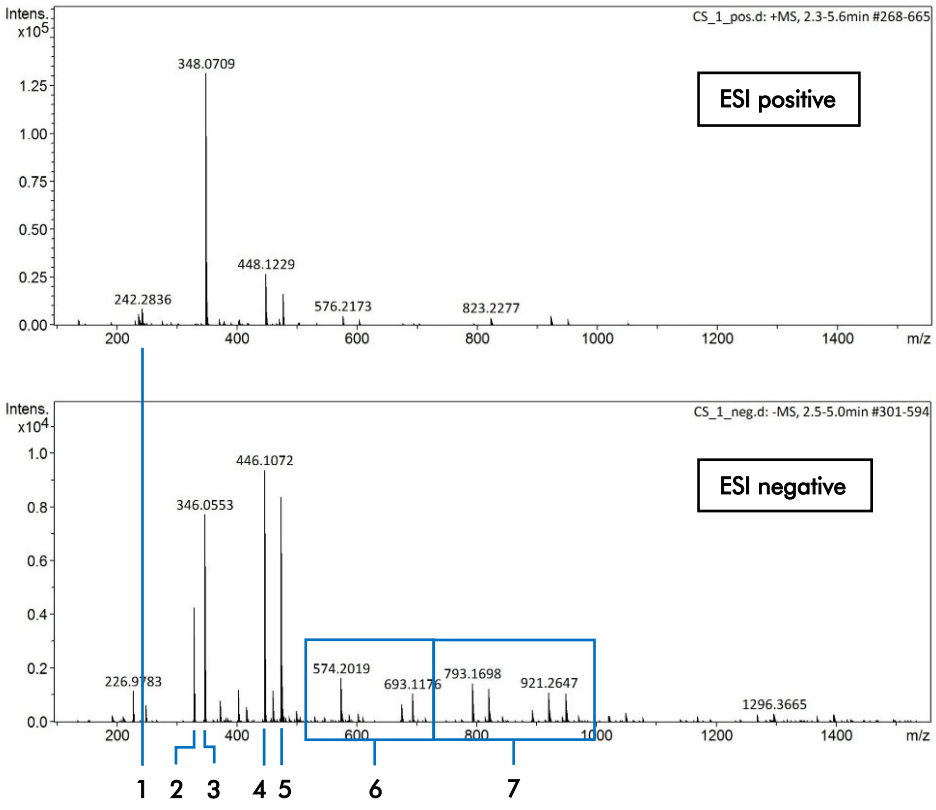
6 Appendix

6.1 Supplementary information



#	Compound	m/z calculated		m/z found	
		$[M+H]^+$	$[M-H]^-$	ESI pos.	ESI neg.
1	Tetrabutylammonium	242.47 [M]		242.284	-
2	cAMP	330.213	328.199	-	328.0
3	AMP	348.227	346.213	348.071	346.055
4	Gly-AMP	405.279	403.265	405.093	403.077
5	Isobutylformiate-adducts	+100 of previous		mostly found	
6	Combinations of previous signals i.e. typical artifacts				

Supplementary Figure 1. Mass spectrometric analysis using electrospray ionization (ESI) of the Gly-AMP product.



#	Compound	m/z calculated		m/z found	
		$[M+H]^+$	$[M-H]^-$	ESI pos.	ESI neg.
1	Tetrabutylammonium	242.47 [M]		242.284	-
2	cAMP	330.213	328.199	-	328.0
3	AMP	348.227	346.213	348.071	346.055
4	AMP-isobutylformiate	448.804	446.790	448.123	446.107
5	Lys-AMP	475.395	474.388	476.165	474.150
	Lys-AMP + H^+	477.409	475.395	477.167	475.152
6	Isobutylformiate-adducts	+100 of previous		mostly found	
7	Combinations of previous signals i.e. typical artifacts				

Supplementary Figure 2. Mass spectrometric analysis using electrospray ionization (ESI) of the L-Lys-AMP product. This analysis resulted in additional peaks as lysine contained two amino moieties that could easily be ionized compared to alanine and glycine, both of which contained only one.

6.2 Acknowledgements

First and foremost, I'd like to express my most fulsome gratitude towards my supervisor Prof. Dr. Andres Jäschke. It was a pleasure to turn his vision of this project into impactful results. His management style was most agreeable, and I always enjoyed being part of his research group. I highly valued the degree of freedom that was granted to me.

I also highly appreciate the efforts of Prof. Dr. Dieter Braun to keep the origin of life research cluster running and for maintaining a collaborative and vivid team spirit.

Thanks to Prof. Dr. Kerstin Göpfrich and Prof. Dr. Friedrich Frischknecht for their interest in my work and for acting as my examiners.

This project will continue mostly in the hands of Philip Slawetzki and Paolo Zucchetti and I could not have asked for any better suited successors. I am excited to see how they will continue developing this topic.

The time I spent with all my colleagues during the last years was amazing. I enjoyed both the professional cooperation as well as the great informality. I always felt very welcome or even like being among friends.

It would be hard to imagine how our work would be possible without the heroic support of our secretary Viola Funk as well as the laboratory technicians Heiko Rudy and Tobias Timmermann. Thank you for all your support; you made a lot of thinks a lot more doable!

In order of appearance, I would like to thank all interns that contributed to this project. All of them showed so much enthusiasm and commitment to research in the origin of life. I enjoyed working with every single one; their presence was inspiring and their work fruitful.

Thanks to Philipp Höflich who spent his bachelor thesis with me. His ingenuity contributed to the optimization of the ligation-based library preparation. After he left, the procedure underwent almost no changes.

Suria Itzel Morales Guzmán absolved an extensive internship working with me for 3 month; I did not spent more time with any other intern. She worked on the optimization of the adenylate synthesis including her independent work on the use of Reagent H. She also assisted me tirelessly in the chromatographic purification of adenylates as well as a multitude of aminoacylation experiments.

I am grateful for the flawless cooperation with Lina Bingel during her research internship. She was a great help while we were working on the synthesis of adenylates, optimized the amount of adenylate and periodate for the oxidative fixation and conducted first mass spectrometric analyses.

The outstanding independence of Jennifer Sauerland deeply impressed me. She contributed notably to chromatographic purification, adenylate synthesis and aminoacylation experiments – even in my absence.

It was a joy working with Mara Behnke. Not only due to her remarkable baking skills but also for her invaluable assistance in preparing and multiplexing most of the 80 aminoacylation samples that we submitted to sequencing and that made up almost all sequencing results presented in this thesis.

Lastly, I'd like to thank Niko Jakob for his efforts during the synthesis of the cytidylate, the following aminoacylation, sequencing, and data analysis but especially for baptizing the subgroup the "Origin of Life Department".

6.3 Abbreviations

A	adenine
aa	amino acid
aa-AMP	aminoacyl-5′-adenylate
aaRS	aminoacyl tRNA synthetase
AMP	adenosine monophosphate
APB	3-(Acrylamido)phenylboronic acid
Boc	N-tert-Butoxycarbonyl (protecting group)
bp	base pairs (length of double stranded DNA and RNA)
C	cytosine
CMP	cytidine monophosphate
DNA	deoxyribonucleic acid
G	guanine
ImpA	adenosine 5′-phosphorimidazolide
IVT	in-vitro transcription
MS	mass spectrometry
mRNA	messenger RNA
NMR	nuclear magnetic resonance
nt	nucleotides (length of single stranded DNA and RNA)
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
rApp-DNA	5' adenylated DNA
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT	reverse transcription
RT-PCR	reverse transcription polymerase chain reaction
T	thymine
TBA	tetra-n-butylammonium hydroxide
TBA-AMP	tetra-n-butylammonium salt of adenosine monophosphate
TBA-CMP	tetra-n-butylammonium salt of cytidine monophosphate
TFA	trifluoroacetic acid
tRNA	transfer RNA
U	uridine

Amino acid	three letter code	one letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophane	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
Selenocysteine	Sec	U
Pyrrolysine	Pyl	O

Table 10. Full names of all 22 proteinogenic amino acids alongside their common three and one letter abbreviations.

6.4 References

1 Bang, M. L. *et al.* The complete gene sequence of titin, expression of an unusual ≈700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ Res* **89**, 1065-1072 (2001).

2 Gubala, V., Betancourt, J. E. & Rivera, J. M. Expanding the Hoogsteen edge of 2'-deoxyguanosine:: Consequences for G-quadruplex formation. *Org Lett* **6**, 4735-4738 (2004).

3 Kapp, L. D. & Lorsch, J. R. The molecular mechanics of eukaryotic translation. *Annu Rev Biochem* **73**, 657-704 (2004).

4 Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics* **6** (2015).

5 Sharp, S. J., Schaack, J., Cooley, L., Burke, D. J. & Soil, D. Structure and transcription of eukaryotic tRNA gene. *Critical Reviews in Biochemistry* **19**, 107-144 (1985).

6 Cusack, S. Aminoacyl-tRNA synthetases. *Curr Opin Struc Biol* **7**, 881-889 (1997).

7 Ibba, M. & Söll, D. Aminoacyl-tRNA synthesis. *Annu Rev Biochem* **69**, 617-650 (2000).

8 Gerber, A. P. & Keller, W. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science* **286**, 1146-1149 (1999).

9 Lorenz, C., Lünse, C. E. & Mörl, M. tRNA Modifications: Impact on Structure and Thermal Adaptation. *Biomolecules* **7** (2017).

10 Andersen, G. R., Nissen, P. & Nyborg, J. Elongation factors in protein biosynthesis. *Trends Biochem Sci* **28**, 434-441 (2003).

11 Korostelev, A. A. Structural aspects of translation termination on the ribosome. *Rna* **17**, 1409-1421 (2011).

12 Sonenberg, N. & Dever, T. E. Eukaryotic translation initiation factors and regulators. *Curr Opin Struc Biol* **13**, 56-63 (2003).

13 Bashan, A. *et al.* Structural basis of the ribosomal machinery for peptide bond formation, translocation, and nascent chain progression. *Molecular Cell* **11**, 91-102 (2003).

14 Koonin, E. V. & Novozhilov, A. S. Origin and Evolution of the Genetic Code: The Universal Enigma. *Iubmb Life* **61**, 99-111 (2009).

15 Sengupta, S. & Higgs, P. G. Pathways of Genetic Code Evolution in Ancient and Modern Organisms. *J Mol Evol* **80**, 229-243 (2015).

16 De Duve, C. The second genetic code. *Nature* **333**, 117-118 (1988).

17 Moras, D. Structural and functional relationships between aminoacyl-tRNA synthetases. *Trends Biochem Sci* **17**, 159-164 (1992).

18 Trifonov, E. N. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139-151 (2000).
19 Men, A. E., Wilson, P., Siemering, K. & Forrest, S. Sanger DNA sequencing. *Next Generation Genome Sequencing: Towards Personalized Medicine*, 1-11 (2008).
20 Kircher, M., Heyn, P. & Kelso, J. Addressing challenges in the production and analysis of illumina sequencing data. *Bmc Genomics* **12** (2011).
21 Metzker, M. L. APPLICATIONS OF NEXT-GENERATION SEQUENCING Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**, 31-46 (2010).
22 Atsumi, T., Mccarter, L. & Imae, Y. Polar and Lateral Flagellar Motors of Marine Vibrio Are Driven by Different Ion-Motive Forces. *Nature* **355**, 182-184 (1992).
23 Okuno, D., Iino, R. & Noji, H. Rotation and structure of F0F1-ATP synthase. *J Biochem* **149**, 655-664 (2011).
24 Macklem, P. T. & Seely, A. Towards a Definition of Life. *Perspectives in Biology and Medicine* **53**, 330-340 (2010).
25 Leggett, M. J., McDonnell, G., Denyer, S. P., Setlow, P. & Maillard, J. Y. Bacterial spore structures and their protective role in biocide resistance. *J Appl Microbiol* **113**, 485-498 (2012).
26 Harrison, S. C. Principles of virus structure. *Fields virology* **1**, 53-85 (2001).
27 Ryu, W.-S. Virus life cycle. *Molecular virology of human pathogenic viruses*, 31 (2016).
28 Tan, I. S. & Ramamurthi, K. S. Spore formation in. *Env Microbiol Rep* **6**, 212-225 (2014).
29 Saccheri, I. & Hanski, I. Natural selection and population dynamics. *Trends Ecol Evol* **21**, 341-347 (2006).
30 Benner, S. A. Defining life. *Astrobiology* **10**, 1021-1030 (2010).
31 Ruiz-Mirazo, K., Briones, C. & de la Escosura, A. Prebiotic Systems Chemistry: New Perspectives for the Origins of Life. *Chemical Reviews* **114**, 285-366 (2014).
32 Bouvier, A. & Wadhwa, M. The age of the Solar System redefined by the oldest Pb-Pb age of a meteoritic inclusion. *Nat Geosci* **3**, 637-641 (2010).
33 Pearce, B. K., Tupper, A. S., Pudritz, R. E. & Higgs, P. G. Constraining the time interval for the origin of life on Earth. *Astrobiology* **18**, 343-364 (2018).
34 Yin, Q. Z. *et al.* A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* **418**, 949-952 (2002).

35 Boehnke, P. & Harrison, T. M. Illusory late heavy bombardments. *Proceedings of the National Academy of Sciences* **113**, 10802-10806 (2016).

36 Cohen, B. A., Swindle, T. D. & Kring, D. A. Support for the lunar cataclysm hypothesis from lunar meteorite impact melt ages. *Science* **290**, 1754-1756 (2000).

37 Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature* **435**, 466-469 (2005).

38 Ryder, G., Koeberl, C. & Mojzsis, S. J. Heavy bombardment of the Earth at~ 3.85 Ga: The search for petrographic and geochemical evidence. *Origin of the Earth and Moon* **475** (2000).

39 Ohtomo, Y., Kakegawa, T., Ishida, A., Nagase, T. & Rosing, M. T. Evidence for biogenic graphite in early Archaean Isua metasedimentary rocks. *Nat Geosci* **7**, 25-28 (2014).

40 Rosing, M. T. 13C-depleted carbon microparticles in> 3700-Ma sea-floor sedimentary rocks from West Greenland. *Science* **283**, 674-676 (1999).

41 Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J. & Brasier, M. D. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat Geosci* **4**, 698-702 (2011).

42 Robertson, M. P. & Joyce, G. F. The origins of the RNA world. *Cold Spring Harbor perspectives in biology* **4**, a003608 (2012).

43 Weiss, M. C. *et al.* The physiology and habitat of the last universal common ancestor. *Nat Microbiol* **1**, 16116 (2016).

44 Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124-2128 (1999).

45 Penny, D. & Poole, A. The nature of the last universal common ancestor. *Curr Opin Genet Dev* **9**, 672-677 (1999).

46 Nirenberg, M. W. *et al.* in *Cold Spring Harbor Symposia on Quantitative Biology.* 549-557 (Cold Spring Harbor Laboratory Press).

47 *Nobel Prize in Physiology or Medicine for the interpretation of the genetic code and its function in protein synthesis*, <https://www.nobelprize.org/prizes/medicine/1968/summary/> (1968).

48 Koonin, E. V. & Novozhilov, A. S. Origin and Evolution of the Universal Genetic Code. *Annual Review of Genetics, Vol 51* **51**, 45-62 (2017).

49 Bock, A., Forchhammer, K., Heider, J. & Baron, C. Selenoprotein Synthesis - an Expansion of the Genetic-Code. *Trends Biochem Sci* **16**, 463-467 (1991).

50 Namy, O. *et al.* Adding pyrrolysine to the genetic code. *Febs Lett* **581**, 5282-5288 (2007).
51 Gilis, D., Massar, S., Cerf, N. J. & Rooman, M. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol* **2** (2001).
52 King, J. L. & Jukes, T. H. Non-Darwinian Evolution. *Science* **164**, 788-+ (1969).
53 Di Giulio, M. The origin of the genetic code: theories and their relationships, a review. *Biosystems* **80**, 175-184 (2005).
54 Hasegawa, M. & Miyata, T. On the antisymmetry of the amino acid code table. *Origins of life* **10**, 265-270 (1980).
55 Rumer, I. B. Codon systematization in the genetic code. *Doklady Akademii Nauk SSSR* **167**, 1393-1394 (1966).
56 Volkensh.Mv & Rumer, Y. B. Systematics of Codons. *Biophys-Ussr* **12**, 6-& (1967).
57 Wetzel, R. Evolution of the Aminoacyl-Transfer-Rna Synthetases and the Origin of the Genetic-Code. *J Mol Evol* **40**, 545-550 (1995).
58 Crick, F. H. The origin of the genetic code. *Journal of molecular biology* **38**, 367-379 (1968).
59 Chechetkin, V. R. Block structure and stability of the genetic code. *Journal of Theoretical Biology* **222**, 177-188 (2003).
60 Freeland, S. J., Wu, T. & Keulmann, N. The case for an error minimizing standard genetic code. *Origins Life Evol B* **33**, 457-477 (2003).
61 Haig, D. & Hurst, L. D. A Quantitative Measure of Error Minimization in the Genetic-Code. *J Mol Evol* **33**, 412-417 (1991).
62 Trifonov, E. N. The Triplet Code From First Principles. *Journal of Biomolecular Structure and Dynamics* **22**, 1-11 (2004).
63 Sonneborn, T. M. in *Evolving genes and proteins* 377-397 (Elsevier, 1965).
64 Woese, C. Models for the evolution of codon assignments. *Journal of Molecular Biology* **43**, 235-240 (1969).
65 Woese, C. R. On Evolution of Genetic Code. *P Natl Acad Sci USA* **54**, 1546-+ (1965).
66 Hopfield, J. J. Origin of Genetic Code - Testable Hypothesis Based on Transfer-Rna Structure, Sequence, and Kinetic Proofreading. *P Natl Acad Sci USA* **75**, 4334-4338 (1978).
67 Ishida, T. & Sueoka, N. Rearrangement of Secondary Structure of Tryptophan Srna in Escherichia Coli. *P Natl Acad Sci USA* **58**, 1080-& (1967).

68 Kim, S. *et al.* The general structure of transfer RNA molecules. *Proceedings of the National Academy of Sciences* **71**, 4970-4974 (1974).

69 Lindahl, T., Adams, A. & Fresco, J. R. Renaturation of Transfer Ribonucleic Acids through Site Binding of Magnesium. *P Natl Acad Sci USA* **55**, 941-+ (1966).

70 Richards, E. G., Simpkins, H., Geroch, M. E. & Lecanidou, R. Optical Properties and Base Pairing of Escherichia-Coli 5s Rna. *Biopolymers* **11**, 1031-+ (1972).

71 Robertus, J. *et al.* Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* **250**, 546-551 (1974).

72 Schwarz, U., Menzel, H. M. & Gassen, H. G. Codon-dependent rearrangement of the three-dimensional structure of phenylalanine tRNA, exposing the T-ϕ-CG sequence for binding to the 50S ribosomal subunit. *Biochemistry-Us* **15**, 2484-2490 (1976).

73 Yoshida, M., Kaziro, Y. & Ukita, T. The modification of nucleosides and nucleotides: X. Evidence for the important role of inosine residue in codon recognition of yeast alanine tRNA. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis* **166**, 646-655 (1968).

74 Ellington, A. D. & Szostak, J. W. In vitro Selection of Rna Molecules That Bind Specific Ligands. *Nature* **346**, 818-822 (1990).

75 Joyce, G. F. Amplification, Mutation and Selection of Catalytic Rna. *Gene* **82**, 83-87 (1989).

76 Klug, S. J. & Famulok, M. All You Wanted to Know About Selex. *Mol Biol Rep* **20**, 97-107 (1994).

77 Tuerk, C. & Gold, L. Systematic Evolution of Ligands by Exponential Enrichment - Rna Ligands to Bacteriophage-T4 DNA-Polymerase. *Science* **249**, 505-510 (1990).

78 Illangasekare, M., Sanchez, G., Nickles, T. & Yarus, M. Aminoacyl-Rna Synthesis Catalyzed by an Rna. *Science* **267**, 643-647 (1995).

79 Sklarz, B. Organic Chemistry of Periodates. *Quarterly Reviews* **21**, 3-& (1967).

80 Steinschneider, A. & Fraenkel.H. Studies of Nucleotide Sequences in Tobacco Mosaic Virus Ribonucleic Acid .3. Periodate Oxidation and Semicarbazone Formation. *Biochemistry-Us* **5**, 2729-+ (1966).

81 Illangasekare, M. & Yarus, M. Specific, rapid synthesis of Phe-RNA by RNA. *P Natl Acad Sci USA* **96**, 5470-5475 (1999).

82 Chumachenko, N. V., Novikov, Y. & Yarus, M. Rapid and Simple Ribozymic Aminoacylation Using Three Conserved Nucleotides. *J Am Chem Soc* **131**, 5257-5263 (2009).

83 Turk, R. M., Chumachenko, N. V. & Yarus, M. Multiple translational products from a five-nucleotide ribozyme. *P Natl Acad Sci USA* **107**, 4585-4589 (2010).
84 Turk, R. M., Illangasekare, M. & Yarus, M. Catalyzed and Spontaneous Reactions on Ribozyme Ribose. *J Am Chem Soc* **133**, 6044-6050 (2011).
85 Illangasekare, M. & Yarus, M. Small aminoacyl transfer centers at GU within a larger RNA. *Rna Biol* **9**, 59-66 (2012).
86 Danger, G., Boiteau, L., Cottet, H. & Pascal, R. The peptide formation mediated by cyanate revisited. N-carboxyanhydrides as accessible intermediates in the decomposition of N-carbamoylamino acids. *J Am Chem Soc* **128**, 7412-7413 (2006).
87 Danger, G. *et al.* 5(4H)-Oxazolones as Intermediates in the Carbodiimide- and Cyanamide-Promoted Peptide Activations in Aqueous Solution. *Angew Chem Int Edit* **52**, 611-614 (2013).
88 Danger, G., Plasson, R. & Pascal, R. Pathways for the formation and evolution of peptides in prebiotic environments. *Chemical Society Reviews* **41**, 5416-5429 (2012).
89 Leman, L., Orgel, L. & Ghadiri, M. R. Carbonyl sulfide-mediated prebiotic formation of peptides. *Science* **306**, 283-286 (2004).
90 Pressman, A. D. *et al.* Mapping a Systematic Ribozyme Fitness Landscape Reveals a Frustrated Evolutionary Network for Self-Aminoacylating RNA. *J Am Chem Soc* **141**, 6213-6223 (2019).
91 Berg, P. The chemical synthesis of amino acyl adenylates. *Journal of Biological Chemistry* **233**, 608-611 (1958).
92 Armstrong, D. W., Seguin, R., Saburi, M. & Fendler, J. H. Synthesis of Amino Acyl Adenylates Using the Tert-Butoxycarbonyl Protecting Group. *J Mol Evol* **13**, 103-113 (1979).
93 Lewinsohn, R., Paechtho.M & Katchalsky, A. Polycondensation of Amino Acid Phosphoanhydrides .3. Polycondensation of Alanyl Adenylate. *Biochim Biophys Acta* **140**, 24-+ (1967).
94 Igloi, G. L. & Kossel, H. Affinity Electrophoresis for Monitoring Terminal Phosphorylation and the Presence of Queuosine in Rna - Application of Polyacrylamide Containing a Covalently Bound Boronic Acid. *Nucleic Acids Res* **13**, 6881-6898 (1985).
95 Evans, M. E., Clark, W. C., Zheng, G. Q. & Pan, T. Determination of tRNA aminoacylation levels by high-throughput sequencing. *Nucleic Acids Res* **45** (2017).
96 Huang, H. & Rabenstein, D. L. A cleavage cocktail for methionine-containing peptides. *Journal of Peptide Research* **53**, 548-553 (1999).

97 Sauerland, J. *Optimization of Amino Acid Activation and Aminoacylation to Elucidate the Origin of the Genetic Code* (Heidelberg University, 2022).
98 Guzmán, S. I. M. *Characterization of a novel screening method for self-aminoacylating tRNA precursors* (Heidelberg University, 2022).
99 Dawson, R. & Elliott, W. Buffers and physiological media. *Data for biochemical research. Oxford University Press, Oxford*, 200-205 (1959).
100 Bowman, J. C., Lenz, T. K., Hud, N. V. & Williams, L. D. Cations in charge: magnesium ions in RNA folding and catalysis. *Curr Opin Struc Biol* **22**, 262-272 (2012).
101 Bingel, L. *Reproducibility and Optimization of novel RNA library preparation procedure* (Heidelberg University, 2022).
102 Clamp, J. & Hough, L. The periodate oxidation of amino acids with reference to studies on glycoproteins. *Biochemical Journal* **94**, 17 (1965).
103 Bertoldo, M., Zampano, G., Suffner, L., Liberati, E. & Ciardelli, F. Oxidation of glycogen "molecular nanoparticles" by periodate. *Polymer Chemistry* **4**, 653-661 (2013).
104 Abdelakher, M. & Smith, F. Oxidation of Glycogen with Periodic Acid. *J Am Chem Soc* **81**, 1718-1721 (1959).
105 Tayri-Wilk, T. *et al.* Mass spectrometry reveals the chemistry of formaldehyde cross-linking in structured proteins. *Nat Commun* **11** (2020).
106 Thavarajah, R., Mudimbaimannar, V. K., Elizabeth, J., Rao, U. K. & Ranganathan, K. Chemical and physical basics of routine formaldehyde fixation. *Journal of oral and maxillofacial pathology* **16**, 400-405 (2012).
107 Hentzen, D., Mandel, P. & Garel, J.-P. Relation between aminoacyl-tRNA stability and the fixed amino acid. *Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis* **281**, 228-232 (1972).
108 Schenborn, E. T. & Mierendorf, R. C. A Novel Transcription Property of Sp6 and T7 Rna-Polymerases - Dependence on Template Structure. *Nucleic Acids Res* **13**, 6223-6236 (1985).
109 Milligan, J. F., Groebe, D. R., Witherell, G. W. & Uhlenbeck, O. C. Oligoribonucleotide Synthesis Using T7 Rna-Polymerase and Synthetic DNA Templates. *Nucleic Acids Res* **15**, 8783-8798 (1987).
110 FerreDAmare, A. R. & Doudna, J. A. Use of cis- and trans-ribozymes to remove 5' and 3' heterogeneities from milligrams of in vitro transcribed RNA. *Nucleic Acids Res* **24**, 977-978 (1996).
111 Olsen, D. B., Carroll, S. S., Culberson, J. C., Shafer, J. A. & Kuo, L. C. Effect of Template Secondary Structure on the Inhibition of Hiv-1 Reverse-Transcriptase by a Pyridinone Nonnucleoside Inhibitor. *Nucleic Acids Res* **22**, 1437-1443 (1994).

112 Suo, Z. C. & Johnson, K. A. Effect of RNA secondary structure on the kinetics of DNA synthesis catalyzed by HIV-1 reverse transcriptase. *Biochemistry-Us* **36**, 12459-12467 (1997).

113 Cheong, C. J., Varani, G. & Tinoco, I. Solution Structure of an Unusually Stable Rna Hairpin, 5'ggac(Uucg)Gucc. *Nature* **346**, 680-682 (1990).

114 <By Yikrazuul - own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=10376954> (2024).

115 *NEBNext® Multiplex Oligos for Illumina®*, <https://www.neb.com/en/-/media/nebus/files/manuals/manuale7335_e7500_-e7710_e7730.pdf?rev=2e735fd18b544d46b36ee0e88353ef5c&sc_lang=en&hash=DF0E393E7F8745A46F771D9944A44FD8> (2025).

116 Sambrook, J. & Russell, D. W. Isolation of DNA fragments from polyacrylamide gels by the crush and soak method. *CSH Protoc* **2006** (2006).

117 Thuring, R., Sanders, J. & Borst, P. A freeze-squeeze method for recovering long DNA from agarose gels. *Anal Biochem* **66**, 213-220 (1975).

118 Tautz, D. & Renz, M. An Optimized Freeze-Squeeze Method for the Recovery of DNA Fragments from Agarose Gels. *Anal Biochem* **132**, 14-19 (1983).

119 *G-CAPSULE™*, <https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/product/documents/400/385/z742614pis.pdf?srsltid=AfmBOopcbdQEMeP82A4l-D2JwoHR9D3m3q-HpngtsUXzNHiQYJC_zN0F> (2025).

120 Fadouloglou, V. E. Electroelution of nucleic acids from polyacrylamide gels: A custom-made, agarose-based electroeluter. *Anal Biochem* **437**, 49-51 (2013).

121 Spiess, A. N. & Ivell, R. A highly efficient method for long-chain cDNA synthesis using trehalose and betaine. *Anal Biochem* **301**, 168-174 (2002).

122 Kurschat, W. C., Müller, J., Wombacher, R. & Helm, M. Optimizing splinted ligation of highly structured small RNAs. *Rna* **11**, 1909-1914 (2005).

123 Biolabs, N. E. *Substrate-based Ligase Selection Chart*, <https://www.neb.com/en/tools-and-resources/selection-charts/substrate-based-ligase-selection-chart> (2024).

124 Winz, M. L. *et al.* Capture and sequencing of NAD-capped RNA sequences with NAD captureSeq. *Nature Protocols* **12** (2017).

125 Hafner, M. *et al.* Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**, 3-12 (2008).
126 McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112-129 (2017).
127 Höflich, P. *Optimization of a Novel Library Preparation Procedure for Illumina Sequencing* (Heidelberg University, 2021).
128 Behnke, M. *Examination of the kinetics of aminoacylation through DNA library synthesis enabling the elucidation of the origin of the genetic code* (Heidelberg University, 2023).
129 Jakob, N. *Internship Lab Report* (Heidelberg University, 2024).
130 Wheeler, T. J., Clements, J. & Finn, R. D. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *Bmc Bioinformatics* **15** (2014).
131 Shen, Y. N., Pressman, A., Janzen, E. & Chen, I. A. Kinetic sequencing (k-Seq) as a massively parallel assay for ribozyme kinetics: utility and critical parameters. *Nucleic Acids Res* **49** (2021).
132 Hou, Y. M. CCA Addition to tRNA: Implications for tRNA Quality Control. *Iubmb Life* **62**, 251-260 (2010).
133 Kuzmic, P. Program DYNAFIT for the analysis of enzyme kinetic data: Application to HIV proteinase. *Anal Biochem* **237**, 260-273 (1996).
134 Peacock, J. R. *et al.* Amino acid-dependent stability of the acyl linkage in aminoacyl-tRNA. *Rna* **20**, 758-764 (2014).
135 Su, M. *et al.* Triplet-Encoded Prebiotic RNA Aminoacylation. *J Am Chem Soc* **145**, 15971-15980 (2023).
136 Kluger, R., Li, X. F. & Loo, R. W. 1996 Bader Award Lecture - Aminoacyl ethyl phosphates. Biomimetically activated amino acids. *Canadian Journal of Chemistry* **74**, 2395-2400 (1996).
137 Spencer, P. C. *Synthesis and reactions of protected aminoacyl ethyl phosphates*, (1998).
138 Katritzky, A. R., Wang, M. Y., Yang, H. F., Zhang, S. M. & Akhmedov, N. G. 1-(α -Boc-aminoacyl)benzotriazoles:: Stable chiral α -aminoacylation reagents. *Arkivoc*, 134-142 (2002).
139 Jash, B., Tremmel, P., Jovanovic, D. & Richert, C. Single nucleotide translation without ribosomes. *Nature Chemistry* **13**, 751-+ (2021).
140 Liu, Z. W., Ajram, G., Rossi, J. C. & Pascal, R. The Chemical Likelihood of Ribonucleotide--Amino acid Copolymers as Players for Early Stages of Evolution. *J Mol Evol* **87**, 83-92 (2019).
141 Sharma, G. & First, E. A. Thermodynamic Analysis Reveals a Temperature-dependent Change in the Catalytic Mechanism of Tyrosyl-tRNA Synthetase. *Journal of Biological Chemistry* **284**, 4179-4190 (2009).

142 Murakami, H., Ohta, A., Ashigai, H. & Suga, H. A highly flexible tRNA acylation method for non-natural polypeptide synthesis (vol 3, pg 357, 2006). *Nat Methods* **3**, 655-655 (2006).

143 Ohuchi, M., Murakami, H. & Suga, H. The flexizyme system: a highly flexible tRNA aminoacylation tool for the translation apparatus. *Curr Opin Chem Biol* **11**, 537-542 (2007).

144 Greenwald, J., Kwiatkowski, W. & Riek, R. Peptide Amyloids in the Origin of Life. *Journal of Molecular Biology* **430**, 3735-3750 (2018).