Sequence self-selection by the network dynamics of random ligating oligomer pools

Patrick W. Kudella

Dissertation

München 2021

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Sequence self-selection by the network dynamics of random ligating oligomer pools

Patrick W. Kudella

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

vorgelegt von
Patrick W. Kudella
aus Rastatt

München, den 11.02.2021

Erstgutachter: Prof. Dr. Dieter Braun
Zweitgutachter: Prof. Dr. Ulrich Gerland

Tag der Abgabe: 11.02.2021
Tag der mündlichen Prüfung: 09.04.2021

1 Abstract

1.1 Selbst-Selektion von Sequenzen durch die Netzwerkdynamik von zufällig ligierenden Oligomer-Ensembles

Ein Problem bei der Erforschung des Ursprungs des Lebens ist der Schritt von kurzen Oligomeren mit zufälliger Basenfolge zu längeren aktiven Komplexen wie Ribozymen. Chemie und physikalische Nichtgleichgewichtsysteme haben mögliche Wege für die Synthese von kurzen Strängen durch aktivierte Nukleotide oder sogenannte "activation agents" gefunden. Es wurde auch gezeigt, dass mittellange, autokatalytisch aktive Komplexe Reaktionsnetzwerke bilden können, die für darauffolgende Reaktionen und die Entstehung anspruchsvollerer Systeme notwendig sind. Der Zwischenschritt ist schwieriger: Kurze Stränge mit zufälliger Sequenz müssen länger werden, wobei gleichzeitig eine Verringerung der Sequenzentropie auftreten muss, die die Bildung von doppelsträngigen Komplexen nicht beeinträchtigt. Dieser Vorselektionsschritt ist aufgrund des riesigen Sequenzraums der Oligomerstränge notwendig und muss zusätzlich den physikalischen Rahmenbedingungen einer Umgebung ähnlich der frühen Erde entsprechen. Die ursprünglichen Reaktionen hingen vermutlich lediglich von den inhärenten Parametern der Oligomerstränge wie der Hybridisierung zu Doppelstrangkomplexen und grundlegenden physikalischen Mechanismen, wie etwa Temperaturänderungen ab.

Obwohl kurze Stränge wahrscheinlich durch einen nicht-templierten polymerisations-ähnlichen Verlängerungsmechanismus entstanden sind, wird bei längeren Strängen die Weitergabe von Sequenzinformationen von einem Strang zum anderen als essentiell angesehen. Chemische Ligation ist im Zusammenhang mit dem Ursprung des Lebens wahrscheinlich, aber die Ausbeute und Spezifität ist bekanntermaßen gering. Als Modell für diesen Mechanismus wird in dieser Studie daher eine moderne DNA-Ligase biologischen Ursprungs verwendet, um die Ligationsreaktion des Modell-Oligomers (DNA) ohne einen möglichen Sequenz-Bias zu ermöglichen. Nach 1000 Temperaturzyklen, die für die Dissoziation und Rehybridisierung der verlängerten Doppelstränge notwendig sind, zeigen die resultierenden Reaktionsprodukte mehrere deutliche Selektionsmerkmale. Im Vergleich zu zufälligen Strängen gleicher Länge weisen die entstandenen Reaktionsprodukte eine deutlich reduzierte Sequenzentropie auf. Diese Stränge können in zwei Gruppen kategorisiert werden: A-Typ- und T-Typ-Stränge mit einem Verhältnis von etwa 70:30 % der jeweiligen Base. Da die Stränge nur dann als Templat oder Substrat in der templierten Ligation reagieren können, wenn sie sich in der einzelsträngigen Konformation befinden, hemmt diese Selektion die Bildung von selbstfaltenden Oligomerprodukten. Gleichzeitig wird die Bildung von doppelsträngigen Komplexen nicht reduziert, da es sich bei den A- und T-Typ-Gruppen überwiegend um Ensembles von Komplementsequenzen handelt. Eine Auswahl der häufigsten Teilsequenzen aus jeder Gruppe als neuer Startpool zeigt eine höhere Liagtionsaktivität des Pools im Vergleich zu einem Pool aus zufälligen Strängen. Während die Analyse der Doppelstrang-Typen und der genauen Dynamik im experimentellen System begrenzt ist, können alle Parameter in einer eng verwandten theoretischen Simulation detailliert analysiert werden. Diese Simulation zeigt einen eindeutigen Zusammenhang der wichtigsten Reaktionsraten im System: die Hybridisierungs-On-Rate, die Ligations-Rate und die Hybridisierungs-Off-Rate.

Mit diesen Parametern konnte das Auftreten einer lokalen Minimum-Maximum-Eigenschaft in der Konzentration-über-Stranglänge-Analyse identifiziert und das Auftreten des Merkmals im Experiment vorhergesagt werden. Das Ergebnis längenabhängiger dynamischer Regime führte auch zum Verständnis der längenabhängigen dominanten Wachstumsmodi der Produktstränge. Die Analyse der entstehenden 24mer mit einer einfachen Simulation, die auf diesen Wachstumsmodi basiert, zeigt ein häufiges Sequenzmuster an Ligationsstellen: `ATAT`. Ein kleiner Bias in Richtung des selbstkomplementären `AT`-Sequenzmotivs am 3'-Ende der Stränge im ursprünglichen 12mer-Pool kann als wahrscheinliche Ursache identifiziert werden.

Diese Studie zeigt, wie längere, neu ligierte Produktstränge in einem einfachen Modellsystem eine reduzierte Entropie aufweisen, während die wichtigen Hybridisierungseigenschaften der Oligomerstränge erhalten bleiben. Die Dynamik der Elongation hängt von den mikroskopischen Raten der Komplexbildung, Ligation und Dissoziation ab, während die Frequenz der Temperaturzyklen ein Ratenlimit für das gesamte System bildet und die Reaktionsdynamik wesentlich verändert.

1.2 Sequence self-selection by the network dynamics of random ligating oligomer pools

One problem on the research on the Origin of Life is the step from short oligomers with random sequence of bases to longer, active complexes like ribozymes. Chemistry and non-equilibrium physics have found pathways for the formation of short strands, mediated by activated nucleotides or activation agents. It has also been shown, that medium length autocatalytically active complexes can form reaction networks necessary for downstream reactions and the emergence of more sophisticated systems. The intermediate step is difficult though: Short random sequences need to be extended, while introducing a reduction of sequence entropy, that does not inhibit double-stranded complex formation. This pre-selection step is necessary due to the vast sequence space of oligomer strands and must conform to an Origin-of-Life-like environment. The ancient reactions only depended on the inherent parameters of the oligomer strands such as the hybridization to double-stranded complexes and basic physical mechanisms such as temperature changes.

Although short strands likely emerged by a non-templated polymerization-like extension mechanism, passing sequence information from one strand to another is assumed essential in longer strands. Chemical ligation is probable in the context of the Origin of Life, but the yield and specificity is known to be low. As a model for said mechanism, this study utilizes an evolved DNA ligase to facilitate the ligation reaction of the model-oligomer (DNA) without introducing a sequence bias. After 1000 temperature cycles, necessary for the dissociation and rehybridization of elongated strands, the resulting reaction products showed several distinct features of selection. Compared to random strands of the same length, the set of emerging reaction products have a significantly reduced sequence entropy. The emerging strands can be categorized in two groups: A-type and T-type strands with a ratio about 70:30 % of each base. As strands can only act as a template or substrate in the templated ligation reaction when they are in the single-stranded conformation, this selection inhibits the formation of self-folding oligomer products. At the same time the formation of double-stranded complexes is not reduced, as the A-type and T-type groups are predominantly sets of reverse complement sequences. Selecting only the most common subsequences from each group as a new starting pool shows a greater fitness of the selected pool for the emergence of new ligation products compared to a pool of random strands. While the analysis of the complex types and the exact dynamics is limited in the experimental

system, all parameters can be accessed and analyzed in detail in a closely related theoretical simulation. This simulation found a distinct relation of the major reaction rates in the system: the hybridization on-rate, the ligation-rate and the hybridization-off rate. With those parameters the emergence of a local minimum-maximum feature in the concentration-over-strand-length analysis could be identified and the feature's appearance predicted in the experiment. The understanding of the dynamical regimes of the reaction also led to the understanding of length-dependent dominant growth modes of the product strands. Analyzing the emerging 24mers with a simple simulation based on these growth modes produced the same common ligation site sequence pattern `ATAT`. A small bias towards the self-complementary `AT` sequence motif at the 3'-end of strands in the original random sequence 12mer pool was identified as the likely cause.

This study found,that longer product strands in this simple model system had a reduced entropy while retaining the important hybridization properties of the oligomer strands. The dynamics of the elongation are dependent on the microscopic rates of complex formation, ligation and dissociation, while the temperature cycling frequency imposes a rate-limit for the entire system and substantially changes the reaction dynamic.

Contents

2 Introduction

2.1 Background

In all of science, coherent concepts as well as complete models and explanations are the ideal results of research. This is also the case of the Origin of Life, and the story most researchers are working on is a unified theory of evolution. Starting from nothing but the simplest of molecules in the prebiotic world, and ending up with a system, that could be classified as "alive" (whatever that means in detail - because as it turns out, it is difficult to precisely state what makes a system alive[i]). One way to tackle this enormous and interdisciplinary field of research is splitting the major theory in manageable questions. An example of a prominent question would be: How could the first amino acids emerge from simple atoms and without precursors or more complicated molecules? A possible answer was given by Miller and Urey [85] in 1953 with the help of their famous experiment. They enclosed water, methane (CH_4), ammonia (NH_3) and hydrogen (H_2) in a sealed vessel and applied a high voltage electrical discharge similar to lightning. The reactants in the vessel formed at least five different amino acids, which were identified by paper chromatography.

This example is a rather chemistry-based question on the way to the Origin of Life, but chemistry and physics usually go hand in hand in the real world. A question that can be explained by physics would be the "concentration problem" [33] during Early Earth, for which "geochemical extrapolations indicate a dilute prebiotic ocean" [7]. The products from the Miller-Urey experiment must have had VERY low concentration, which in turn must have resulted in an incredibly low reaction rate with other substances. Mast *et al.* showed in 2013, that water-filled elongated compartments in submerged porous volcanic rock with temperature gradients are able to accumulate species in diluted solutions up to several orders of magnitude [83]. The scenario of those so-called hydrothermal micro-environments [102] can be found even today, for example in the so-called "Lost City" close to the mid Atlantic Ridge [69]. The underlying physical effect is known as thermophoresis, which itself is caused by different microscopic effects such as convection and electrostatics [4, 15, 82].

This interplay of different physical phenomena enables far greater reaction rates that are impossible without such an increase in concentration. Most importantly, those environmental conditions and physical processes are all rather simple, like a difference in temperature or a difference in density. Today's evolved cells function through multi-step and multi-component metabolisms like glycolysis in the Warburg effect [57] or by specialized enzymes like ATP syn-

[i] being **alive**: With "Defining Life" Steven Benner gave a summary of the typically employed definitions of life in his 2010 published review [11]. He starts by describing a conference with the goal of defining life in a simple, short way. The scientific consortium found as many definitions that met those criteria as contradicting examples. And this is essentially the state of the art; some properties of living systems are straight forward, like the need to reproduce, but those properties need intricate additional specifications to be universally applicable. E.g. the statement "life is the ability to reproduce" falls with an example from the above mentioned conference, that a single bunny, despite being undoubtedly alive, cannot reproduce by itself. Consequently, the definition of "alive" is still complicated and might vary in different scientific fields. The "NASA definition" is often used in the context of the Origin of Life and captures the idea, while not excluding some critical points: a "self-sustaining chemical system capable of Darwinian evolution" [11].

thase which provide energy-rich adenosine triphosphate (ATP) [17]. But on Early Earth, such advanced and complicated physical "machineries" were simply not existent (yet). Therefore, all early chemical reactions needed to be simple, without a lot of intermediate steps. The physical mechanisms and processes needed to be simple as well, and only rely on the environment and the reactants themselves [6, 9, 82, 90, 97, 126].

For research on the Origin if Life, it must be noted that non of the results can be verified, as THE Origin of Life on our earth happened an almost unimaginable long time ago, and it is impossible to reconstruct the exact processes, molecules and reactants. For dinosaurs [20], the proof of existence is comparably simple when a giant, complete skeleton is found and reliable analysis methods like C-14-carbon analysis [2] reveal an age of several million years [19]. For small information carrying polymers, like a possible pre-RNA, this is impossible. While there are artificial ways to fossilize DNA [92, 132] this was not applicable in an Early Earth scenario. Additionally, the predecessors of RNA and DNA might have been completely different molecules than expected or emerged by pathways, that both might have since gone extinct [60, 61, 73].

Therefore, research on the Origin of Life is research on *AN* Origin of Life [41, 73]. Usually, results describe in detail one of the intermediate questions as shown above. One open question, which is in comparison to other problems like the emergence of the first amino acids (see above) still very much open to debate, is the transition from short random sequence oligomers to longer oligomers capable of enzymatic activity. The question is part of the "RNA world hypothesis" [27, 94, 126] with the core concept, that RNA can catalytically help its own reproduction [6, 56, 65]. These so-called ribozymes have a minimal length of 30 to 41 bp [14, 110] and thus a sequence space of more than $4^{30} \approx 10^{18}$. However, the subset of functional and yet catalytically active sequences is very small [39]. The high number in the variety of sequences would not allow nature to assemble the required structures from mono- or multimers by "trying". Consequently, the evolution from single nucleotides to networks of functional sequences must have been followed a selection mechanism towards lower entropy systems. Here, the elongation and a selection might have happened at the same time, presumably in the same process.

For elongation of short polymers there are two basic mechanisms called polymerization and ligation. In the most basic form of polymerization, a single strand is extended by the addition of single bases, often in the context of activated nucleotides. In templated polymerization a double-stranded section of a two-strand-complex is used as the so-called primer. Starting from that primer, the single-stranded part of the complex is replicated in a base-by-base fashion (by a polymerase). In contrast for ligation, two so-called substrate strands hybridize (partially) on a third template strand and are then connected (by a ligase). For both elongation mechanisms medium-length strands are necessary. Their formation is well understood by means of self-activated random ligation [58, 59, 116], which results in a pool of random sequence medium-length RNA- or DNA-like oligomers [18, 40, 66, 87].

The issue with polymerization reactions as the major elongation mechanism are the **error catastrophe** [ii] and the **tyranny of the shortest** [iii]. Therefore, an elongation reaction with a polymerase will yield a quickly decreasing, long tailed, very likely random sequence oligomer length

[ii] **error catastrophe:** Each replication mechanism has an error rate > 0, which means that the template strand is not rebuilt 1:1, but slightly altered (for example by ligating a strand that includes a single non-complementary base). By that mechanism the function of the replicated faulty oligomer might be lost, especially if the altered region is the active site of e.g. a ribozyme. If the error rate is low compared to the replication rate and the degradation rate, the original template will survive compared to the altered replicates. But above a certain error threshold the original template will disappear and all replicates are faulty, altered sequences [36, 38, 114, 115].

[iii] **tyranny of the shortest:** Under selection pressure for fast growth, replication by polymerization can lead to an increase of the replication rate at the cost of oligomer product length. Mills, Peterson and Spiegelman demonstrated, that RNA polymerases isolated from bacteriophages could replicate their intact RNA. The resulting molecule lost

distribution, as argued above. In the desired step from medium length random strands to longer length "less-random" strands, polymerization reactions are therefore unlikely to be successful.

Ligation reactions on the other hand have comparably low error rates, pass (sequence-) information from the template strand on to the substrate strands and typically don't suffer from the two error modes mentioned above. In the context of Early Earth, ligation and polymerization reaction are presumably equally likely. Both rely on either activation agents like EDC [3, 35, 113] or activated oligomers/ nucleotides [42, 76]. They also both rely on the hybridization of the reverse complement sequence onto the substrate strand, an internal property of the polymer itself. Those described properties fit the prerequisites assumed viable on Early Earth, as all driving mechanisms are simple and only depend on the environment or internal properties of a reactant.

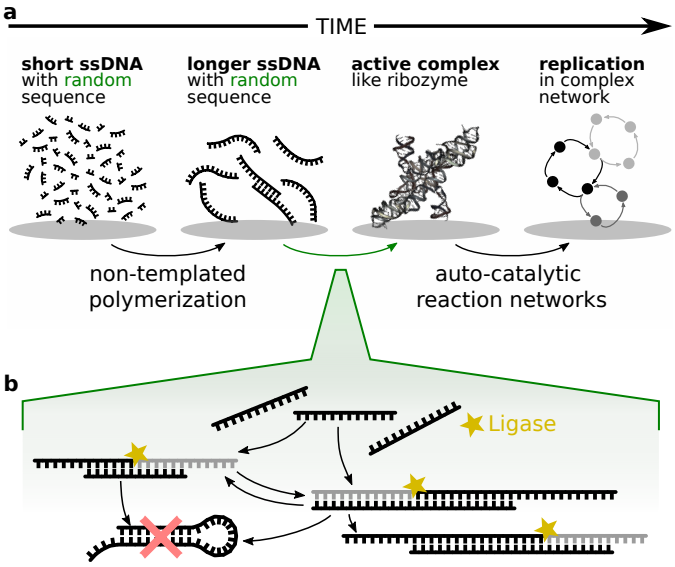

Figure 2.1 Evolution of oligomer ensembles and complexes:
a Schematic timeline of events at the dawn of life: While the formation of short oligomers and the selective capabilities of protein and enzyme networks is well documented, there is little known about the intermediate step.
b Processes expected to occur in random sequence oligomer pools. Templated ligation plays an important role in preserving sequence information in offspring oligomer sequences.

On a hypothetical timeline of Early Earth, short random oligomers presumably existed before and at least concurrently to the hypothesized simultaneous elongation and sequence selection of strands. While the elongation of strands to longer oligomers still need further selection steps to act as catalytically active complexes, those oligomers likely promoted hybridization of multiple strands into double-stranded complexes, simply by their length and the elongation process that relies on double stranded complexes. Reaction networks made from such double-stranded complexes undergo additional evolutionary steps [101] and lead to collaborative network-dynamics [5, 54, 124]. At one point those networks presumably mimicked ribozyme- or enzyme-like functions and evolved further towards replication networks and ultimately cells (Figure 2.1).

85 % of its original length but replicated 15 times faster, which indicates a higher efficiency for the interaction with the replicase enzyme [67, 86].

2.2 The experiment

The starting point of the experiment is the presence of short, random sequence oligomers[iv] that can form double-stranded structures by Watson-Crick basepairing [28] (see Section 5.1) and a ligation mechanism. For the first ligatable complexes to form, at least three single strands need to hybridize, with one as the template strand and two substrate strands, as shown in Figure 2.1b. If the experiment works as intended and includes and elongation as well as a selection mechanism, emerging oligomers consist of several ligated original stands and have a lower entropy than random strands of the same length. Due to limitations in the systems temperature and the length-dependent ligation performance of the ligation mechanism, the experimental system needs to be designed with certain properties.

- **DNA instead of RNA**
  In contrast to the presumed ideal oligomer contender RNA, as assumed in the RNA-world hypothesis, this study utilized DNA as the model oligomer. DNA is more robust in a laboratory environment, especially at elevated temperatures and salt conditions, but still shows properties directly comparable to RNA [107] that are likely also found in early oligomers, such as temperature-dependent base pairing and formation of 3D complexes. Formation of double-stranded complexes is common in DNA, e.g. the human genome is typically stored in a supercoiled conformation. The basis is the so-called Watson-Crick basepairing, which describes the formation of bonds between complementary bases (see Section 5.1).
- **Ligation by enzyme**
  The formation of a linkage between the two substrate strands in an Origin-of-Life-context was presumably happening due to activated oligomers or activation agents, as argued above. In comparison to the ligation by an (evolved) enzyme, chemical ligation is slow, error prone and yields all kinds of intermediate products [35] that might interfere with further reactions down the line. This is no problem for the real Origin of Life, as there was plenty of time and volume for such reactions. Non-functional and incorrect replicates might have simply died out at some point. Anyhow, in a laboratory environment enzyme driven ligation is fast, doesn't form intermediate products, has a low error rate and (ideally) no sequence bias (see Section 5.5). This makes the use of an enzyme the ideal model to study even extended timescales inaccessible by chemical ligation. An evolved Taq DNA ligase is therefore used as a surrogate ligation-mechanism, which allows focusing on the inherent properties of the random pool and not on the chemical mechanism of ligation. In a ligatable complex, the ligase forms a standard phosphate backbone linkage that is identical to all other backbone-linkage in the substrate strands.
- **Characteristic length**
  The length of short oligonucleotides was probably a long-tailed size distribution with only few longer strands and a majority of short strands, that stems from the polymerization reactions that lead to said oligomers [83]. As described above, it is impossible to distinguish if a strand was ligated from to shorter strands, as there is no feature to identify the ligation site. This is especially true, when a strand can be made from several different monomer lengths (e.g. a 20 nt strand could be made from 4+16, 5+15, 6+14, 7+13, ...). Therefore, it is convenient to start with a fixed strand length. All longer strands are part of a discrete length distribution and ligation sites can easily be identified. Additionally, longer strands have a significantly larger sequence space. This effectively decreases the concentration of strands

[iv] The non-enzymatic synthesis of nucleotides and the formation of short oligomers in an Origin-of-Life-like setting has been shown before [53, 72, 79, 93, 95].

with the correct hybridization sequence (also see Section 5.3). The evolved ligase enzyme is derived from a bacterium in which it repairs single-strand nicks in double-stranded DNA. A high specificity of the ligase is necessary, for which it needs a minimum double-stranded region. Because the sequences are only made from two of the four bases common in DNA (`A` and `T`) (see next point), the melting temperature of short duplex strands is very low. It's in fact so low, that the ligase is not active anymore. Therefore, a minimum length of 12 nt DNA is chosen and yields successful ligation reactions (see Section 5.3.2 and 5.5).

- **Limiting the sequence space**
  As mentioned above, the sequence space spanned by the length of the strands is essential in the hybridization dynamics of the system. The elongation mechanism of templated ligation relies on the presence of duplex complexes emerging by hybridization from the originally single-stranded pool of strands. While there was an immense amount of time at the Origin of Life for trial and error in the complex formation, there is a simple way to increase the hybridization on-rate k_{on} in the laboratory without significantly inhibiting or altering the binding mechanism of the DNA strands. Building the DNA from only two instead of four bases results in a so-called binary alphabet of sequences. The length-dependent sequence space shrinks by a factor of 2^{length}. As an example: 12 nt strands made from bases `A` (adenosine) and `T` (thymine) only reduces the sequence space from $4^{12} \approx 16.7 * 10^6$ to $2^{12} = 4096$. An additional upside of this limit in the sequence space is that the "monomer" pool can be thoroughly sampled with next generation sequencing (NGS). Sequence and pattern analysis in binary alphabets is also simpler to visualize and understand (see Section 5.13).
- **Reaction tube instead of hydrothermal microenvironment**
  The presumably closest experimental environment in comparison to an Origin of Life setting would be a liquid-filled pore in a temperature gradient [1, 62, 68, 82, 83, 88, 89, 105]. This setup wouldn't only accumulate diluted reactants but also cycle its components through higher temperatures from time to time. However, this might introduce additional selection mechanisms, that might be difficult to distinguish from mechanisms based on the hybridization of strands. Therefore, experiments are conducted in a reaction tube, where the entire sample is heated and cooled homogeneously, inhibiting convection, condensation and accumulation of diluted substances.

On the first glance some of the set limitations seem quite drastic: Utilizing only two of the possible four bases? Will this result in accurate or at least useful results? And only one set length of oligomers? Doesn't this limit the possible conformation even more?

As argued up until here, there have been no comparable experiments yet. It is unclear if a reduction in sequence entropy occurs, if said reduction is due to emerging sequence patterns, or if it's simply a reduction of the amount of random strand motifs. The dynamics of the strand elongation are unclear as well. Would the length distribution of the emerging longer strands follow a simple exponential decay, or would the length distribution be more complicated? Those simple limitations can easily be relaxed in future studies if the results of this study can find first useful interpretations.

In experiments with short polymers an electrophoresis based analysis method is very common. This method is explained in detail in Section 5.11, but the basics are that longer polymers have a lower velocity in the 2D sieve-like gel. The polymers are pulled through the gel by an electric field and are separated by length. If the experiment described here leads to the emergence of longer strands (in a measurable quantity), the polyacrylamide gel electrophoresis (PAGE) anal-

ysis will show them as bands at longer lengths than the origin 12 nt long "monomer" strands of the original pool.

And indeed, Figure 2.2a shows that after 200 temperature cycles the first 36 nt and 48 nt long strands are observable. For more temperature cycles, the amount of longer strands growths

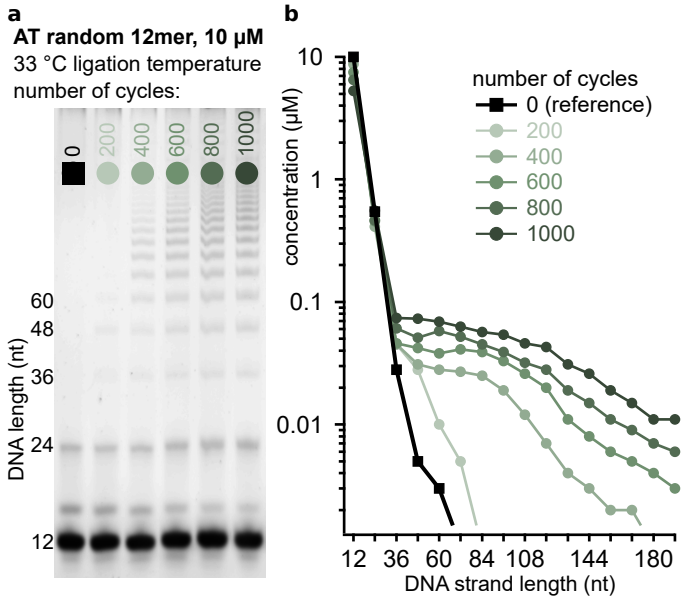

Figure 2.2 Random sequence templated ligation yields oligomers of 12xN length:
a Photo of a denaturing PAGE gel with SYBR post-stained ssDNA. For higher cycle times more oligomers are build.
b Custom LabView software can extract the concentrations of oligomer products in each lane. The concentration as a function of length is not following a powerlaw behavior as it would be expected for random ligation without additional acting effects.

and even very long oligomer strands can be observed (length of 192 nt and more). Panel b in Figure 2.2 shows the quantified concentration of each band. This method is explained in detail in Section 5.6 and highlights the non-monotonously and especially non-exponentially falling length distribution of the emerging product strands.

By some mechanism, presumably as shown in Figure 2.1b, longer strands are ligated from the shorter strands. The non-trivial length distribution already indicates, that the extension dynamics depend on more than one parameter (like concentration) and is most likely also dependent on the specific strand distribution itself.

The following chapters will analyze the next generation sequencing (NGS) data and find a reduction in sequence entropy that stems from multiple effects: an auto-selection of non-folding strands and sequence patterns at and in between ligation sites. Additionally, the influence of experimental parameters on the elongation dynamics such as ligation temperature T_{lig}, melting temperature T_{melt}, ligation time step lengths t_{lig} and sequence space are tested and analyzed in detail. Finally, the results of this study are correlated with the results of a collaborative study simulating the microscopic hybridization and ligation rates of a templated ligation-based setup. Here, the importance of the ratio of the length-dependent elementary reaction rates and extension mechanisms could be identified.

3 Results

The transition from short, random sequences to a set of autocatalytically active complexes is based on a combined extension- and selection-mechanism. In the introduction Section 2.2 it was first shown, that the experimental setting works and leads to the formation of longer strands. The dynamics of the experiment are analyzed in details later, starting at Section 3.7, but this elongation is not consequential in this context, if there would be no selection. Therefore, the next chapters use data obtained by next generation sequencing (NGS, also-called deep sequencing) to analyze the original 12mer sequence pool, a reduction in sequence entropy for longer strands, and the underlying mechanism, that lead to the reduction.

3.1 Nomenclature

Some terms in this work do need additional explanation, because the definition used here might be different than the original or usual use case.

- **Monomer:** Usually, a monomer is the shortest single building block of an oligomer, such as a single sugar with an attached base is the monomer used to extend existing polymers like RNA or DNA in a standard polymerase chain reaction. In this work, a monomer is also the smallest building block of the reaction, but this smallest building block is (unless specified otherwise) a 12mer single-stranded DNA strand. To highlight the difference and stress the reference to the 12mer DNA instead of the single sugar-base molecule, monomers might be written with quotation marks: "monomer".
- **Complex:** A complex is a double-stranded hybridized structure made from one, two or more single strands. In the simulation section, complexes describe all types of single- and double-stranded strands, that can be hybridized or single-stranded.
- **Oligomer (product):** Complexes made from at least two "monomers" that are ligated by the Taq DNA ligase are called oligomer or oligomer product.
- **Ligation Site:** The bond in between two complexes, two "monomers" or a complex and a "monomer" in an oligomer product.
- **Subsequence:** A sequence in a complex with the length of a "monomer" and the position in between two ligation sites or the strand-start/-end and a ligation site. In the context of sequence motifs, it refers to the region around the actual bond ($\pm$ 1 to 6 bases).
- **Submotif:** A sequence of a length x. In contrast to the subsequence, a submotif can start and end at any position on an oligomer or "monomer".
- **Base:** The base of a single nucleoside, like the *purines* adenine and guanine ot the *pyrimidines* cytosine, thymine (and uracil in the case of RNA). In case this study mentions specific bases, they are highlighted: `A` for a nucleoside with adenine. The sequence in longer individual strands is also highlighted: `ATTATATT` for a specific 8 nt sequence motif.

The theory, methods and results presented in this thesis are based on the study in ref. [74], published in PNAS in 2021. The overall argumentation and images are therefore comparable, although extended with new analysis, findings and additional results and interpretation.

3.2 Entropy reduction in oligomers products

Whenever a random ensemble of possible states is reduced to fewer states, the system entropy is lowered. This isn't different for the emerging oligomer sequences, if the extension is accompanied by an selection mechanism. For plotting differences in entropy, the Shanon entropy [111] nomenclature is used, with adaptions introduced by Derr *et. al* [29]: initially, this formulation was used to describe the variety of amino acids in proteins. For the comparison of different ensembles it was extended to incorporate information about the sequence length s in a strand of length L:

$$H_k(s) = -\sum_i p_i \log_2(p_i) \tag{3.1}$$

Here, i is the index of a unique string of length k, with k in a range of 1 to (basecount)^k. p_i is the frequency of the ith k-mer in s, as described by Derr. In order to quantify the entropy reduction, the maximum entropy value per subsection has to be defined. The maximum entropy could simply be produced by simulating a random sequence pool made from a large number of computer-generated random sequence strands. Alternatively, a set value could be calculated by assuming perfectly random strands (as it is done in the discussion Section 4.6). But here, a possible reduction compared to the randomness of the pool is of interest. Therefore, the entropy of all sequenced 12mer strands is calculated as the baseline for each 12mer subsequence position. Entropy values extracted from the experimental (longer) sequence data are then divided by this maximum entropy to give a relative entropy. A value of 1 describes an ensemble made from the same random sequences without any entropy reduction and thus a maximum entropy; a value of 0 describes the reduction to only one specific sequence.

To compare the resulting oligomers and the original pool, the entropy for all oligomers are calculated as the mean entropy per 12mer subsequence. As Figure 3.1 a shows, the 12mer random sequence pool divided by its own entropy is 1, as expected (first datapoint at length 12 nt). Starting from 24mer oligomers, the mean relative entropy decreases, and has a lower value for

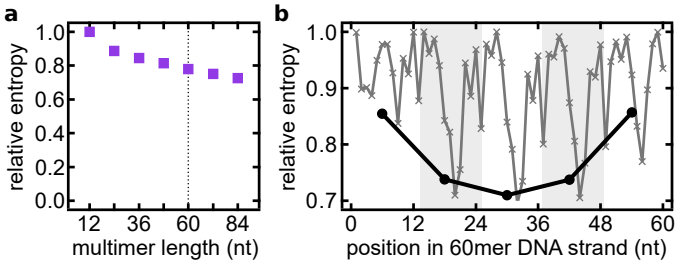

Figure 3.1 Reduction in sequence entropy for oligomer products:
a Relative entropy reduction of oligomers. While 12mers have no reduction of entropy, as expected for a random sequence ensemble, longer strands show significant reduction of ensemble sequence entropy. Longer strands show a greater reduction.
b Analyzing the 12mer subsequences and the entropy reduction per base reveals that in the strand center sequences have a lower entropy compared to the start and end sequences. Single positions show their lowest entropy also in the center of the 12mer subsequences, in between ligation sites.

longer oligomers, with 84mers having a relative entropy of about 0.75 compared to random sequence 12mers. The entropy reduction per 12mer subsequence is dependent on the position of the subsequence in the oligomer, as exemplarily shown for 60mers in Figure 3.1 b. Center sub-

sequences have a significantly lower entropy compared to start and end of the oligomer. This feature is similar for all lengths of oligomers. Comparing the entropy for single-bases depending on their position within the oligomer also shows a pattern. While there seems to be only minor reduction of entropy at 12mer subsequence junctions, the entropy in the center of the subsequences is noticeably lower. This pattern is seen for all center sequences and to a smaller extend for the start and end subsequences.

As the initial 12mer pool of DNA strands has the same entropy as an artificial computer-generated random sequence ensemble of strands with the same length, this suggests the absence of a bias of certain sequence motifs on the scale of the monomers. The clear reduction in 12mer subsequence variety within longer oligomers points to the existence of an underlying selection mechanism.

3.3 Base composition of oligomer products

Despite the above shown significant reduction in sequence entropy, the amount of different sequences is still very large. But said reduction is probably due to a reduction in sequence motifs, that ultimately make up the whole strand. Because there are only two bases in a binary alphabet strand, the formation of patterns are likely to change the base-composition of the strands. Plotting the A-to-T ratio of all 12mers of the original pool results in an about binomial shaped distribution. This distribution is expected and can be thought of as a probability distribution of flipping a coin, with one side corresponding to `A` and the opposite to `T`. And 12 coin-flips create one strand. A set of many strands will form a binomial distribution, as it is most likely, that the majority of strands has about 50 % of each base, while only few strands are made almost completely from one base. Here, the frequency over base-composition plots are treated like probability density functions (PDFs) to allow for comparison of different oligomer lengths. Every distribution is the sum (in PDFs, it would be an integral to account for all possible states on the x-axis) of all probabilities P_N of strands of length N to find a certain base-fraction $\mathrm{d}_{A:T}$:

$$\int P_N(A:T)\mathrm{d}_{A:T} = 1. \tag{3.2}$$

Remarkably, the emerging 24mer strands show a very different distribution. Instead of a peak at about 50 % `A` and 50 % `T` the distribution is bimodal, with peaks at about 0.35 and 0.65 A:T content. The two peaks are characterized as A-type and T-type sequence groups.

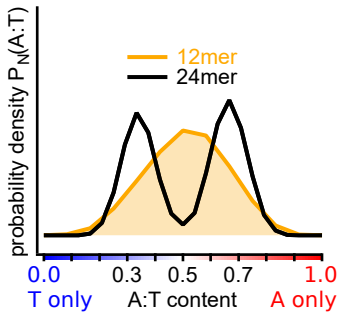

Figure 3.2 Base composition analysis, A:T-ratio of 12mer monomers and 24mer oligomers: 12mer strands that make up the pool have a binomial shaped AT-distribution with its center at about 50 % `A` and 50 % `T`. 24mer sequences show a bimodal distribution, with A-type sequences (about 70 % `A` and 30 % `T` and *vice versa*).

The drastic change in the form of the distribution points to an underlying selection mechanism in the formation of the 24mers. As mentioned in the introduction, DNA and other oligomers can form double-stranded structures, like the well known double helix structure of DNA. But strands can also form double strands with themselves, like so-called hairpins. A strand in a hairpin formation has a high k_{on}-rate compared to two free strands, because of the close special proximity (see Section 5.2). Hairpins in closed form are then unlikely to participate in templated ligation reactions, as they cannot hybridize to other strands and thus can't act as substrate nor template. In the 2018 publication of Tkachenko and Maslov [119] they use a kinetic model for a hypothetical oligomer system with templated ligation as the extension mechanism. For this study and the corresponding publication [74] this model was extended to incorporate a possible formation of hairpins. The details are described in Section 5.9.2, but a short summary, every strands is analyzed for it's maximum length hairpin and the energy of formation is calculated based on real-world parameters. Next to the formation of complexes by two or more single-stranded complexes, the self-folding is another pathway to double-stranded complexes. Starting from an initial small amount of seed 24mers with random sequence (binomial shaped distribution) as template seed material, emerging 24mer strands transition into a bimodal shaped distribution.

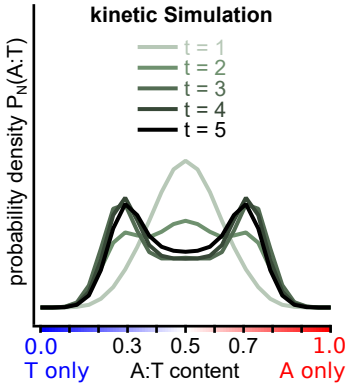

Figure 3.3 Kinetic simulation of 12mer templated ligation induced base-composition shift: Kinetic simulation of 12mer strands under templated ligation reaction, adapted from [119]. Because strands can fold on themselves and form hairpins with low dissociation or melting rates, oligomers with self-complementary sequence motifs become less abundant. The binary alphabet sequences with 50:50 A:T ratio become especially less abundant over time, as they are more likely to have a reverse complement internal sequence motif.

The simulation calculates the strand concentrations for several time steps, which show rapidly increasing strand concentrations for later time points. To compare the distribution of different the time points, the distributions are again treated like PDFs, as shown in Figure 3.3. Similar to the experiment, the formation of hairpin structures in the simulation hinders a strand from taking part in further templated ligation reactions, both as template and as substrate. In a binary sequence alphabet, the probability to find a sequence motif that can form an internal hairpin is highest for a similar amount of complementary bases in a strand, namely an A:T fraction of 0.5. Strands with a A:T fraction of 0.3 or 0.7 can still act as template and substrate, but are unlikely to form hairpins. In a system where templated ligation is the only basis of an potential selection mechanism, the strands with about as much `A` as `T` are therefore in a disadvantage and aren't replicated or incorporated as often. As a result, the original binomial distribution of 24mers, that is extended with a pool of random sequence 12mers, shifts to a bimodal distribution, again with about A:T ratios of about 70:30 and *vice versa*.

At a closer look, the distribution for 24mers in Figure 3.2 is not completely symmetrical, but with a slight bias towards too much A-type sequences. For the emerging oligomers, this even more pronounced: A-type sequences in Figure 3.4 start to be significantly more abundant in comparison to T-type sequences. A detailed analysis of the original pool reveals, that while it

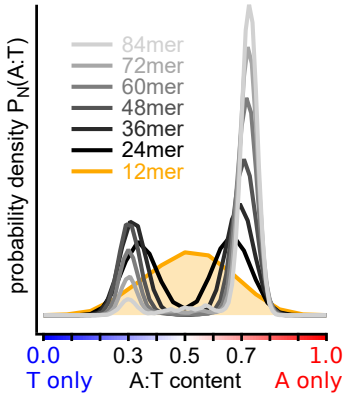

Figure 3.4 Base composition analysis, A-type bias for long oligomer products:
Starting from 36mers, the bimodal distribution shifts the peak heights at 0.3 and 0.7 A:T fraction. A-type sequences become significantly more abundant compared to T-type sequences.

is remarkably close to the expected binomial distribution, the measured distribution is a little shifted. Figure 3.5 shows the 12mer A:T distribution compared to a computer generated random sequence distribution. Both experimental curves for the uncycled reference sample and after 1000 temperature cycles is shifted towards too much A-type sequences. As the majority of sequences have about 50 % of each base, the difference to a random sequence distribution is only about 10 %. But sequences with either a high number of `A` or `T` are already rare in comparison, and the shifted curve in Figure 3.5 b shows that especially T-type sequences are by 10 % to 500 % underrepresented. The relative error becomes large for the high ratios, due to the binomial distribution. For the 88009 analyzed 12mer strands in the NGS data set after 1000 temperature cycles, there are only 5 `A`-only and no `T`-only strands found, while about 20 would be expected for a binomial distribution with the same amount of total strands. The overall strand composition distribution indicates that the 12mer pool is indeed made from random sequence strands. But a similar composition could still be achieved by only 13 different 12 nt motifs. Therefore, a possible entropy reduction in 12mers is analyzed in Figure 3.6, similarly to Section 3.2. The entropy of the analyzed set of nearly 90000 strands does not differ from a random sequence distribution, as indicated by the gray circle. But on a base-by-base level, the strands do show a reduction in entropy towards the 3'-end of strands. In Figure 3.12 all strands have a distinct pattern of `AT` at the last two positions, which is indicated here by the local entropy reduction. The reason is likely a bias during DNA synthesis: strands are build in a solid state synthesis by *biomers* starting from the 3'-end. The reaction is supplied with similar amounts of the precursors for both `A`- and `T`-nucleotides. A reduction in strand entropy towards especially the 3'-end points to a bias in the synthesis reaction. Figure 3.6 suggests that at least the last three bases exhibit that bias, which is getting weaker the more bases are attached. This mechanism will be important again in the simulation of randomly templated strands that are only based on the pool here in Figure 3.17 (more details in Section 5.16).

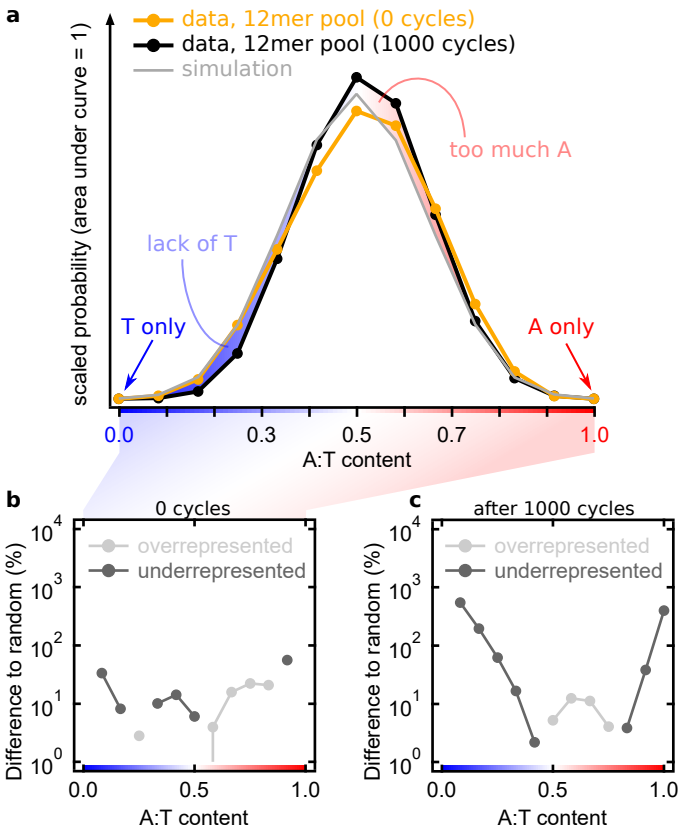

Figure 3.5 Initial 12mer pool base composition:

a Comparing the initial 12mer pool without temperature cycles to the remaining 12mers after 1000 temperature cycles and to a computer-generated random ensemble of strands reveals a slight bias towards A-type sequences for the random sequence pools. The 12mer uncycled pool is also different than the cycled 12mer pool. Subsequences with 6:6, 7:5 and 8:4 A:T ratios are overrepresented, while 4:8, 3:9 and 2:10 A:T compositions are underrepresented. **b** Sequences on either end of the A:T ratio are rare due to the binomial shape of the probability function. For a total strand count of 8533 analyzed 12mer strands (with 0 temperature cycles), there should only be about 1 instance of the strands `AAAAAAAAAAAA` and `TTTTTTTTTTTT` each, but during NGS analysis none are found. Strands with ratios of 3:9, 7:5 to 10:2 A:T are slightly overrepresented, while the other ratios are underrepresented. **c** For a total strand count of 88009 analyzed 12mer strands (after 1000 temperature cycles), there should only be about 20 instances of the strands `AAAAAAAAAAAA` and `TTTTTTTTTTTT` each, but during NGS analysis only 5 `A`-only strands are found. Strands with A:T ratios of 6:6 up to 9:3 are overrepresented, while all other strands are strongly underrepresented.

Extending this argument of a biased synthesis reaction, the analysis of sequence motifs for all positions in 12mer shows, that the abundant pattern `AT` is prominent at the last position. Instead of the expected 25 % abundance for each motif, `AT` has a frequency of almost 34 %. The motifs `TA` and `TT` are less abundant than expected with about 19 % and 20 %. The shifted binomial distribution for the A:T fraction is also reflected here. `TT`-motifs are rare in almost the entire strand, while the other three motifs are about equally abundant.

In Figure 3.7 the mean frequency of those four 2 nt motifs are analyzed for all strand lengths. For the initial 12mer strands, the motifs `AA`, `AT`, and `TA` are about equally abundant (`AA` at 26.6 %, `AT` at 25.9 %, and `TA` at 25.7 %) while `TT` motifs are less frequent with 21.7 %. For

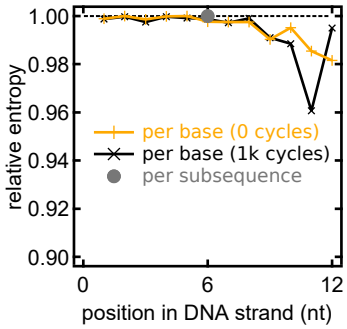

Figure 3.6 Initial 12mer pool sequence entropy:
For 12mer sequences the entropy is not reduced on the scale of the strand length (gray circle). Analyzing each position compared to a truly random sequence reveals a bias at the last few bases. After 1000 temperature cycles, the entropy is lower for the strand positions 10 and 11.

24mers and all longer oligomers the alternating motifs `AT` and `TA` both stabilize at a frequency of about 21 %. Despite an initial increase for the `TT` motif in 24mers, the abundance of `AA`-motifs growths to 49 % in 84mer oligomers and `TT`-motifs become very rare with less than 10 %.

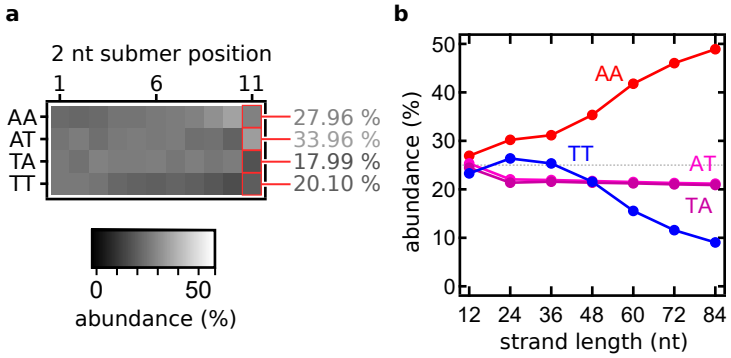

Figure 3.7 2 nt sequence motifs:
a Analyzing all four possible 2 nt long sequence motifs `AA`, `AT`, `TA` and `TT` for each position in 12mer reveals several local biases. The last 2 nt position at bases 11 and 12 has a frequency of almost 34 % for the motifs `AT` while `TA` and `TT` are found less frequent than expected (25 %).
b The total abundance of all four 2 nt motifs in the 12mer subsequences of the ligation products shows, that the fraction of the two alternating motifs `AT` and `TA` are about constant for all product lengths at about 20 %. For short products, abundances of `AA` and `TT` increase, but starting at 36mers, diverge with `AA` become significantly more abundant.

Long oligomer strands as shown in Figure 3.4 experienced multiple templated ligation reactions. Each templated ligation reaction includes the hybridization of at least three complexes to a ligatable complex. This process is statistical and could be compared to well known mathematical model of drawing one of many objects from a vast pool of objects. In case the pool is inhomogeneous and some objects are more common than others, the assembled complexes will be biased towards that specific object. As long strands are the product of multiple ligation reactions, a small inhomogeneity in the initial pool is amplified.

Section 5.9.1 describes a theoretical model that takes a bias towards A-type strands in the initial pool into account, to simulate the evolution of the A:T fraction distribution. The model

calculates the maximum hairpin length per strand and the number of possible non-overlapping alignments to get an expression relating the maximum hairpin length to the strand length. Approximating A-type and T-type distributions with a Gaussian curve each and introducing the composition bias β for one of them yields a distribution only dependent on β, the peak position and the oligomer length. Compared to the experimental data in Figure 3.4 the T-type peak is

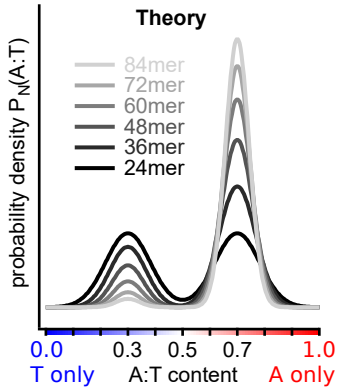

Figure 3.8 Theoretical description of A-type abundance in oligomers caused by a biased initial pool: The great abundance of A-type sequences in long oligomers can be described by an imbalance of the initial monomer pool (bias parameter $\beta = 2$). Here, an abundance of A-type strands is assumed.

not as low and the peaks only get slightly more narrow. In the data, the region with about 50 % of either base is especially underrepresented and the peaks are substantially narrower, which shows, that the bias in the initial pool is not the only reason for the biased bimodal base composition distribution.

The suppression of hairpins seen in the data and the kinetic simulation for 24mers also act on longer oligomers. Analyzing the A:T fraction of 72mers reveals and abundance of homo-A-type or homo-T-type strands. Figure 3.9 shows how A-type ("a") and T-type ("t") 12mer subsequences are usually either homo-type of only have one subsequence of the other type, mostly in the first or last position. Those strands build the major peaks at a:t-ratios of 0.3 and 0.7 of the distribution in Figure 3.9 b while, two small peaks mark the 2:1 and 1:2 a:t-ratio of 72mers. As expected for hairpins, the order of subsequences is homo-X-type and changes after either two, three or four subsequences to the inverse base-type.

Measuring the maximum internal hairpin length in all sequenced strands reveals a shift in the distribution of most likely internal hairpin stem length, as shown in Figure 3.10 b. The longest internal hairpin for all strands of a certain length start with a sharp binomial distribution around their most likely hairpin stem length. Starting from 48mers, the distribution develops a shoulder for longer and longer hairpin stems. This is due to the behavior described in Figure 3.9: the more subsequences in each oligomer are of the opposing type, the higher the chance that the oligomer can fold into a secondary structure. The abundance of homo-type oligomers causes the most likely maximum hairpin stem length to increase by only a factor of 1.89 while the strand length increases by a factor of seven.

Equation (5.4) describes the expected hairpin length as a function of the strand length, as shown in Figure 3.11 extrapolating the predicted hairpin stem lengths in oligomers. With the composition bias factor $p = 0.5$ the theoretical line crosses exactly the data point for 12mers, suggesting that the majority of 12mers are unbiased or only slightly biased, as discussed in Fig-

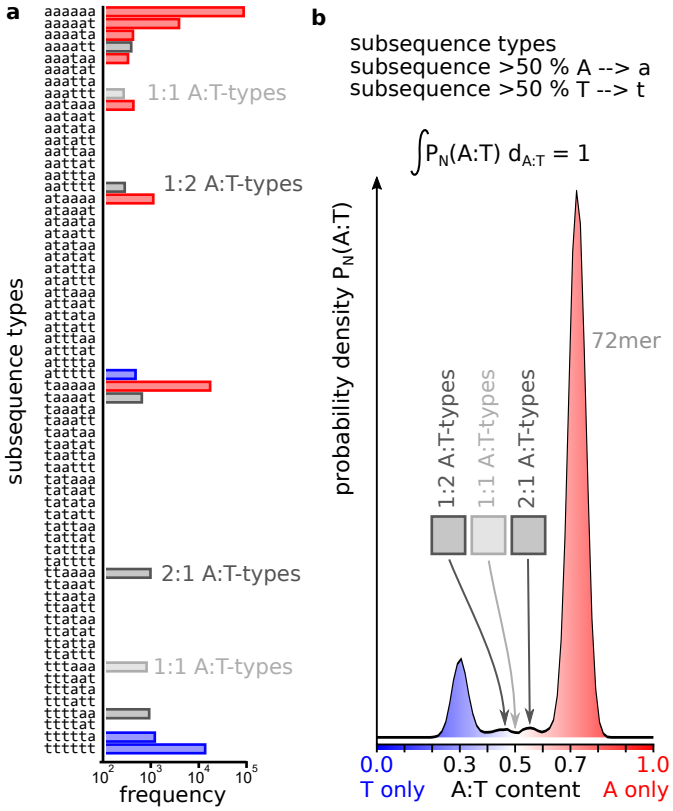

Figure 3.9 Subsequence order in 72mers:

a A-type 12mer subsequence noted as "a", T-type as "t". Most 72mers are either homo-A-type or homo T-type sequences. Some strands shift from one base-type to the other.

b For long oligomers, the PDF over the base-composition shows additional peaks that represent strands with an internal shift from A-type to T-type or *vice versa*.

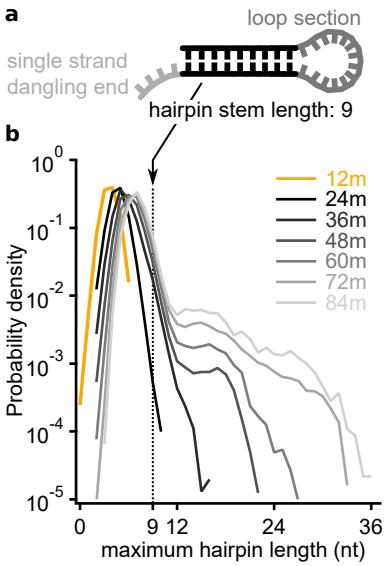

Figure 3.10 Most likely internal hairpin stem length:
a Schematic illustration of the internal hairpin formation.
b The maximum possible hairpin stem length for all strands grows only by a factor of 1.89 (from 3.7 to 7) while the strand length grows by a factor of 7 (from 12 to 84). For 48mers and longer oligomers the PDFs show an increasing amount of very long hairpins. Those are likely the rare non-homo-type oligomers shown in Figure 3.9b.

ure 3.5. Data points for 36mer oligomers and longer strands are on the line for a bias factor $p = 0.785$. This ratio is close to the maximum of the experimental A:T-ratio in oligomers, as shown in Figure 3.4.

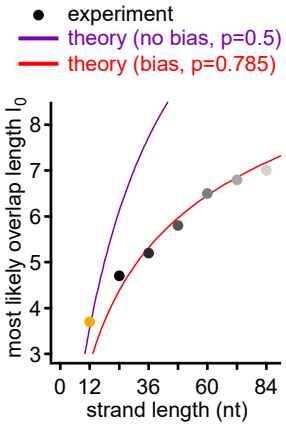

Figure 3.11 Most likely internal hairpin stem length modeled with base-bias factor:
The lengths of the most likely hairpin stem length as a function of the strand length are described by equation (5.4). This equation includes a factor describing the base-composition bias. 12mers lay on the unbiased curve, while oligomers are in good agreement with a bias factor of 0.785.

Overall, the elongation of 12mer monomers by templated ligation closely seems to follow a simple model introducing a composition bias for longer strands. The self-segregation into A-type and T-type sequences is caused by the inhibition of hairpin structures. Those strands fail to act

as template or substrate and are thus not taking part in the templated ligation reactions and are not part of oligomers or rebuild as templates. The compositional bias in the A:T distribution of the initial monomer strands likely leads to an abundance of A-type strands in longer oligomers.

3.4 Submotif sequence selection

Although the analytical and kinetic models discussed above include a composition bias and can therefore explain the entropy reduction by changes in the A:T distribution, the experimental data has narrower distributions. This suggest a stronger selection by an additional selection of sequence motifs in the already reduced sequence space.

Figure 3.12 marks the probability to find each base at a certain position within 12mer and oligomer strands. High probabilities of finding `A` are marked in red, finding `T` in blue and positions without a clear preference are marked white. As expected, 12mer strands are about random with a tint of red hinting the slight bias in the pool towards `A`. The last two position on the 3'-end are slightly red (position 11) and blue (position 12) indicating the already known abundance of the 2 nt sequence motif `AT` (see Figure 3.7a).

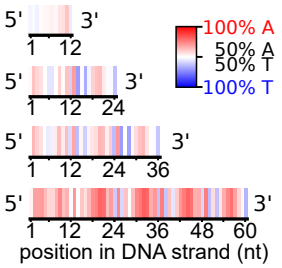

Figure 3.12 Base probability per position in monomers and oligomers:
Starting from the random sequence 12mer pool, its base sequence is about random with a little too much `A`, and a bias towards `AT` on the 3'-end. Longer strands show the emergence of `AT`-alternating patterns on the ligation site and a bias towards `A` in between ligation sites. For 60mers the uneven A-type and T-type distribution discussed above is visible with the overall red tinted base sequence.

In contrast to the rather flat base probability profile for 12mers, 24mer reaction products show a distinct pattern: at the ligation site, the sequence pattern is alternating between `A` and `T`. On the start and end of the strand, poly-A and poly-T motifs start to emerge. A similar sequence pattern can be seen for 36mers. For 60mers the bias towards A-type sequences is apparent in the overall red tint of the base probability.

For 12mers the sequence space is $2^{12} = 4096$ what makes it impossible to plot histograms for all sequences. But reducing the sequence motif length to 6 nt also reduces the sequence space to $2^6 = 64$ which enables plotting histograms and makes the motifs more readable at the same time. Comparing 6 nt sequence motifs on the ligation site to the regions in between ligation sites in 36mer oligomers (for better statistics, this can be done for longer sequences, but then the bias of A-type and T-type sequences is changing the result, see below for Z-score ligation landscapes) shows a comparable pattern to Figure 3.12. On the ligation site alternating `AT` sequence patterns are up to eight times more common than in between ligation sites (Figure 3.13). Between the ligation sites, pattern with poly-A and poly-T are up to six times more common.

The position-dependent sequence pattern of poly-A, poly-T, and `AT`-alternating might be one reason for the difference between simulation and theory in Figure 3.4 and Figure 3.8. In Fig-

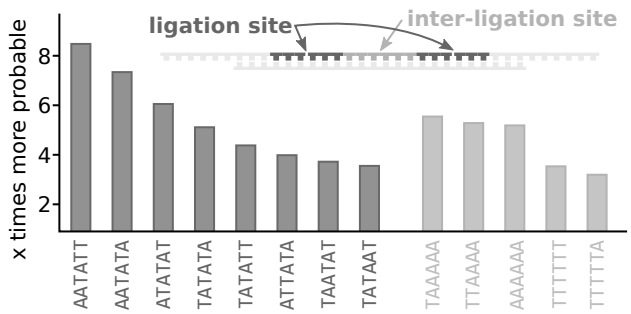

Figure 3.13 Ligation site and inter-ligation site sequence motifs:
`AT`-alternating sequence motifs are up to eight times more common on ligation sites for 36mers, than between ligation sites. On the other hand, poly-A and poly-T motifs are very common in the center strands, but not on ligation sites.

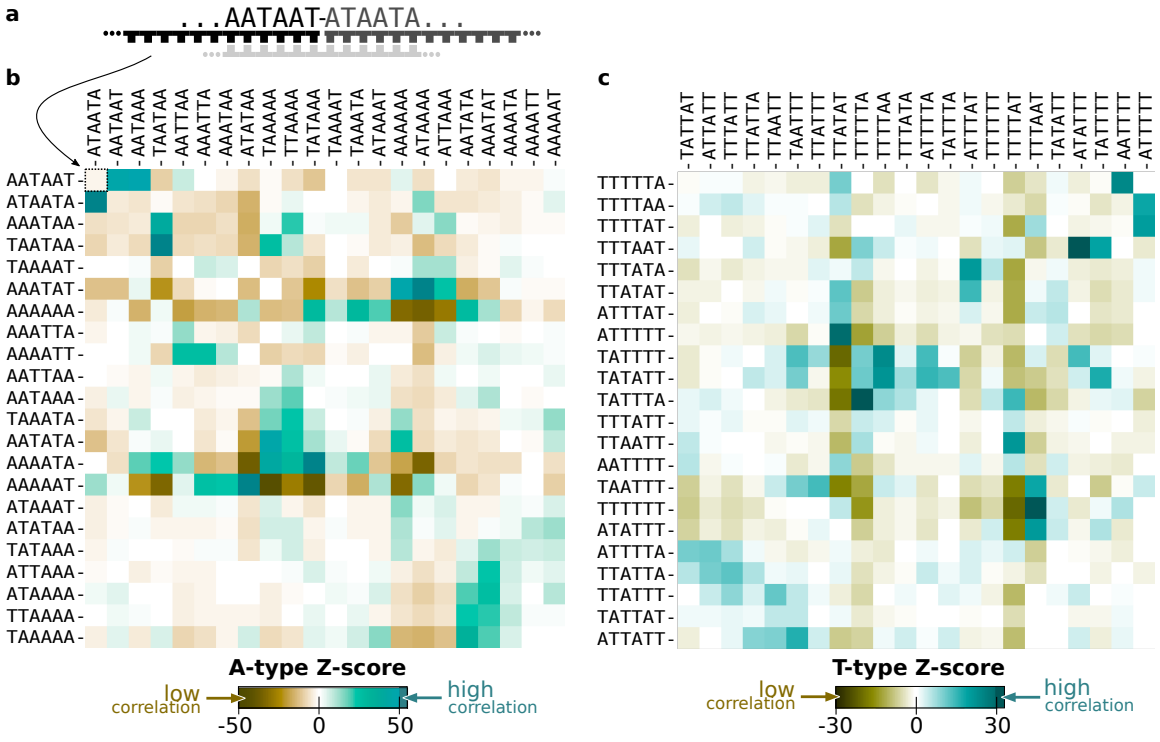

Figure 3.14 Z-score ligation site sequence motifs in A-type and T-type groups:
a Sketch of the ligation site position within double-stranded complexes. The light gray strand is the reverse complement sequence template strand.
b Z-score ligation site sequence landscape for A-type oligomers: Common motifs are colored teal, uncommon motifs ocher. Again, ligation site motifs are governed by the alternating `AT` motif, while poly-A and poly-T are rare. In those long strands first fingerprints of ligation site sequence shifts appear in the form of poly-A ligation site strands, enclosed by alternating `AT` motifs (bottom right corner of the plot).
c Z-score ligation site sequence landscape for T-type oligomers: Overall T-type strands show a similar preference for alternating `AT` ligation site sequence motifs. Due to the abundance of A-type strands in 72mers, the overall Z-score is lower.

ure 3.14 b and c the Z-score of all internal junction between the 4^{th} and the 5^{th} subsequence in long 72mer oligomer products are shown. The Z-score is usually used to compare datasets with different scales and distributions, that are then shifted and normalized. Positive Z-score values lay above the mean of the distribution, negative values are below the mean of the dataset. For a sample with index i, j, the Z-score is

$$Z_{ij} = \frac{x - \mu}{\sigma} = \frac{N_{ij}^{\text{observed}} - N_{ij}^{\text{expected}}}{\sqrt{N_{ij}^{\text{expected}}}} \tag{3.3}$$

with $N_{ij}^{\text{expected}} = \frac{N_i N_j}{N_{\text{total}}}$. Here, the probabilities for the most and least abundant 6 nt submotifs right before and after the 4^{th} junction (ligation site) are shown. Sequence motifs on the x- and y-axis are shown, if the Z-score of the junction Z_{ij} is either significantly higher than 0 (dark teal color) or substantially lower than 0 (dark ocher color). Because of the segregation into A-type and T-type sequences and their different abundances and complementary motifs the ligation landscapes are plotted separately for both groups. As seen before in Figure 3.12 and Figure 3.13, the most abundant sequence motifs directly on the junction are `AT`-alternating. In the beginning of the left hand side motifs and on the end of the right hand side motifs poly-A and poly-T are common. The junction landscape shown in Figure 3.14 follow that trend, but highlight the vast sequence motif space around the ligation sites. Sequence motifs with alternating `AT`, but two similar bases directly on the ligation site like `AAAATA-ATATAA` (z-score $= -36.552$) are significantly underrepresented compared to the alternating motif like `AAAATA-TATAAA` (z-score $= 69.079$). The emergence of this frequent pattern is likely due to the inhomogeneity of sequence motifs in the initial 12mer pool, as shown in Figure 3.7. With the simple simulation that is based on the abundance of the sequence motifs only (see Figure 3.17) a similar self enhancing sequence pattern appears for the templated region (this explained in more detail below).

In the bottom right of the A-type ligation site sequence landscape a region of teal colored Z-scores indicate abundant motifs with poly-A directly on the ligation site, such as `ATAAAA-AAATAT`. This motif is not expected and at the first glance distinctly different in comparison to the motifs analyzed above and in Figure 3.13. On a second glance, the motif seems common after all, but appears to be in the wrong order. The reason is an effect that is suspected to emerge when long single-stranded templates become abundant: Other than for short templates, there is no restriction for substrate strands on where to bind on a long template. A 12mer substrate might hybridize directly on a region that originally was a junction in the template strand. Those strands might also indicate the extension of original primer sequences which, in order to be the reverse complement of a "poly-A"-"AT-alternating"-"poly-A" (`?????AAAAATATATAAAAA????`) strand, needs to have the alternating `AT` part at its center and the poly-T regions on the start and end (`TTTTATATATTT`). Both cases are rare in comparison to the common junction region that are separated by poly-A and poly-T stretches.

Analyzing the abundance of such motif shifts reveals that the longer the strand, the greater the chance to find at least one ligation site with a poly-A or poly-T motif, as shown in Figure 3.15a. The frequency increases in a linear fashion, which is explained by Figure 3.15b: all junctions for all oligomer products have the same probability of about 15 % to include a motif shift on a ligation site. The location of the ligation site motif shift is not homogeneous though. Figure 3.16 shows, that junctions near the end or start of the strand have a higher probability to differ from the AT-pattern compared to the center junctions. Especially the first junction has a higher probability for ligation site motif shifts.

With Figure 3.8 showing, that a small bias in the initial sequences of the 12mer pool can lead to dramatic changes for longer strands that have experienced several ligation events, the emer-

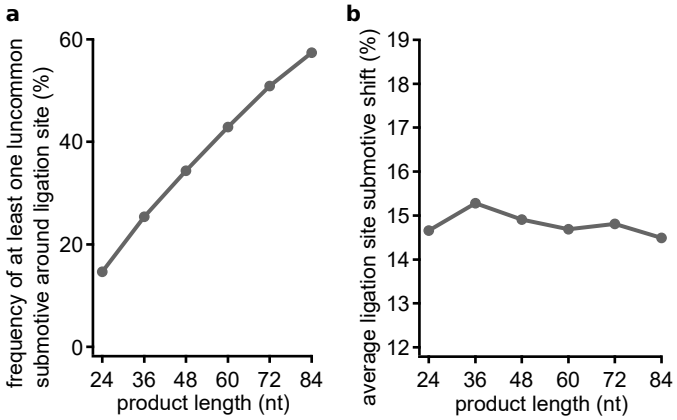

Figure 3.15 Ligation site motif shift:

a Frequency of finding motifs with poly-A (`AAAA` , `AAAAA` , `AAAAAA`) or poly-T (`TTTT` , `TTTTT` , `TTTTTT`) directly on the ligation site as a function of oligomer product length. The increase is about linear, with about 60 % of 84mers having at least one of such motif shifts. For 24mers, its only about 15 %.

b The frequency of a motif shift per junction is about equal for all lengths. In 15 % of junctions the common junction motif AT-alternating is replaced by poly-A or poly-T.

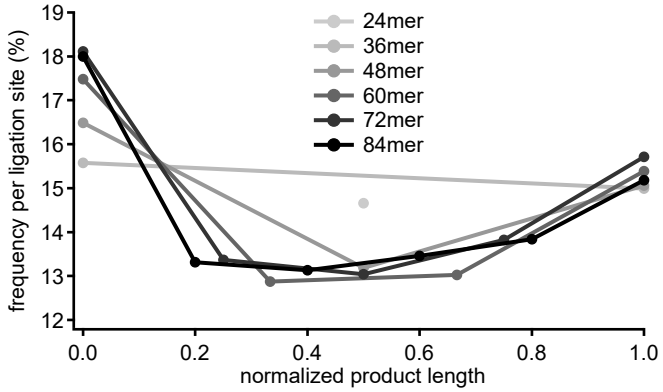

Figure 3.16 Ligation site motif shift per position in oligomers:

The probability to find a ligation site motif shift is higher for the end and especially for the first junction compared to the center junctions. The probabilities are similar for 36mers up until 84mers.

gence of abundant ligation site motifs might lay in the "monomer" pool as well. The analysis of the randomness of the 12mer initial pool in Section 3.3 and Figure 3.5 shows a homogeneous distribution of the base compositions with only a slight bias towards A-type strands. But the base-frequency per position in Figure 3.12 indicates, that there is a high probability of sequences of all lengths to end in `T` and for 12mers in `AT`. A more detailed and meaningful analysis for such possibly biased pools is the frequency of subsequence motifs. In Figure 3.17b the 2 nt motifs at every position (1 to 11) of all 85722 sequenced 12mers are analyzed. The last position shows an increased probability of the sequence `AT` compared to the other three possible motifs `AA`, `TA`, and `TT`. The most common motif at the second to last position is `AA`. The overall lack of `T` in all 12mer strands can be seen in the low frequency of `TT` for positions 4 to 11. In comparison, a completely random distribution has almost perfectly uniform 25 % frequency for all motifs at all positions, as shown on the right in Figure 3.17b.

To explore a potential influence of the ligase that might change the sequence motifs, especially at the ligation site, the now known abundances of position-dependent subsequence motifs are used as the input for a templated ligation simulation. Figure 3.17 illustrates the basic simulation steps and results. The extension mechanism is based on the experiment and the known extension modes (see Section 3.9). Therefore, the simulation is subdivided into two steps which are both done by Monte-Carlo-style particle-based non-kinetic simulations.

In the first step "initial" 24mer strands are calculated by simulating ligation of two 12mer substrates on a third 12mer acting as the template strand. The simulation randomly chooses a 12mer strand from the NGS data (1 out of 85722 sequenced strands, which makes it a weighted selection, as some of the possible 4096 possible sequences are sampled multiple times). For the ligation reaction to occur, only the abundance of X nt long motifs at the start, center ($2X$ nt), and end give rise to the simulation-ligation rate. The frequency of the X nt 3'-end is multiplied with the frequency of a X nt 5'-start. E.g. for 2 nt submotif length, a center subsequence motif `AATA` templates the sequence motif `TATT`. The frequency of the 5'-start motif `TT` is 26.10 %, the frequency of the 3'-end motif `TA` is 17.99 % which would give a ligation rate of about 4.7 %. The highest possible ligation rate occurs for the templating sequence `AAAT` with about 8.9 %, the lowest rate for the templating sequence `TTTA` with about 4.1 %. The length of motifs can be changed from $X = 1$ to $X =$ strand length$/2$. For longer motifs the simulation includes a parameter to increase the ligation rate by the same factor for all interactions, because here, the calculated ligation rate depends on the motif length. The initial 24mer pool is made from a specified amount of such complexes; 100 strands for the simulation shown in Figure 3.17d. Features already known from the 12mer pool are now prominent in the 24mers: the increased abundance of the 3'-end motif `AT` is clearly visible at position 11 and the lack of `TT` at almost all positions is also reflected in the "initial" 24mers.

In the second simulation step the ligation reaction rate is calculated in a similar way, but the template selection is different. From the simulation in Section 3.9 it is known, that the predominant extension mechanism for strands, in the presence of longer template strands, is the so-called "primer-extension" mode (see Section 3.9.2). Basically, the abundance of (long) template strands facilitates the growth by extending an already bound substrate strand by a second substrate strand. For 24mers this means that a 24mer template binds a 12mer which is then extended by another 12mer. The complex is then most likely a double blunt-ended complex. In the simulation here, the second step follows this behavior discovered in the kinetic simulation. Only existing 24mers can now act as a template. The templating section is the center section of the existing 24mers, which is the old ligation site of the initial 24mers from simulation step one, as shown in Figure 3.17e. All newly formed 24mers are added to the list of existing 24mers and can act as a template in the subsequent simulation cycles. In each cycle, all 24mers have the

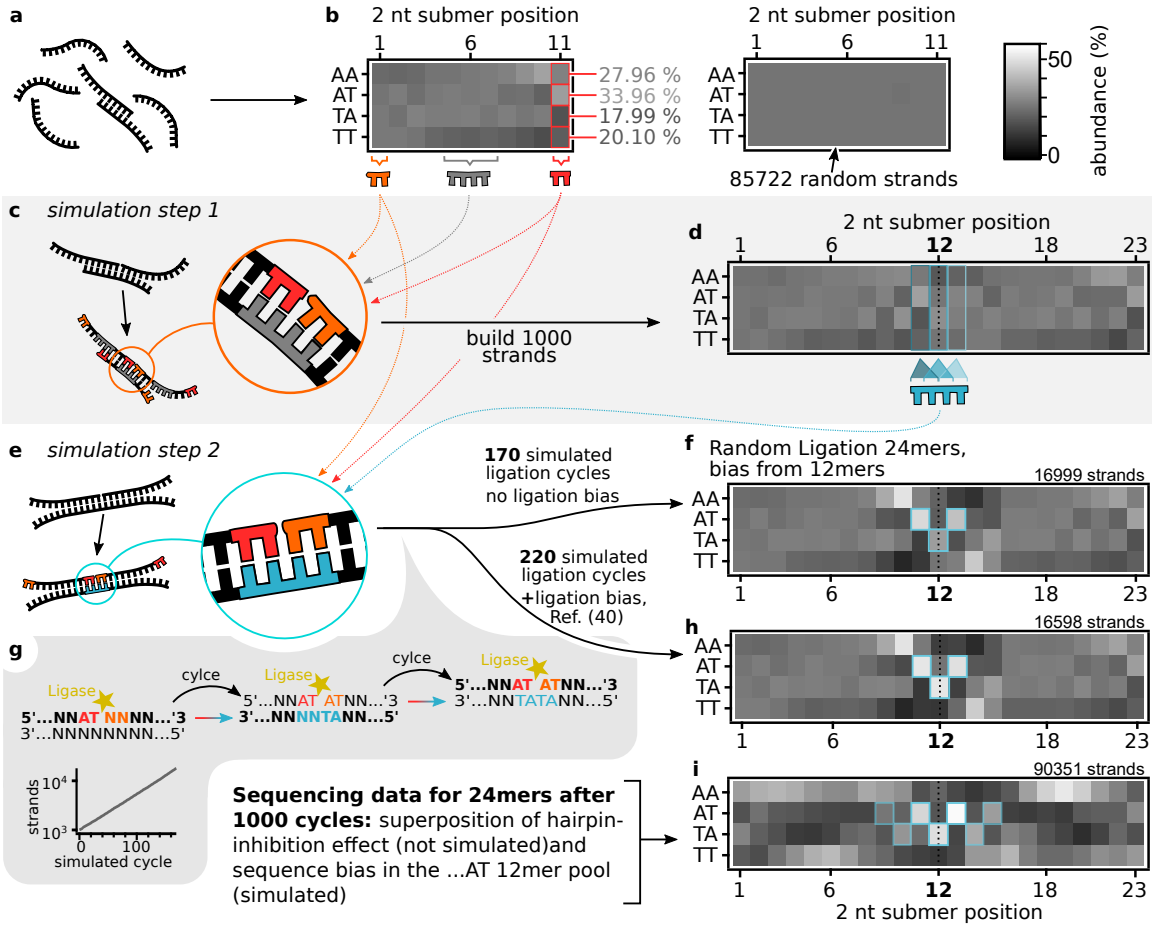

Figure 3.17 12mer sequence motif based templated ligation simulation:

a, b The 2 nt subsequence motifs in all 12mer strands analyzed by NGS show a position dependent bias for certain sequences. 12mers are more likely to end in `AT` than the other three possibilities.

c, Simulation step 1: In the experiment, the initial 24mers presumably emerged due to complexes made from three 12mer strands, as shown in Figure 3.17. Th simulation utilizes the abundances of center-, as well as 5'-start and 3'-end motifs of the initial 12mers in panel b. "Initial" 24mers are build by randomly hybridizing two 12mer strands onto another 12mer acting as the template strand. The effective ligation rate only given by the abundance of the submotifs, as highlighted in orange, red and gray.

d The initial 24mers basically show a similar pattern compared to the 12mers in panel b.

e, Simulation step 2: The predominant growth mode for oligomers in the presence of long templates is the "primer-extension" mode. In the simplest case, one 24mer template binds one 12mer strand and extends it with another 12mer strand.

f The second simulation step is run for multiple cycles, here for 170. The emerging 24mer strands show a clear `ATAT` pattern at the ligation site.

g Due to the bias towards `AT` at the 3'-end 24mers are more likely to have said motif at position 11. If this strand acts as a template, it will induce the reverse complement motif (which is also `AT`) at position 13 of a new 24mer. This strand is then more likely to have the ligation site motif `ATAT`, due to the abundance of the 3'-end bias of the 12mer pool.

h With an assumed bias potentially induced by the ligase, the overall ligation rate shrinks. The worst case scenario for the Taq DNA reduces the ligation rate of substrate `A-A` motifs more, than the other four possibilities. The outcome, anyhow, is very similar to an unbiased ligation rate.

i NGS data for 24mer strands after 1000 temperature cycles in the experiment show a clear `ATAT` motif preference at the ligation site. The additional patterns can readily be explained by the inhibition of hairpin structures and a less position-specific hybridization.

chance to template a new 24mer strand: the rate of ligation is again the frequency off finding the reverse complement of X nt motifs at the 5'-start and 3'-end of the 12mer strands in the original pool. With each cycle, exponentially more sequences are build that can act as new templates, as shown in Figure 3.17g, right side. The resulting 24mers are analyzed for the abundance of all possible 2 nt sequence motifs for all positions, as before (see Figure 3.17f). The center section shows a distinct pattern of `...ATAT...` from position 11 to 13. The dominance of the `AT` motif is due to the initial bias in 12mer strands in the original pool. Figure 3.17g shows a sketch of the simplified mechanism. A strand with said motif `AT` at the 3'-end is ligated to another strand. The resulting 24mer can act as a template in subsequent ligation steps with its motif `ATNN` and will template the motif `NNTA`. Importantly, the motif `AT` is its own reverse complement: `AT` hybridizes on `AT`. In a new ligated strand the second substrate therefore needs the motif `AT` at the 5'-start. At the same time, the 3'-end motif of the first substrate strand is still more likely to be the motif `AT` as well. The likelihood of an emerging 24mer with the ligation site motif `ATAT` is thus high.

Compared to the NGS data for 24mers after 1000 temperature cycles in Figure 3.17h the pattern at the ligation site is very similar. But the NGS data shows more features not explained by this simple simulation:

- **A-type / T-type;** as shown in Figure 3.2 the 24mer strands are already segregated into A-type and T-type groups that prevent the formation of hairpin structures. As a result, the sequence motifs `AA` and `TT` are more abundant in the NGS data. This effect is not included in the simple simulation here.
- **Ligation site shift;** in the experiment, the hybridization location is not as defined as in the simulation. The simulation results in Section 3.9.2 suggest a dominant primer extension mode, but the exact hybridization position can vary. The experiment allows the 12mer substrates to hybridize at positions ± 1, ± 2, *etc.* of the original ligation site. The sketch in Figure 3.17g discussed above would also induce the abundant self-complementary motif `AT` at the substrate positions 9 and 15 when shifted, but with a reduced frequency, as the ligation site is less common (see Figure 3.15). In the NGS data of 24mer strands in Figure 3.17h this additional shift can clearly be seen in the extended `AT` alternating pattern.

As discussed in more detail in Section 4.6, a sequence-dependent ligation rate could be responsible for the emergence and amplification of specific sequence motifs, especially at the ligation site.

3.5 Position-dependent subsequence correlations

Despite the slightly larger abundance of ligation site motif shifts in the outermost junctions of oligomer products strands, the AT-fraction distribution in Figure 3.4 and the patterns in Figure 3.12 lead to the assumption, that all strands must be very similar to one another. With the sample Pearson Correlation Coefficient (sPCC) the similarity between all different subsequences in an oligomer can be compared to an oligomer of another length in order to detect possible correlations on the length-scale of 12mers. To calculate the sPCC the abundance of all subsequences in a certain position within a x-mer is plotted *versus* the abundance of all subsequences in certain position within a y-mer. Figure 3.18 shows how such plots look like for low (Figure 3.18a) and for high (Figure 3.18b) correlations. Each data point marks one sequence with at least one occurrence in at least one oligomer. For visualization purposes, the `A` and `T` abundance of each sequence are marked from red to blue, as indicated.

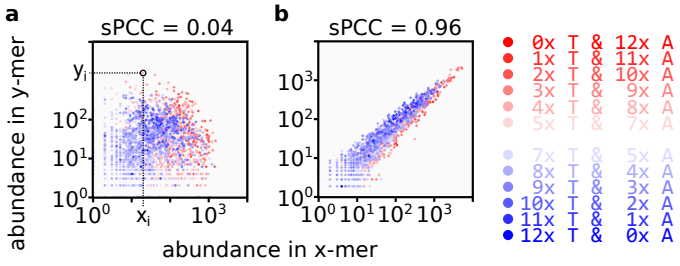

Figure 3.18 sample Pearson Correlation Coefficient for correlating and non-correlating subsequence positions:
Plotting the abundance of all subsequences found in a x-mer at a certain position *versus* the abundance of all subsequences in a certain position of a y-mer with color coded A-type (red) and T-type strands (blue).
a Without a correlation, the distribution has no particular shape. A subsequence common in the x-mers is unlikely to also be abundant in the y-mer. The sPCC value for the distribution plotted here is 0.04, indicating a very low correlation.
b With a sPCC value of 0.96 this distribution indicates a strong correlation between the x-mer and y-mer subsequence positions: Subsequences common in the x-mer are also common in the y-mer.

The abundance of subsequences in low correlation distributions are different in x-mers and y-mers: common subsequences in the y-mer are unlikely to also be common in the x-mer. Therefore, the distribution has no particular shape. If the x-mer subsequences and the y-mer subsequences are similar in abundance, the resulting distribution is a straight line through the origin. Subsequences common in the x-mer are then also common in the y-mer. Subsequences that are rare in the x-mer are also likely rare in the y-mer. To obtain a quantitative measure of the correlation, the straightness of the distribution is analyzed and rated between 1, perfect positive correlation, 0, no correlation, and -1 perfect negative correlation with

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3.4}$$

Here, n is the sample size, the individual sequences are called i with their abundances x_i, y_i and the samples mean abundances $\overline{x}$, $\overline{y}$. A-type and T-type groups form separate distributions that are slightly shifted, because A-type strands are more common in longer oligomers.

In Figure 3.19 all subsequences from the seven possible subsequence positions in 84mers are compared all subsequences from the six possible subsequence positions in 72mers. Figure 3.19a schematically indicates the position of the sPCC value of the 3$^{\text{rd}}$ position in the 72mer and the 5$^{\text{th}}$ position in the 84mer. High sPCC scores indicating a high correlation are dark blue in Figure 3.19b, while low sPCC scores indicating a low correlation are pink and white. The first (or start-) subsequence in 72mers and 84mers are highly correlated with a sPCC score of 0.99. Similarly, the last subsequences are also highly correlated with a sPCC score of 0.99. Subsequences between the second and the second-to-last position correlate with a sPCC score of at least 0.87 and up to 1.00. Interestingly, the sPCC scores correlating subsequences between those three groups are low: Comparing the first subsequences in 72mers to the last subsequences in 84mers gives a sPCC score of only 0.09 indicating a very low correlation. Also, the correlation score of the first and last subsequences with center subsequences are typically only about 0.4.

Despite the DNA strands only consisting of bases `A` and `T` and a reduced sequence space due to the A-type and T-type groups, there are three distinctly different subsequence motif groups

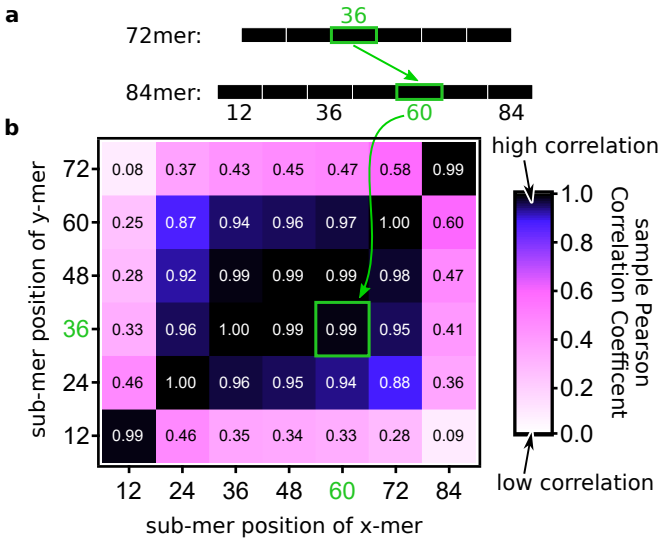

Figure 3.19 sPCC score matrix comparing all subsequence positions in 72mers and 84mers:
a Schematic sketch indicating how to read the sPCC score matrix in panel b.
b Dark regions mark high correlations with high sPCC scores, light regions mark low correlations with low sPCC scores. Start-subsequences, end-subsequences and center-subsequences form self-similar groups in which sequences have similar abundances. The light color on the edges indicates a low correlation between the three different groups, segmenting an oligomer into three distinct parts.

emerging in oligomer products. Start-subsequences, end-subsequences and center-subsequences form inherently self-similar groups with similar abundances of subsequences. But comparing the three groups to one another, shows only very low to no correlation, especially for the comparison of the start-subsequences to the end-subsequences. In this context, a hypothesis comes to mind: a strands is elongated by this random sequence templated ligation, as long as the attached 12mer or other attached oligomer products are similar to the center subsequences. But as soon as a 12mer similar to the start- or end-subsequence groups is attached, the growth in this direction of the oligomer product stops. The functionality of start and stop codons is known from protein synthesis: three consecutive bases in an mRNA strand code for a specific amino acid [55,100]. Start and stop codons mark the start and the end of the synthesized protein [55,103]. In this simple model of DNA elongation by templated ligation, the start- and end-subsequences are more of a stalling factor for the growth dynamics than a functional mechanism in evolved cell metabolisms. But it shows, that even simple dynamics are capable of producing sophisticated function in their products.

Due to the asymmetry in oligomer products with an abundance of A-type strands, those strands govern the comparison of subsequences in different oligomer lengths shown in Figure 3.19. Therefore the comparison of 36mers are chose to analyze the difference of A-type and T-type strands, as 36mers still have a comparably similar concentration in both groups (see Figure 3.4). As expected, there is no positive and only a slight negative correlation of subsequences in all positions in Figure 3.20 a. T-type strands only contain a very small amount of A-type subsequences and *vice versa*, as shown in Figure 3.9. Therefore, subsequences common in one base-type must be uncommon or even absent in the respective other one. But taking all 36mer T-type strands, calculating their reverse complement strands and comparing this set of strands to the sequenced A-types reveals a sPCC matrix (Figure 3.20 b) similar to the comparison of different lengths. This

suggest, that A-type and T-type strands are mostly reverse complement groups. Sequences common in one group will probably be found as the reverse complement in the other group.

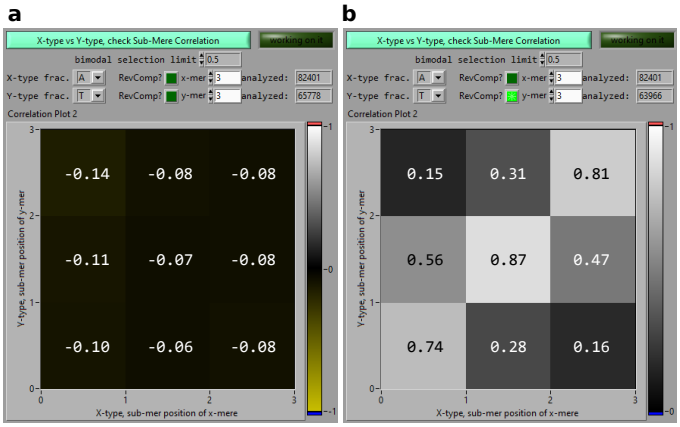

Figure 3.20 sPCC score matrix comparing A-type and T-type 36mers:
Screenshots from the LabView sequence analysis tool.
a Comparing A-type 36mers to T-type 36-mers reveals no correlation at all. This is expected, as A-type subsequences are rarely found in T-type strands and *vice versa*.
b In contrast, building the reverse complement for all T-type strands and comparing those to the sequenced A-type strands reveals a similarly shaped sPCC matrix as known for the comparison of different lengths. This reveals, that A-type and T-type strands are mostly reverse complement sets of strands.

3.6 GC-random templated ligation reaction

The choice of building the model oligomer from DNA with only bases `A` and `T` stems mainly from the lower melting temperature of the Watson-Crick basepair `A` - `T` (see Section 5.16). Strands made from `C` and `G` have higher T_{melt} which might lead to difficulties with the ligase activity and persistence. Additionally, poly-G motifs are known as "sticky" bases that also form double-stranded sections with other poly-G strands [125]. Anyhow, the emergence of oligomer products from short GC-only ssDNA by templated ligation should be expected, as the mechanism of templated ligation is robust and the enzyme should not have a preference for AT-strands. As described in Section 5.3.2, 12mer AT-only is the shortest "monomer" length that can be ligated in the experimental settings here. But because of the increased binding energy of GC-only strands, 10mer GC-only "monomers" did get ligated and produced oligomer products comparable to the 12mer GC-only samples. To ensure quality sequencing results in *Illumina* -NGS-columns samples should not have a bias towards some bases, or base-fractions. Typically, different barcoded samples are combined in a single flowcell to diversify the reads and reduce evaluation artifacts. In the sequencing analysis for the GC-only random sequence samples the relative sample concentration of the GC-only samples was about three times the concentration of AT-only samples. Still, the read count for the GC-only strands with a similar read quality as for the AT-only sample is significantly lower, as shown in Figure 3.21. For AT-only samples the read count even for 12mer monomers, a length at which the efficiency of the NGS kit is low, produce over 80000 good reads and for long strands like 84mers over 200000 good reads. With over a factor of over 50 less reads in the best case scenario (and over 10000 in the worst case, for long strands) the accuracy of statistical analysis and quality of results is greatly reduced in the analysis of GC-only samples.

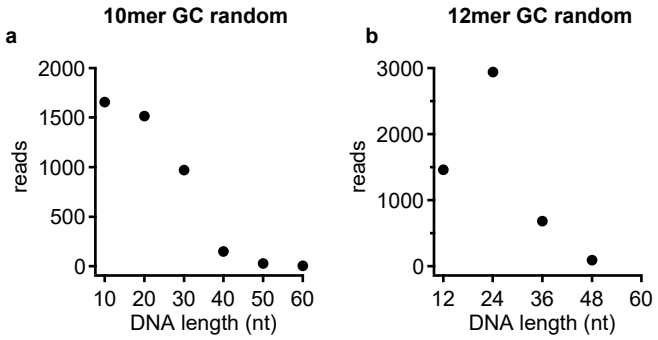

Figure 3.21 Sequencing read counts for GC-random samples:
a Although 10mer GC-only strands do produce oligomer products under random templated ligation reactions, the read count of high quality reads is at least a factor of 50 lower than in AT-only samples.
b 12mer GC-only samples do also show a significantly lower read count compared to AT-only samples.

As in the experiments with AT-only DNA, the randomness of the initial pool for the GC-only strands influences the elongation sequence dynamics. Both, 10mer and 12mer GC-only are shifted towards G-type strands, as shown in Figure 3.22a, b. The 10mer sample is shifted more compared to the AT-only sample, although strands with 50:50 G:C are still the most abundant. The GC-fraction of the 12mer sample is significantly shifted towards G-type stands with strands made from eight `G` and four `C` being the most abundant. Figure 3.22c, d highlight that asymmetry. While all C-type strands are underrepresented, strands with a very low G:C fraction are especially rare. G-type strands are overrepresented, and in contrast to the AT-only sample, high G:C-ratio strands are strongly overrepresented by up to 100 % compared to a random pool (binomial distribution). For the 12mer GC-only sample, this asymmetry is even more pronounced and even strands with G:C-ratios of 6:6 and 5:7 are less abundant than expected for a random pool. The most obvious difference of both GC-only samples compared to the AT-only sample in Figure 3.5a is that while low G:C-ratio sequences are rare, high G:C fraction strands are very common. The broad region of strands with about 50 % of either base are also underrepresented instead of overrepresented. This might be an artifact from DNA synthesis, favoring base `G` over `C` during the solid state build method.

The G:C-ratio for oligomer products is less reliable than the AT-only sample due to the significantly lower read counts for GC-only samples. Figure 3.23 shows the G:C-ratio of emerging 20mer and 24mer products. The 10mer "monomer" sample shows a clear bimodal distribution for 20mers. The peaks are centered at around 0.35 and 0.65 G:C fraction, close to the results from AT-only. 30mers and 40mers, despite their low read counts, also fall in the same bimodal distribution. In contrast to the A-type sequences being more abundant (which is probably due to the asymmetry in the initial AT-pool), the systems seems to favor the C-type oligomers despite their lack in the 10mer GC-only random sequence pool. Here, the reason might lay in the before mentioned nature of poly-G strands. Complexes with regions of poly-G tend to form duplexes with similar strands [125] and DNA triplex-structures [24, 44, 46]. This binding mechanism is absent in AT-only samples and might provide a multitude of additional hybridization regions for poly-G strands. As a result C-type sequences might elongate better, because G-type strands tend to stick together and therefore can't take part in templated ligation reactions. The 12mer sample also shows hints of a bimodal distribution for its "dimers" but the peak of G-type strands is still greater than the C-type peak. With the analysis from Figure 3.23b and the strong asymmetry

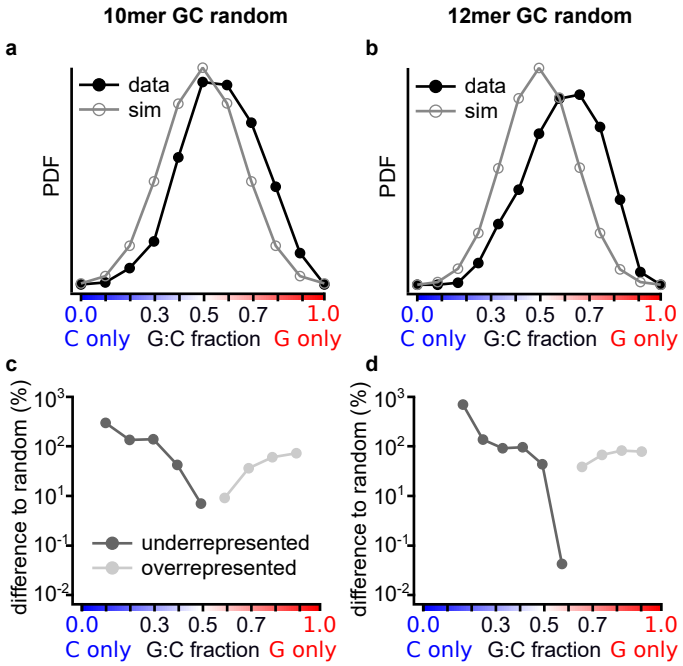

Figure 3.22 Initial 10mer and 12mer GC-only random strands are shifted towards G-type strands:
a 10mer GC-only random strands are shifted towards G-type sequences.
b 12mer GC-only strands have a significant shift towards G-type strands. Strands with about 50:50 G:C are also underrepresented, which is not seen for AT-only and 10mer GC-only samples.
c C-type strands are eight to 300 % (for 9, 10 `C`) underrepresented, while G-type strands are overrepresented by up to 100 %.
d Long C-type strands are up to 1000 % underrepresented, while strands with at least eight bases `G` are overrepresented by 200 %.

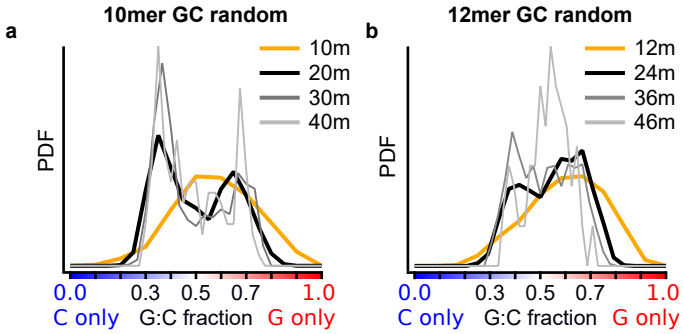

Figure 3.23 G:C-ratio evolution:
Oligomers longer than "dimers" are indicated by thin lines. The low read count for GC-only samples prevents a detailed analysis.
a Oligomers in the 10mer pool show the emergence of a bimodal G:C-ratiowith centers at about 0.65 and 0.35, comparable to the AT-samples. The G-type peaks a smaller than the C-type peaks, despite an overall shift towards G-type strands in the initial pool.
b 24mers indicate the formation of a bimodal G:C-distribution. Again, low read counts prevent a more detailed analysis.

seen in Figure 3.5b one might suspect a superposition of both effects: The system might favor

C-type strands because of their lack of triplex or `G` - `G` duplex formation but the 12mer GC-only sample has so much more G-type strands, that the first mechanism might only be visible for longer strands or higher cycle counts.

Overall, the system seems to select (in this case select means favorably produce) strands that can act as templates for the random templated ligation reactions, that at the same time do not experience further hybridization to themselves (hairpins) or similar strands (poly-G).

3.7 Reduced complexity pool experiments

The time scale of the experiment is governed by the ligation temperature time step. This step needs to be long enough, so complexes can form and also be ligated. Experiments discussed in Section 3.8.3 lead to the assumption, that the limiting factor is indeed the formation of the complex. The subsequent ligation is fast in comparison (a more detailed analysis is done in the kinetic simulation model in Section 3.9). A simple possibility to analyze sequences in a later stage of the original experiment is by either letting the experiment run for an extended amount of time (which is difficult, because the ligase will degrade and the overall time scale will be very large) or by reducing the sequence space of the random sequence pool. A smaller sequence space with a similar total strand concentration increases the relative hybridization partner concentration of single strands. This in turn increases the complex formation rate, also-called on-rate k_{on}.

In the following section several reduced complexity pools are analyzed: a 12 nt length "6mer" ($2^6 = 64$) sample, a sample made from eight sequences selected from the resulting pool in Section 3.2 and a 12mer self-complementary strand capable of forming three-strand-complexes with just one sequence.

3.7.1 x64 pool, "double-bases"-monomers

One fairly obvious idea to reduce the sequence space of a pool is to make the strands shorter. With a lower exponent in basecount$^{\mathrm{length}}$ the sequence space will shrink significantly. From a binary alphabet 12mer to a 6mer the sequence space decreases from 4096 by a factor of 64 to 64 only. But as discussed above the experimental conditions here do not allow for 6mer strands, as the melting temperature of three-strand complexes are too low and 3 nt overhangs are not ligated by the Taq DNA ligase. Instead, 12mers are used which only act as a 6mer by having "double-bases": a 6mer with sequence `ATATTT` is mimicked by a 12mer `AATTAATTTTTT`. This allows for the formation of complexes with hybridization of only three "double-bases", with the same effective length and melting temperature as in the 12mer AT-random case.

For the stock sample, all 64 possible AT-random 6 nt long sequences are ordered as their double-base version 12mers, concentration-analyzed with the NanoDrop (see Section 5.4) and mixed. As the sequences space is significantly lower than the AT-random space, the total stock concentration is adjusted to 1 µM from 10 µM for one sample. For both concentrations a sweep in ligation temperature from 25 °C to 39 °C is shown in Figure 3.24. The first major difference compared to the AT-random sample is the consistent amount of long oligomers emerging. Due to the lower sequence space but same experimental time scales for hybridization and ligation (for details compare Section 3.8), the amount of oligomer products is higher, and therefore the autocatalytic production for even longer strands stronger. For the 1 µM concentration there is a clear difference visible for all lanes. Interestingly, for higher T_{lig} the amount of short oligomers is remarkably low, compared to low T_{lig}. This is described in more detail in Section 3.8, where the dynamics of the oligomer product distribution is analyzed in context of the experimental

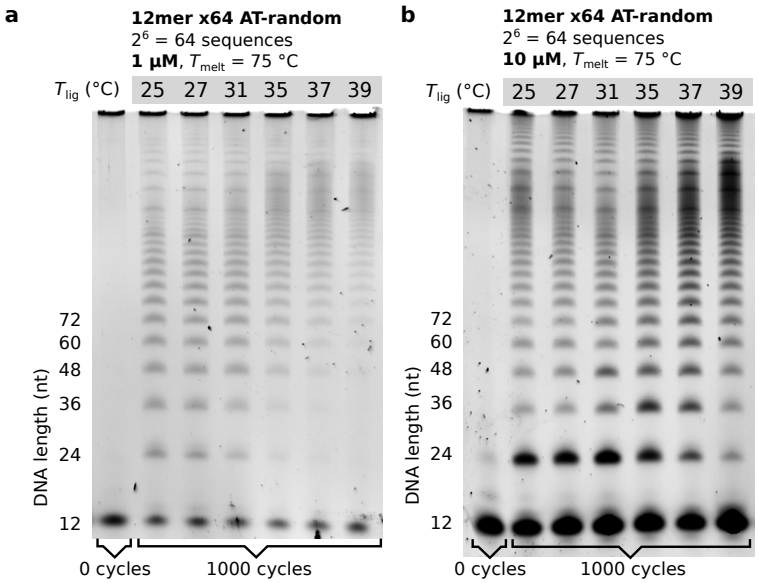

Figure 3.24 PAGE image of x64 "double-base" random sequence 12mer with sequence space 64:
a For 1 µM concentration in the experiment, the oligomer product distribution follows a trend comparable to the AT-random experiment and produces oligomers with length of a multiple of 12. As seen in comparison to the reference lane, the temperature cycled samples have a significantly lower 12mer concentration, indicating a high oligomer production rate for the same cycle count and time at a lower total strand concentration than the 12mer AT-random sample.
b For 10 µM 12mer pool concentration the product concentration is significantly higher and there is less difference between the lanes (and therefore the experimental conditions) visible by eye. Due to close proximity and intensity of the bands the analysis with the concentration quantification tool is not possible.

parameter space. The sample with a 10 µM pool concentration (Figure 3.24 b) results in a high concentration of long oligomers as well. Due to their abundance the CQ-tool can't distinguish between long oligomers which makes the concentration quantification impossible.

The sequencing data for the 10 µM concentration sample suggest an about binomial A:T-ratio, as expected for a random distribution. In comparison with a simulated random distribution, the lack of 50:50 A:T is visible. Additionally, T-type strands and 5:1 A:T-ratio are overrepresented, while 4:2 A:T-ratio are underrepresented. Because the strands are made from "double-bases", uneven A:T-ratios must be read errors from sequencing or synthesis artifacts in the DNA (for more details on error estimation see Section 5.16). Anyhow, the amount of wrong reads is low, as seen by the dark gray dotted line.

One essential feature that develops in 12mer AT-random samples, is the emergence of A-type and T-type strands. This property inhibits strands from folding on themselves which in turn prevents a strand from acting both as template and as substrate in subsequent ligation reactions. As shown in Figure 3.26 the x64-sample also forms a bimodal A:T-ratio distribution from the initial binomial distribution. In panel b all wrong reads are filtered: The A:T-ratio evolution as a function of oligomer length is very similar to the AT-random sample. This reduced sequence space sample also favors A-type strands, which become very abundant in comparison to T-type strands in long oligomers. Interestingly, short oligomer products favor T-type strands (24mer) or are about equally abundant (36mer). The abundance of 5:1-ratio A-type and 6x `AA` might have a stronger influence on long oligomer products than on shorter ones.

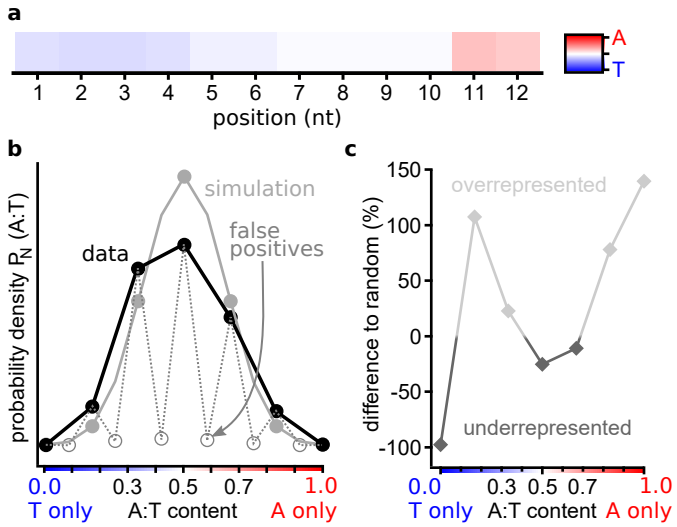

Figure 3.25 x64 "double bases" sample 12mer initial pool sequence motifs:

a Base probability plot for 12mers. As designed, the probability of the double bases is almost similar. The bias towards `T` for the last base in the 12mer AT-random sample is a bias towards `AA` here.

b A:T-ratio distribution for 12mers. While 3:3 and 4:2 A:T-ratio sequences are underrepresented, 1:5 A:T-ratio, 2:4 A:T-ratio and 5:1 A:T-ratio sequences are overrepresented compared to a random distribution. The false reads and incorrectly synthesized sequences (which are indistinguishable errors) are marked in light gray hollow circles.

c The difference to random from panel b in percent. Although T-type strands with about 30 % `A` are overrepresented, A-type strands with five or six "double-bases" `A` are also up to 150 % more common than expected for a random distribution.

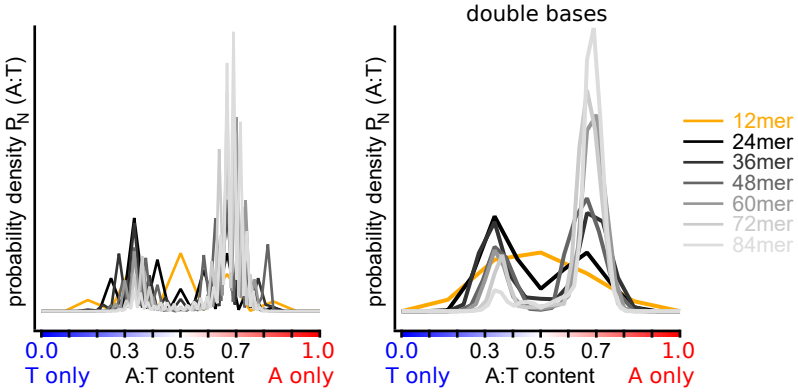

Figure 3.26 A:T-ratio evolution in x64-sample:

a The A:T-ratio produces A-type and T-type strands from the random distribution of 12mer sequences, similar to the AT-random sample.

b Same graph as in panel a but without the data for false reads. Here, the graph looks very similar to Figure 3.4. Although, there is an initial bias towards T-type strands, the oligomer products favor A-type strands.

A so-called de Bruijn graph visualizes directed networks as a 2D mapped landscape. For the visualization of sequencing data the sequences are represented as the nodes and the their abun-

dance defines the size of the node. The thickness of the edges connecting different nodes represent how often two sequences are found after one another in an oligomer strand. With only 64 different sequences, the network visualizing the succession of subsequences in oligomer products can be plotted completely. And although the 84mer strands include all possible 64 sequences and 1701 of all possible 4096 (41.5 %) edges, there are only a few very common ones (edges cutoff <1000 sequenced connections between two subsequences). Figure 3.26 shows the two emerging network families in the 84mer oligomer products. Due to the bias towards A-type strands, the T-type strand network only has three strands above the plotting threshold of 1000 occurrences. In the A-type network family, the edges between three common sequences dominate. Remarkably, the T-type network consists of the exact three reverse complement sequences that are best connected in the A-type network. From the three subsequences in A-type strands, two start with

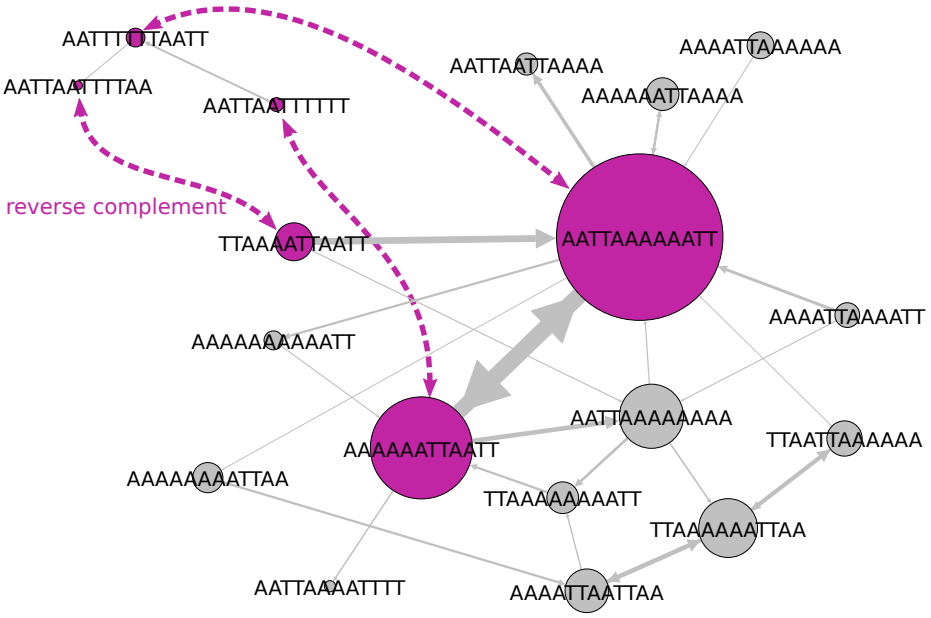

Figure 3.27 de Bruijn subsequence-sequence network for x64-sample 84mer oligomers: The most common sequences and the most common connections are drawn as a de Bruijn graph. The network consists of two families of strands, which are mostly either A-type or T-type. All three sequences in the T-type network are the reverse complement of strands in the A-type network - similar to Section 3.7.2, below.

`AA` and end with `TT`, include one continuous section of six `A`, while all three include the pattern `AATTAA`. The two most common subsequences in A-type 84mer strands are also commonly found in alternating succession (e.g. sequence $S_1 \rightarrow S_2 \rightarrow S_1$). Despite finding all 64 possible sequences in the oligomer products, they are predominantly made from only a few sequences. Those sequences have similar base-patterns and A:T-ratios and form A-type and T-type groups, that inhibit the formation of hairpin structures and can template the respective other group.

3.7.2 x8 pools

Instead of shrinking the sequence space by having 2 nt-long "bases" as in Section 3.7.1, a smaller sequence space pool can be made by simply selecting a subset set of sequences. The most trivial reduced complexity pool is one made from only a few possible sequences that are randomly selected from the 4096 possible AT-only 12mers, as shown in Table 3.2. It is unlikely that a pool

produced by such a random selection will perform particularly well in a templated ligation reaction. But the selection for such reduced complexity pools does not need to be random, it can be aided by the experiments described above.

de Bruijn graph - sequence networks in AT-random 12mer oligomers

As the experiment with AT-only 12mer "monomers" produces a lot of long sequences from the short pool, it stands to reason that the resulting strands might have a "sequence advantage" when it comes to the underlying elongation and selection mechanism. A long strand needs a sequence, that does not fold on itself, but can hybridize to other common strands. In terms of the already shown de Bruijn network graph, such a strand would be especially "fit" the better its subsequences are connected in the space of common subsequences.

Figure 3.28 shows the network for the most common connected subsequences in A-type and T-type oligomer products (A-type: all internal junctions, expect first an last, for all strands longer than 48 nt, all strands are sequenced at least 20 times and have a Z-score>30; T-type: all internal junctions, expect first an last, for all strands longer than 48 nt, all strands are sequenced at least 10 times and have a Z-score>15). The distinction of A-type and T-type is necessary due to the complementary sequences in both groups.

There are several seemingly independent sequence-families in each network, that aren't connected by edges. The A-type network has three large families with distinctly different structures:

- **top** An interconnected subnetwork with several different end-nodes. The highlighted sequence `TAATAAAAAAAT` is one of the eight most common sequences in all oligomers products.
- **center** A small directed subnetwork. Five different sequences have a common sequence `ATAATAAAAAAT` as their ends-node.
- **bottom** A loosely connected network including five of the eight most common sequences.

Interestingly, the T-type network features the same structure and shows a very interconnected, a directional and a loosely connected subnetwork. In the top eight most common sequence for each network, four are found as the reverse complement in the respective other network. The comparison of 36mer A-Type and T-type strands (Figure 3.20) already suggested, that the two groups are made from reverse complement sequences. The de Bruijn networks now clearly show, that not only the sequences, but also the strand motifs and structure in both A-type and T-type groups are closely related. This result is clearly in line with the extension by random templated ligation, the mechanism that produces the oligomer products.

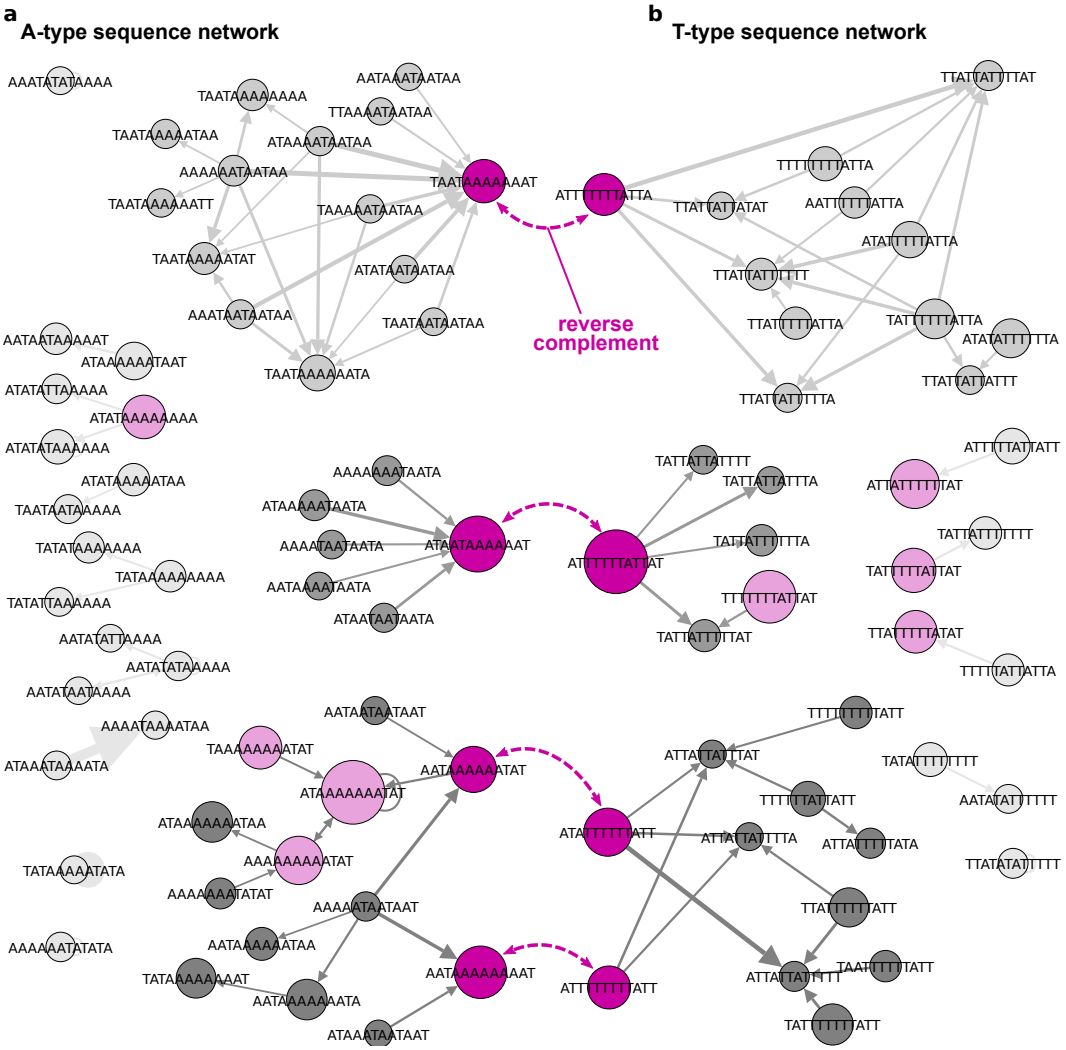

Figure 3.28 de Bruijn graphs of internal subsequence-sequence networks in AT-only oligomer products:

a A-type strands form several subsequence-families, that aren't interconnected. The network-families all have different forms, e.g. the center family is very directed, whilst the top one shows an intricate interconnectivity.

b T-type networks look remarkably similar to A-type subsequence families. A similarity in abundant sequences is not surprising in a complementary sequence group, but the similarity in network-forms suggest a true complementary system.

x8 sequence pools - selected from network, random subset, replicator

The visualization with a de Bruijn graph enables the selection of the best connected and most common 12mer subsequences from the 84mer oligomer products. From A-type and T-type networks the four most common sequences, that are found as the reverse complement in the respective other group, are selected as the "x8 `Network`" pool sample. The sequences are shown in Table 3.1. For comparison of the sequence-emergence- and ligation-ability of this pool two other

Table 3.1 `Network` selection x8 pool, selected by the templated ligation experiment

ID	sequence	modification	sample concentration (μM)
N_1	ATAATAAAAAAT	5' phosphate	1.25
N_2	AATAAAAAAAAT	5' phosphate	1.25
N_3	AATAAAAAATAT	5' phosphate	1.25
N_4	TAATAAAAAAAT	5' phosphate	1.25
N_5	ATTTTTTATTAT	5' phosphate	1.25
N_6	ATATTTTTTATT	5' phosphate	1.25
N_7	ATTTTTTTTATT	5' phosphate	1.25
N_8	ATTTTTTTATTA	5' phosphate	1.25

x8-pools are build for comparison. The second sample called "`Random`" consists of eight randomly chosen 12mers from the 4096 possible binary AT-only 12mers (see Table 3.2). The third sample is

Table 3.2 `Random` subset x8 pool, randomly selected sequences from the 4096 possible 12mer AT-only sequences

ID	sequence	modification	sample concentration (μM)
R_1	AAAATAAAATAT	5' phosphate	1.25
R_2	ATAATTAAATAA	5' phosphate	1.25
R_3	TAAAAATTATTT	5' phosphate	1.25
R_4	TTAAATTTTATA	5' phosphate	1.25
R_5	TATTTAATTTTT	5' phosphate	1.25
R_6	TAAAAATTAATA	5' phosphate	1.25
R_7	AAAATAATTTAT	5' phosphate	1.25
R_8	TTATATAAAATA	5' phosphate	1.25

designed to form three-strand complexes with two overhangs on the substrate side. All strands are themselves made from three sections:

- alternating-homogeneous-alternating
- homogeneous-alternating-homogeneous

This structure resembles the sequence pattern found in all oligomer products: alternating-homogeneous-alternating is similar to Figure 3.13 and homogeneous-alternating-homogeneous acts as the template for the first case, but can also act as the substrate with the first case acting as the template. From the total of eight strands in the `Replicator`-sample four are A-type strands and four are T-type strands, as shown in Table 3.3.

The main difference of those x8-samples compared to the AT-only full random 12mer samples is, as described, the sequence space. In the experiments the rate of oligomer elongation is significantly higher and thus there are clear bands already after 200 temperature cycles. Figure 3.29a

Table 3.3 `Replicator` x8 pool, designed to hybridize and elongate

ID	sequence	modification	sample concentration (μM)
E_1	ATATTTTTTATA	5' phosphate	1.25
E_2	TATAAAAAATAT	5' phosphate	1.25
E_3	AAATATATAAAA	5' phosphate	1.25
E_4	TTTTATATATTT	5' phosphate	1.25
E_5	AAAATATATAAA	5' phosphate	1.25
E_6	TTTATATATTTT	5' phosphate	1.25
E_7	TATTTTTTTTAT	5' phosphate	1.25
E_8	ATAAAAAAAATA	5' phosphate	1.25

shows the PAGE images of the three x8-samples, Figure 3.29b the concentration estimation for all lanes. The `Network` sample shows an almost perfect exponential decrease in oligomer concentration as a function of length, starting from 36mers. For short oligomer product strands, the concentration is the highest for all three samples. The `Random` sample also shows the emergence of oligomer products, but with a low concentration. The `Replicator` sample has an especially high concentrations for long products strands. The `Replicator` is designed to build complexes

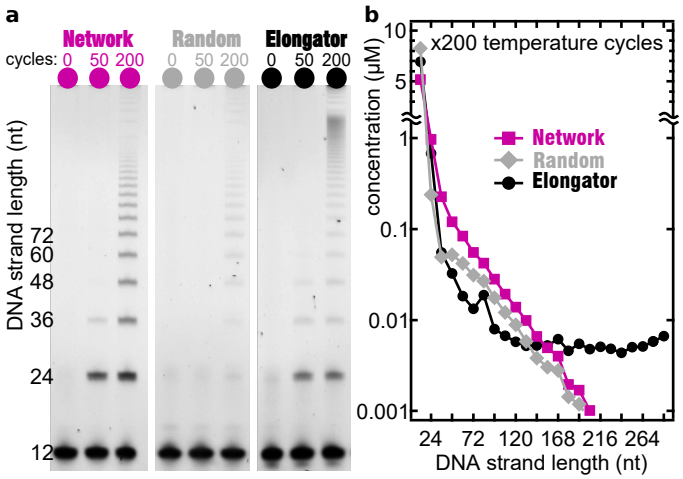

Figure 3.29 PAGE analysis for x8 pool samples, `Network`, `Random` subset, and `Replicator`:
a All samples show the emergence of oligomer products. The bands for the `Random` subset sample are faint, which is not surprising, as the elongation by templated ligation can only be due to chance because of the random choice of strands in this sample. The `Replicator` produces a lot of very long strands. The pool made from sequences selected by the network dynamics of the random templated ligating pool (`Network`) produce the most short oligomer products.
b Concentration quantification of the PAGE image in panel a. The `Network`-sample produces the highest output for short an medium-length oligomers suggesting a worse autocatalytic extension ability for the other two samples, `Random`, and `Replicator`. The `Replicator`-sample, on the other hand, is the "fittest" for the production of very long oligomers, due to its designed overhang that prevents unextendable blunt-end double strands.

with overhangs, onto which reverse complement sequence monomers and other complexes can hybridize. In comparison to the full random pool, this can be interpreted as a reduction in se-

quence space and a reduction in "alignment space". The sequences in **Replicator**-sample only allow for double-stranded oligomers that are either blunt-ended or ligatable. Thus, it is more likely that strands hybridize in as a ligatable complex than for the AT-random 12mer sample. Consequently, the ligation rate is higher for double strands is high. In contrast to the other two x8-samples Network and Random, the concentration distribution is not exponentially decreasing. Due to the presumably (see below for more details) similar and repeated internal structure for the **Replicator**, the emergence of long double strands with overhangs is expected. The longer strands grow, the higher their melting temperature and the less likely they are to dissociate by temperature cycling to 75 °C. Similar to ref. [122] Fig. 2, the ligation rate for complexes with a low melting temperature can out-produce shorter complexes, even in a competition for the same substrate. The **Replicator** might facilitate a growth mode that is not accessible in the other samples, that in turn changes the overall appearance of the concentration over product length distribution (see Section 3.9).

Any random selection of only eight of the 4096 possible AT-12mers has a low probability to produce oligomer products. The Random sample used here should not form complexes suited for templated ligation by the Taq DNA ligase. This prediction by NUPACK (see Section 5.3) apparently doesn't hold in the experimental setting, where there clearly are oligomer products present after 200 temperature cycles. The concentration, anyhow, is significantly lower than for the other two x8-samples.

The Network sample produces the total highest amount of product, especially for short and medium-length DNA of up to 144 nt. The sequences selected by the templated ligation reaction from the 4096 possible AT-only 12mers apparently build an interconnected network even by themselves. Subsequently emerging different oligomer product sequences in turn catalyze the formation of more oligomer products.

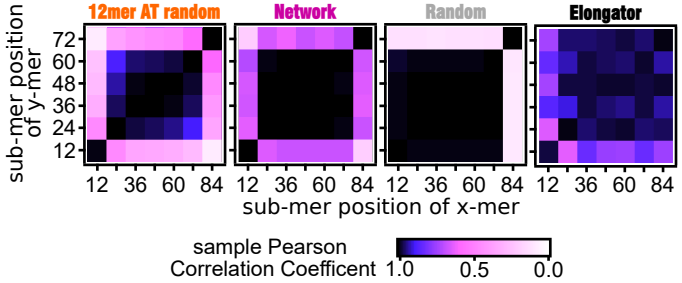

Figure 3.30 sample Pearson Correlation Coefficient for x8 sequence pools:
The sPCC-matrix of the "Network" sample (second from left) looks very similar compared to the sPCC-matrix of the entire 12mer AT-only full random sample (see Figure 3.19, reproduced on the left), despite only containing eight different sequences. The Random-sample shows a similarity for all start and center sequences and the end sequence respectively. The **Replicator** shows repeating patterns expected for the designed elongation mechanism.

By plotting the sPCC matrix comparing the subsequences in 84mers to 72mer for all three x8-samples the structure of the underlying network can be compared: Figure 3.29 shows that the Network sample has a very similar appearance compared to the sPCC matrix of the initial pool (reproduced on the left, see Figure 3.19). There is a distinct starting and a distinct ending sequence, despite the network that enables the selection of the eight sequences in Figure 3.28 not including the first and last subsequence of the original oligomer products. The center sequences are again very similar to themselves. For the Random sample, the sPCC-matrix suggests

strongly self-similar first sequences, with only the end sequences differing. This might indicate that oligomer products predominantly consist of only one sequence in the x8-Random pool. The Replicator sample shows patterns of self-similarity with a distance of 12 nt and 24 nt. With the designed overhangs in this sample the elongation is very specific, which might indicate the elongation by alternating 12mer sequences.

These distinct position dependent self-similarities in the sPCC matrices are again a sign for repeating motifs depending on the position in the strand. Figure 3.31 shows the probability to find A or T at each position in 84mers. Random and Replicator seem to favor A-type strands, while the Network sample has a trend towards T-type strands. In all three cases there are comparable patterns emerging: AT -alternating in the proximity of the ligation sites and a homogeneous section in between. This is especially apparent in the Replicator sample. The Random sample has almost no light or white color-codes. Such a lack of uncertainty in the base-probability points towards a structure made from only one sequence, as there would then only be slight variations due to read errors during the sequencing (see 5.16).

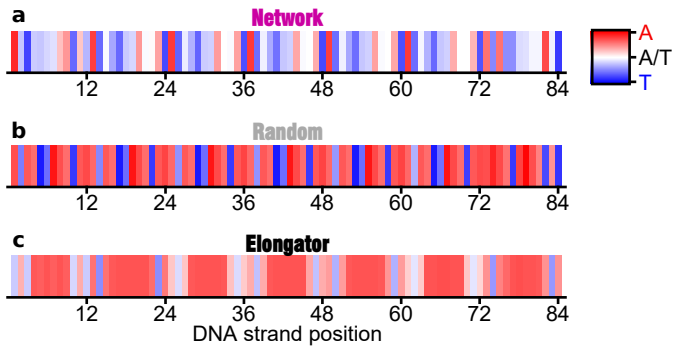

Figure 3.31 Position dependent base probability for 84mers in the x8-samples:
The sequences in the three pools are very different and thus, the sequence probabilities for A and T depending on the position within an 84mer.
a Network slightly favors T-type strands. Here, poly-T motifs are separated by AT -alternating motifs.
b Random A-type sequences are selected that show short poly-A sequences separated by single- and double- T .
c Replicator also favors A-type strands, but with long poly-A stretches and alternating AT -motifs in between. This is plausible as the pool is designed in a corresponding way.

The Network sample shows distinct ATAT at the ligation sites and less specific sequence patterns in between. Interestingly, this graph appears to include more T-type strands than A-type strands. In the AT -random experiment the resulting oligomer products did favor A-type strands. For all 12 nt long sequences in the three x8-pools a NanoDrop analysis (see Section 5.4) was performed to determine the stock concentration and to ensure a homogeneous concentration for all eight sequences in the x8 stock. The analysis of the A:T-fractions in Figure 3.32 shows that the initial pool for Network has a bias towards T-type strands, Random does also show a peak for about 50:50 % A:T and Replicator has, despite the sequence-symmetry of its initial pool, an unsymmetrical distribution. For Replicator and Random the slight bias towards A-type strands in oligomer products is again amplified by the templated ligation, and so is the bias towards T-type monomers in the Network sample. In all three cases the system selects oligomer strands with about 25:75 % A:T and T:A ratios to prevent hairpin formation in template strands.

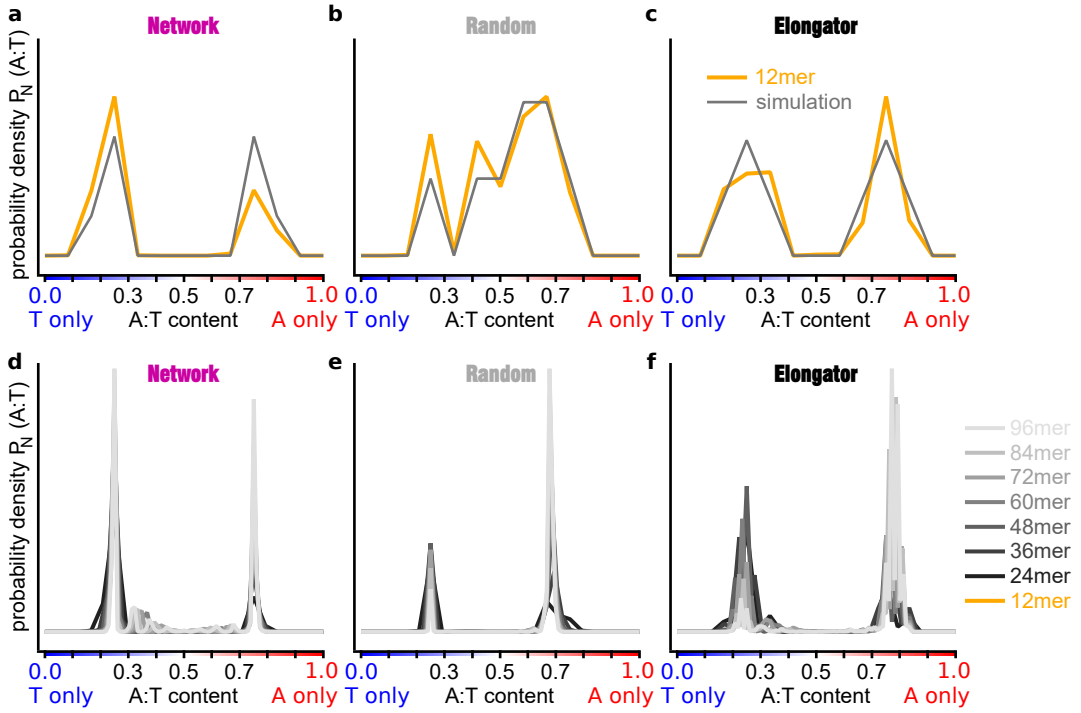

Figure 3.32 A:T-ratio evolution for x8-samples:

a, d `Network` Despite NanoDrop measurements, the T-type concentration in the initial x8 Network pool is higher than the A-type concentration. As argued before in Figure 3.5, a bias in the pool is amplified in the resulting longer oligomer products.

b, e `Random` By chance (meaning the A:T-composition of the eight randomly chosen sequences), the randomly selected pool for this sample is slightly biased towards A-type strands resulting in a bias for longer strands. Compared to the expected monomer pool composition, there is a lack of T-type strands.

c, d `Replicator` These sequences were designed to be reverse complements of each other and to facilitate complex-formation with base overhangs. But as in panel a, a bias in the initial pool towards A-type strands is amplified in oligomer products.

de Bruijn network graphs for x8-samples

In contrast to the AT-random sample, where, in order to analyze the oligomer product subsequences in a de Bruijn graph a threshold filtering step has to be applied, all sequences in the x8-pools have a high read count. The simplicity of the final network is due to the maximum of 64 edges from the eight possible nodes (sequences) per pool.

As expected from the sPCC-matrix (Figure 3.30) and the A:T-ratio analysis (Figure 3.32) the Network-sample de Bruijn network has a tendency towards T-type strands and is fairly interconnected, with all possible 64 edges being sequenced at least once. Therefore, sequences of the same base-type are predominantly connected. Similar to the AT-random oligomers, x8-Network-oligomers might also include reverse complementary sections. For the Random sample there is almost no network. The most common sequence tends to follow itself, as expected from the sPCC-matrix and base-probability analysis before. The lack of edges is no artifact from plotting, but not measured at all, meaning, that all oligomer products are made from either `TATTAATTTTTT` OR `ATAATTAAATAA` (the third largest sequence-count is irrelevant compared to the other two). The **Replicator** sample has two amplified groups,

- `ATAAAAAAAATA` - `TATAAAAAATAT`
- `TATTTTTTTTAT` - `ATATTTTTTATA`.

The groups are the perfect reverse complements of each other and can use the other four strands in the system as template. Both groups form sequence-alternating oligomers and have a very low chance to be found in the same strand as their reverse complement. The ligation sites show `...ATA-TAT...` or `...TAT-ATA...` in both cases. Poly-base motifs including `...AAA-AAA...` and `...TTT-TTT...` at the binding sites are present in the oligomer products but rare in comparison to the two major groups. This is similar for the AT-random data.

The subset of eight sequences, selected by the AT-random network actually manages to replicate distinct features - or more specifically: a small subset of sequences is responsible for several features in the full AT-random system. The designed sample behaves in a predictable way forming A-type and T-type strands and favoring `...ATAT...` motifs at the ligation sites. A small set of random sequences barely forms oligomer products strands, and can only do so with self-repeating patterns. The growth mode of the Network sample is presumably comparable to the AT-random system, while the **Replicator** builds long perfect reverse complements of its oligomers. Those long oligomers might act as stable, under mild thermal cycling inseparable, double-stranded two-component complexes that can be easily extended further (more details in Section 3.9). The growth-mode of the Random sample is likely only possible due to the high sequence-species concentration (compared to AT-random) and unlikely complex conformations with low life-times, that are connected only by chance. For other sets of randomly selected sequences there might be less or more product depending on the actual sequences, but it will likely always be worse than the Network sample, due to the unspecific "by chance" growth mode compared to the proven sequences selected by the AT-random experiment.

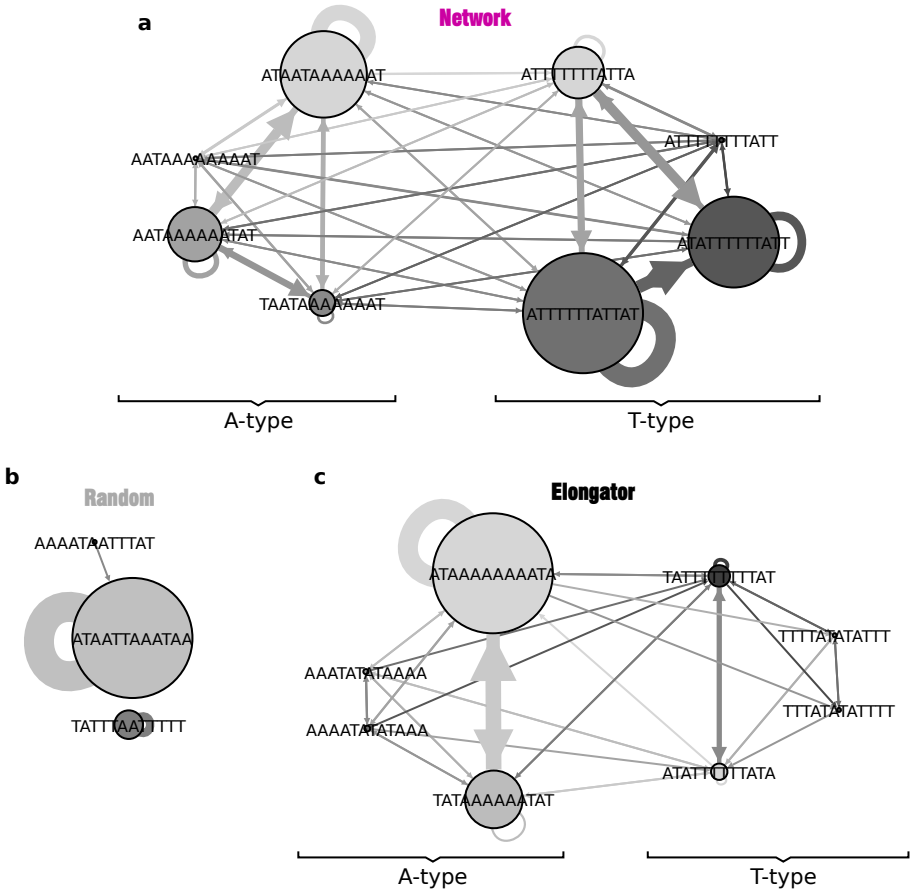

Figure 3.33 de Bruijn network graphs for 84mer oligomer products from the three x8 samples:
Each sequence-complementary network shows the A-type sequences on the left and the T-type sequences on the right. Edge thickness describes the frequency of the transition between two nodes. The size of a node describes the frequency of a subsequence-read in the oligomer products. Complementary sequences in a network are shown at the same mirrored position.
a, Network Shows an intricate network, where all eight sequences are interconnected. The sequence `ATTTTTTATTAT` is most common. There are frequent connections to other T-type sequences. As for the AT-random oligomers, `Network` oligomers tend to stay purely A-type or T-type.
b, Random As expected from the sPCC-matrix, long strands are made predominantly from two sequences, that both tend to follow themselves. From the other six sequences in the pool, five are not found in oligomer products at all.
c, Elongator The network is also interconnected, but has significantly fewer edges. The most common sequences `ATAAAAAAAATA` and `TATAAAAAATAT` are often found in an alternating motif, and so are their reverse complements. Oligomers with poly-`A` or poly-`T` at the ligation site are rare in comparison.

3.7.3 x1 pool

A self-complementary DNA strand is a system favored by theoretical modeling. Here, the sequence information is always the same and as a result the direct implementation of sequence can be neglected. Usually, the bases are simply described by an arbitrary letter like `N` that is defined to form double-stranded complexes by hybridizing with another `N`. The shortest self-complementary sequence would then be `NN`, a dimer. In the experiment, the melting temperature of dimers is far too low for the ligase to be active and probably also in a temperature range in which water is solid. Additionally, there is no real easily accessible self-complementary base in standard DNA. But just like in Section 3.7.1 a self-complementary base can be build as an artificial construct by substitution with several bases: the shortest `N` could be a section like `AT` with the resulting "dimer" `ATAT`. This complex can bind either to itself without overlap, as a complex with three strands or form a hairpin structure by folding onto itself. Again, the length of four bases is too short for the experimental system due to the low melting temperature. Therefore, the encoding of `N` is stretched to more bases and a systems with a total "dimer" length of twelve bases is built: `AAATTTAAATTT`.

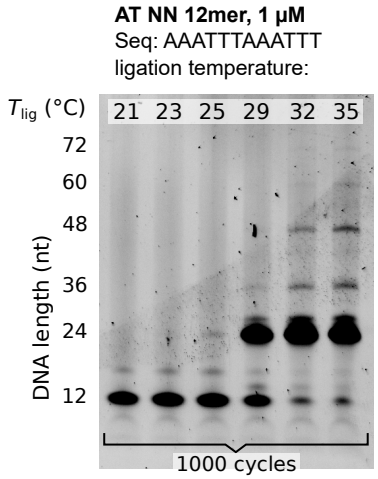

Figure 3.34 `NN` self-complementary DNA strands with length 12 nt:
12mer `AAATTTAAATTT` strands start forming oligomer products at a ligation temperature of 25 °C. Higher T_{lig} lead to the emergence of up to 60mer strands. The ligation is very efficient and after 1000 temperature cycles almost all 12mers experienced at least one ligation reaction as substrate strand.

Figure 3.34 shows a ligation temperature sweep for the 12mer-`NN`-sample as a PAGE analysis. At ligation temperatures of 21 °C and 23 °C there is no oligomer product observed in the lane. Starting with 25 °C 24mer oligomer products emerge. For higher ligation temperatures almost all 12mer "monomer" strands are ligated to at least 24mer strands. The 24mer `NNNN` is already long enough to form hairpin complexes - this inhibits the analysis with the concentration quantification LabView tool described in Section 5.6. For T_{lig} = 32 °C and 35 °C even longer oligomer products can be observed. In the last lane a lower concentration for 36mer is followed by a higher concentration of 48mers. This is even more pronounced in Figure 3.35.

Here, the sample concentration is 10 µM instead of 1 µM. The higher concentration leads to a higher product concentration also for lower T_{lig}. For T_{lig} = 40 °C Figure 3.35 a shows a distinct alternating concentration pattern in oligomer products. Strands made from an even number of 12mer "monomers" are underrepresented compared to their neighboring bands. The primary hy-

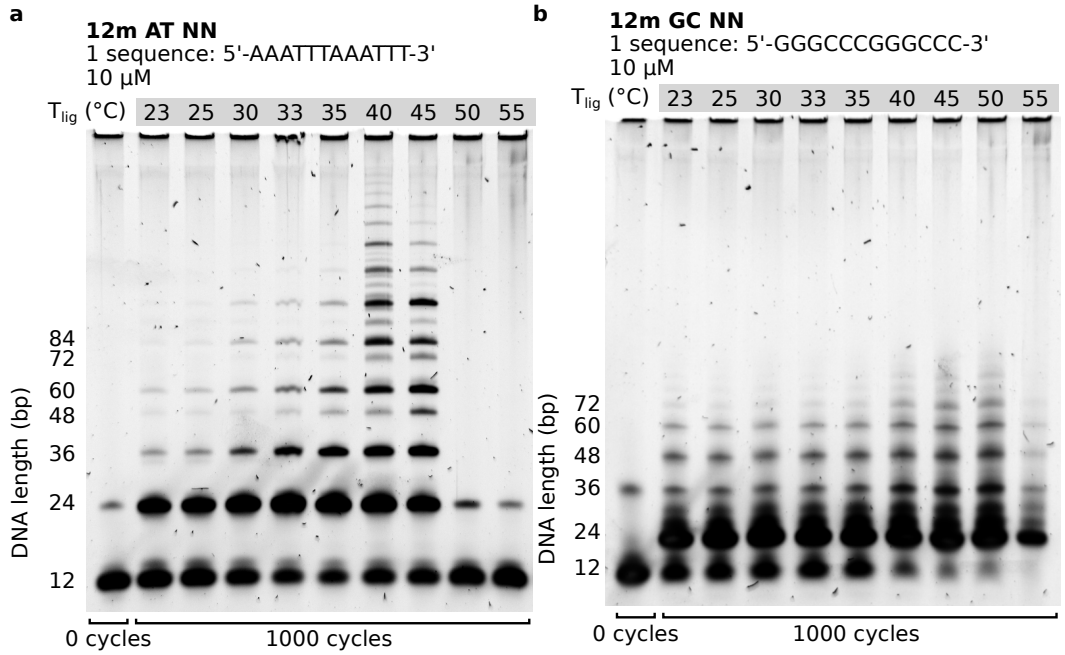

Figure 3.35 12mer AT-only and GC-only random `NN` temperature series:
a Similar to Figure 3.34, but with 10 µM pool concentration. The concentrations distribution is smoother than for 1 µM concentration. The highest output concentration and longest oligomer products lengths are detected for 40 °C and 45 °C. Again, strands with even counts of 12mer "monomers" are less frequent.
b For GC 12mer `NN` strands (`GGGCCCGGGCCC`) also build longer strands. The main difference to the `A` , `T` -based `NN` -sample is the significantly shorter length distribution for oligomer products. Due to the higher binding energy in GC-samples, double-stranded complexes are more difficult to dissociate. Therefore, single-stranded DNA that can act as template or substrate is even more rare.

bridization mechanism in this experiment is likely the formation of hairpin strands. Apparently, strands made form an uneven number of 12mer "monomers" are better at hairpin formation and thus less likely to be extended by another strand.

12mer GC-based- `NN` samples with the sequence `GGGCCCGGGCCC` do also form longer oligomer products. The lanes on the gel are less defined, a common observation for GC-rich DNA on PAGE. Even more pronounced as for the AT- `NN` experiments, the majority of GC- `NN` 12mer "monomers" are ligated and the original pool is drained almost completely for $T_{\mathrm{lig}} = 45\,°\mathrm{C}$ and 50 °C. In those lanes, the asymmetry of even and uneven strands lengths can also be seen, despite the even higher hybridization energy for GC-only DNA. For the experiments with only one sequence the entire pool will at a certain point in time be ligated to another strand or oligomer. In contrast to the random sequence pool, the limiting factor is only the hybridization alignment. Whereas for the random sequence experiments, the hybridization is much more complicated due to the sequence information in the strands. As described above, the hairpin formation of oligomer products governs the overall appearance of the oligomer length distribution. In the AT-random sequence samples and even the reduced complexity systems the suppression of hairpins lead to the emergence of A-type and T-type sequence groups. Here, the hairpins suppress certain lengths, because they would form less stable hairpins and can be elongated easier than uneven-length oligomers.

3.8 Oligomer length distribution dynamics

The experimental conditions that can be parameterized in this simple setup are the ligation temperature T_{lig}, the dissociation temperature T_{melt}, the ligation time in each cycle t_{lig}, the dissociation time in each cycle t_{melt}, the total initial pool concentration, and the sequence space per concentration. In the following the influence of each parameter is tested and its impact analyzed on the ligation dynamics of the system.

3.8.1 Ligation temperature

The ligation temperature T_{lig} might be the most straight forward parameter of the system. As the hybridization of two or more ssDNA strands depends on the binding energy, The system temperature is inherently connected to the the amount and length of the double-stranded section. For low system temperatures the melting curves (compare Section 5.3.2) show a large amount of bound bases, while for higher temperatures the probability to find double-stranded complexes is low. This in turn means, that a elevated T_{lig} the formation of short complexes is less likely and oligomer products, if any emerge at all, are long. For low T_{melt} the probability to find short oligomer products is higher.

Figure 3.36 a shows three PAGE gels with a total of six similar experiments from the same master mix. This mix was then divided into 36 separate tubes that were individually subjected to the temperature cycling, with $T_{\text{lig}} = (25,\ 30{,}35,\ 40,\ 45)\,°\text{C}$ at the same melting temperature of 75 °C and $t_{\text{lig}} = 120\,\text{s}$, $t_{\text{melt}} = 20\,\text{s}$. The oligomer length distributions are very similar for all six experiments. This is confirmed by the CQ-tool (see Section 5.6). For each of the six sets of six lanes the same analysis was done independently. The results for the concentration over length were then plotted together to quantify the difference and calculate the standard deviation.

- The reference samples, which were stored in the fridge for the experimental time, show artifacts at around 24mer length (see Figure 5.12). Otherwise, there is only the 12mer band for the initial pool visible.
- For $T_{\text{lig}} = \mathbf{25\,°C}$ the concentration distribution is exponentially falling, but has the largest amount of short oligomer products (24mer to 48mer), as expected from the temperature-dependent DNA hybridization described above.
- For $T_{\text{lig}} = \mathbf{30\,°C}$ the distribution is also exponentially falling, but only from the 36mers onward. medium-length oligomer products (60mer to 120mer) have the largest concentration for this experimental setup.
- $T_{\text{lig}} = \mathbf{35\,°C}$ is similar to 30 °C, but has an even shallower decrease with a higher concentration of long oligomer products (144mer and larger).
- For $T_{\text{lig}} = \mathbf{40\,°C}$ and $\mathbf{45\,°C}$, the CQ-tool does not quantify oligomer products with a concentration larger than the baseline. However, there are also no bands visible by eye in the corresponding gel lanes.

As predicted, the lowest T_{lig} produces the largest amount of short oligomers. At the ligation temperature of 25 °C the ligase is already active (see Section 5.5) while at the same time the temperature is so low, that even short double-stranded complexes are common in the ligation step. With the high binding probability for almost the entire ensemble of strands, the most likely reaction product are simply "dimers" made from two 12mer "monomers". The probability of a "dimer" to be extended to a "trimer" or longer are exponentially lower.

For very high T_{lig} of 40 °C and 45 °C there are no complexes made from at least three strands, that have a half-life time long enough to be ligated (for more details, see Section 3.9). Conse-

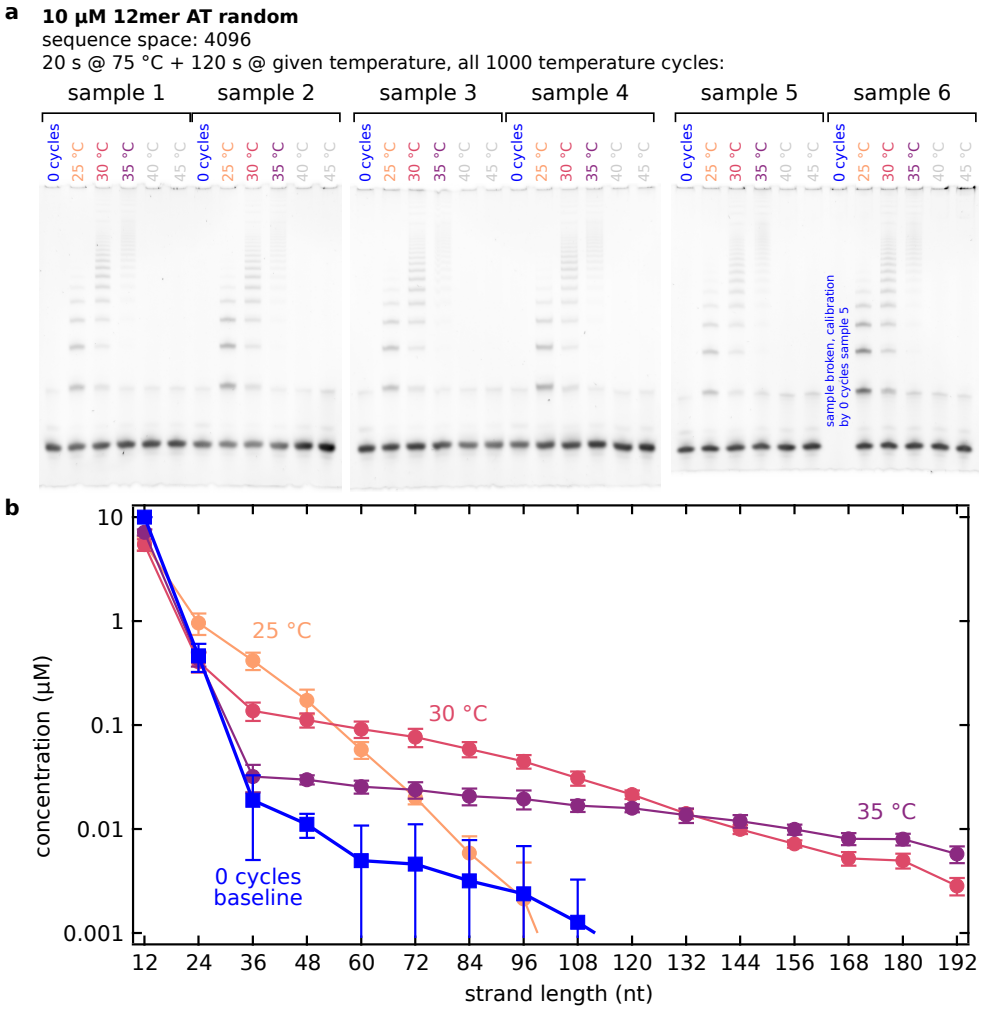

Figure 3.36 PAGE gels and concentration estimation with identical samples at different ligation temperatures:
Fluorescent images of PAGE gels, DNA is post-stained with SYBR gold. All lanes contain an aliquot from the same mastermix, that were exposed to the experimental conditions but in separate tubes. The dissociation temperature is 75 °C for all samples.
a The oligomer product length distribution depends on the ligation temperature. After 1000 temperature cycles, the sample with $T_{\mathrm{lig}} = 25\,°\mathrm{C}$ produces a short-tailed oligomer length distribution. For 30 °C and 35 °C the length distribution becomes long-tailed and at $T_{\mathrm{lig}} = 45\,°\mathrm{C}$ and 45 °C there are no product strands visible.
b Analysis of the PAGE by the concentration quantification tool (see Section 5.6). The analysis is performed and the concentration over length graph was obtain for each of the six experiments. The error bars stem from the difference between each experiment. For $T_{\mathrm{lig}} = 25\,°\mathrm{C}$ there is an about exponential decrease in concentration. For 30 °C and 35 °C the length distributions have a more shallow decrease towards longer oligomers and a lower concentration for short strands.

quently, there are no oligomer products after 1000 temperature cycles. As shown later (in Section 3.8.3), this is a function of ligation time - as the probability of finding and ligating a complex at elevated temperatures becomes significantly lower, it's not zero. Especially for $T_{\mathrm{lig}} = 40\,°\mathrm{C}$ there might be long oligomer product strands after several thousand cycles - assuming the ligase did not degrade.

For medium T_{lig} of 30 °C and 35 °C the temperature during the ligation step is so high, that short double-stranded complexes are rare. Overall, many hybridization events might take place, but the elevated temperature does not allow extended half-life times and strands quickly dissociate. Therefore, short strands are predominantly found in a single-stranded conformation. But by chance some short complexes are ligated and the resulting longer oligomers are now stable binding partners for either multiple monomer strands or other oligomer products. Consequently, the concentration of 24mer and 36mer strands is low, but longer oligomers like 84mers are more abundant for $T_{\mathrm{lig}} = 30\,°\mathrm{C}$ and $35\,°\mathrm{C}$ than for $25\,°\mathrm{C}$.

GC-only random sequence pools of different lengths were submitted to temperature cycling conditions as well. In contrast to the AT-only samples, in which the shortest initial strands that enabled the templated ligation reaction are 12mers, already GC-only 9mers produced oligomer product strands, as shown in Figure 3.37d. The shorter 8mer strands did not ligate at the same experimental conditions. As discussed in Section 5.3.2 and above, the melting temperature of the DNA pool dictates the extension characteristics of each sample. In order to compare GC-only samples to AT-only samples T_{lig} and T_{melt} have to be increased (see Figure 5.6). With the higher binding energy for `C` - `G` pairs compared to `A` - `T` and the less specific binding of `G`, the denaturing conditions on the PAGE gel are less efficient in dissociating dsDNA. This leads to smeared lanes on the gel, especially for longer strands. The larger binding energy, even though T_{melt} was increased to 95 °C, apparently prevented the 12mer (and to a certain extend the 10mer) samples from developing a long tailed product strand distribution. As the PAGE analysis is much clearer for AT-only samples and yield of GC-only samples in *Illumina* -sequencing is very low, most analysis in this study is performed on the AT-only sample.

In the following section the influence of the dissociation temperature T_{melt} is analyzed as a parameter sweep for the AT-only sample at a constant T_{lig}. But the oligomer length distribution int he analysis with varied T_{lig} also depend on T_{melt}. Figure 3.38 shows images of PAGE gels with the same T_{lig} conditions as above but $T_{\mathrm{melt}} = 95\,°\mathrm{C}$. The AT-random sample has the same characteristic as before, but only very low amounts of long oligomers for $T_{\mathrm{lig}} = 30\,°\mathrm{C}$. At 35 °C there are now no oligomers visible at all. Additionally, all lanes have a lower concentration of oligomer products than for $T_{\mathrm{lig}} = 75\,°\mathrm{C}$. For the x64 "double bases" sample, the results are also comparable, but as for the AT-random sample, the overall concentration of oligomers is low and the the long-tailed distributions consist of less long oligomers. Interestingly, for the AT- `NN` sample with a sequence space of just one, there are no longer oligomers than "dimer" (24mer) oligomer products. Here, even the shortest ligated product strands with a length of only 24 nt will most likely form hairpins after being dissociated at T_{melt} in each temperature cycle (see also Section 5.2). Apparently, the higher dissociation temperature is able to separate complexes, that were still bound at $T_{\mathrm{melt}} = 75\,°\mathrm{C}$. Hence, the system needs double-stranded, non-melting complexes to produce more oligomer strands. In Section 3.9 a numerical simulation of a random templating systems is discussed, that introduces the concept of "elongator" complexes: such complexes have long double-stranded sections and dangling ends which are easily extended.

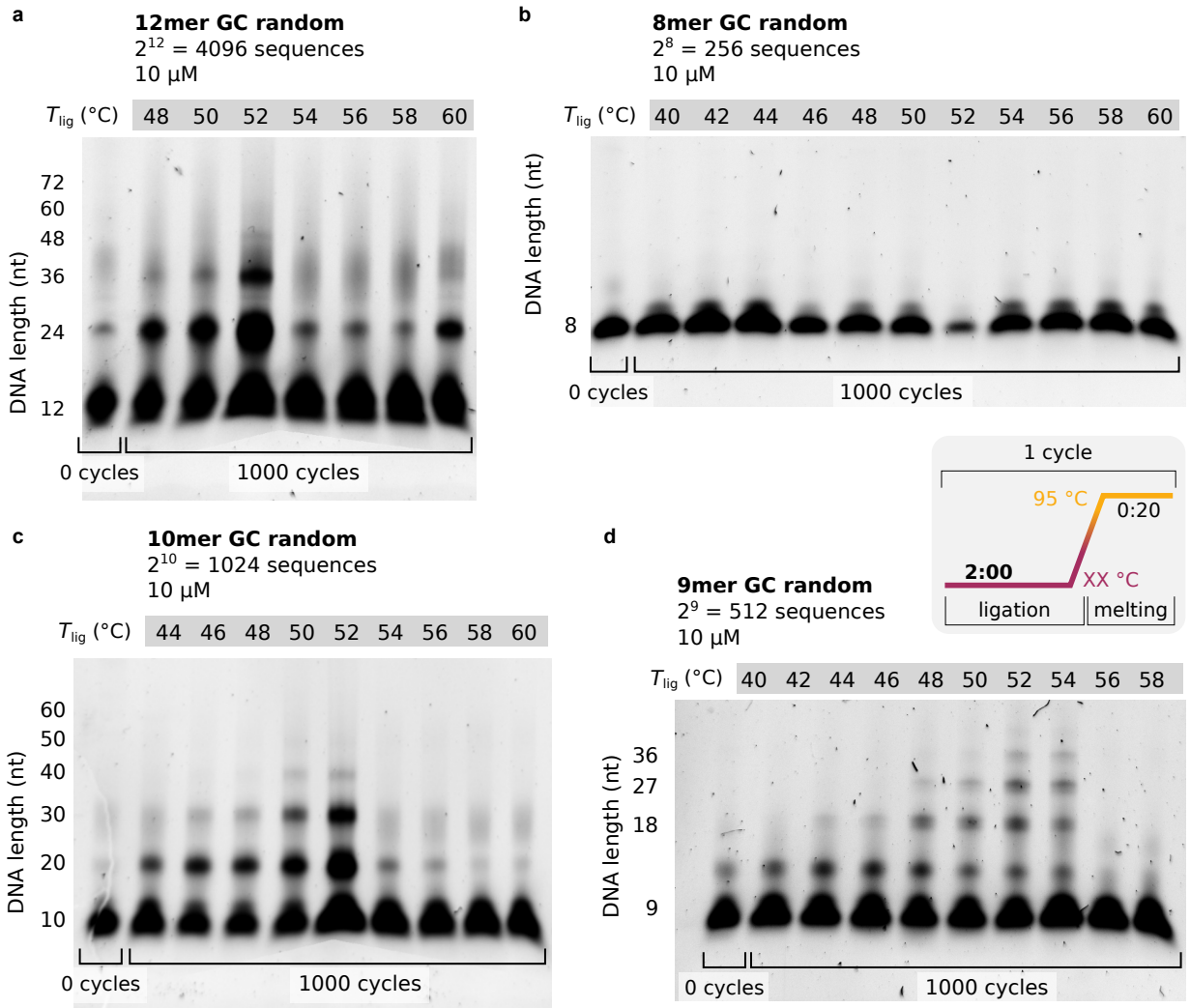

Figure 3.37 GC-only random sequence pool ligation, T_{lig}-series:
a 12mer GC-only random sequence pools give rise to longer strands, as already discussed in the sequencing data, see Section 3.6. Each band has a wide spread smear that points towards incomplete dissociation of dsDNA on the denaturing PAGE in the context of GC-only strands. This likely the reason for the less prominent smear for shorter strands.
b 8mers show no product strands for the given experimental conditions.
c The 10mer sample is remarkably similar to the 12mer sample in panel a, as expected from the sequencing data in Section 3.6.
d 9mers interestingly create a lot of product strands with clearly distinguishable bands in each lane.

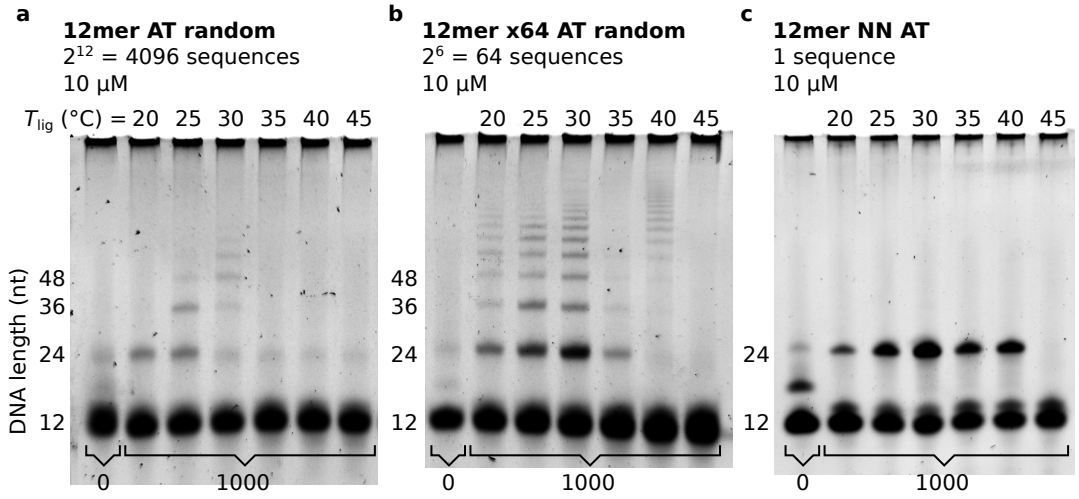

Figure 3.38 Temperature series for AT-random, x64 "double bases", and `NN` sample at 95 °C dissociation temperature:
a For the AT-random sample, the length distribution is very short in comparison to the same experiment with a dissociation temperature of 75 °C, as seen in Figure 3.36. The shift in the length distribution from T_{lig} = 25 °C to 30 °C is also visible here.
b The x64 "double bases" sample produces significantly more and longer oligomer product strands in 1000 temperature cycles, due to the larger relative concentration. The sample in the lane with 35 °C had a tube with a not completely closed lid and therefore shows an inconsistent length distribution. At T_{lig} = 45 °C the ligation temperature is so high, that no product strands can be observed.
c The sample with a single sequence `AAATTTAAATTT` (compare Section 3.7.3) only produces 24mer "dimers" and no longer strands at all.

As a summary, higher T_{lig} leads to long-tailed oligomer product length distributions with a shallow exponential slope. Low T_{lig} conditions facilitate the emergence of short-tailed distributions with steep exponentially decaying length distributions, and high concentrations of short oligomer products.

3.8.2 Dissociation temperature

The dissociation or melting temperature T_{melt} in the experiment is applied at the denaturation step in each temperature cycle. The central idea of this step is to "reset" the system to a single-stranded state and allow for the formation of new complexes. Especially blunt ended double strands should be separated, because they cannot act as template nor substrate in subsequent templated ligation reactions.

In the end of Section 3.8.1 the difference between T_{melt} = 75 °C and 95 °C are shown: for the higher dissociation temperature the emerging oligomer products have a lower concentration and a short-tailed length distribution. Here, the transition from short- to long-tailed oligomer concentration distribution is analyzed with a T_{melt} parameter sweep for the same T_{lig} of 33 °C. In Figure 3.39a, b the AT-random sample has very similar lanes for high T_{melt} of 60, 65, 70, 75 and 80 °C. The length distribution is long tailed and with an about exponentially decreasing slope. For the low melting temperatures of 45 and 50 °C almost no sample is visible by eye, but the CQ-tool measures a steep exponential concentration over length decay for oligomer products. For a melting temperature of 40 °C, there is no product at all. The most interesting lane is for the melt-

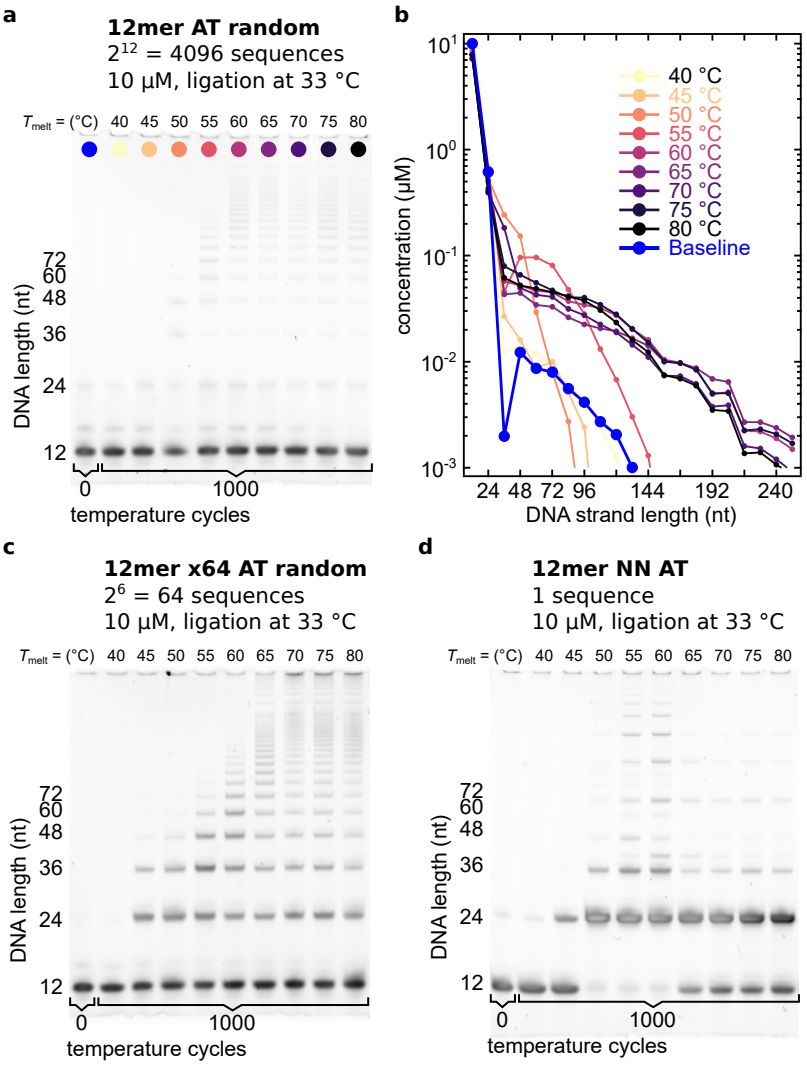

Figure 3.39 T_{melt} (dissociation temperature) sweep for AT-random, x64 "double bases", and `NN` samples:

SYBR gold post-stained images of PAGE gels. All lanes on each gel are aliquots from the same mastermix subjected to different T_{melt} experimental conditions.

a, b AT-random 12mer sample PAGE and concentration quantification: for low T_{melt} there is no oligomer product visible at all and the CQ gives values similar to the baseline. Starting at $T_{melt} = 50\,°C$ the typical about exponential decrease in concentration over length is measured. For $T_{melt} = 60\,°C$ to $80\,°C$ there is no clear difference between the bands and the respective concentration graphs in panel **b**. In between, there is a transition at about $T_{melt} = 55\,°C$. For lower T_{melt} the distribution is a steep exponential decay, while for higher T_{melt} it's a more shallow exponential decay with a larger concentration for long strands. The transition, however, features a local minimum at strand length of 36 nt and a local maximum at 60 nt.

c x64 "double bases" AT-random 12mer sample: there is also no product for $T_{melt} = 40\,°C$. Starting at $45\,°C$, there is a smooth transition from a short-tailed to a long-tailed concentration distribution.

d AT-`NN` sample: the single sequence pool first produces small amounts of 24mers at $T_{melt} = 45\,°C$. For $T_{melt} = 50\,°C$, $55\,°C$, and $60\,°C$ almost all 12mers are elongated or attached to oligomer products. For higher temperatures the lanes and especially the bands for long oligomers don't differ much.

ing temperature of 55 °C, at the transition from short tailed to long tailed length distributions. The oligomer distribution has a medium-length, between the groups of low T_{melt} and high T_{melt}. The main difference is the slope: for long oligomers it's exponentially decreasing but for shorter strands there is a local minimum (36mers), followed by a local maximum (48mers and 60mers). This behavior is not expected for the transition.

Figure 3.40 shows a more detailed resolution of the transition between $T_{\mathrm{melt}} = 50$ °C and 55 °C in 1 °C steps. Indeed, the transition from short- to long- tailed distribution with the emergence and vanishing of the local minimum-maximum feature is smooth as a function of T_{melt}. On this

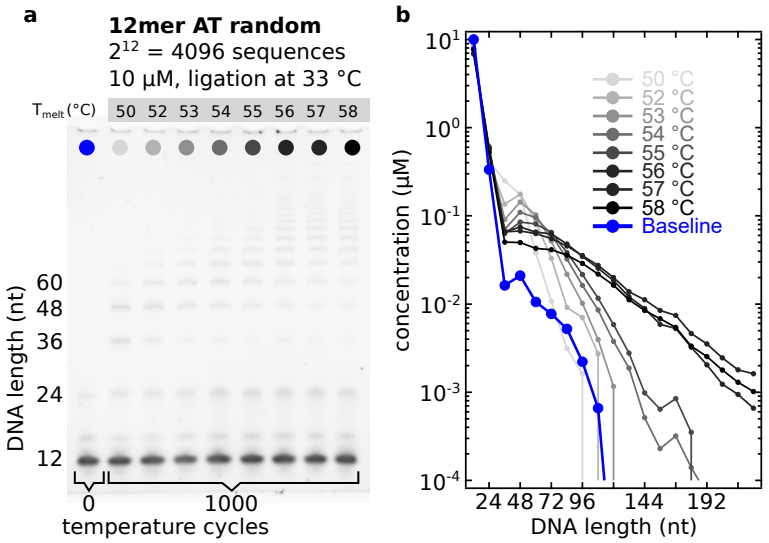

Figure 3.40 T_{melt}-dependent concentration distribution transition in AT-random samples:
All lanes except the reference lane were cycled 1000 times between $T_{\mathrm{lig}} = 33$ °C and T_{melt}. All lanes are an aliquot of the same mastermix.
a SYBR gold post-stained image of a PAGE gel. The bands for 36mer products become less abundant for higher melting temperatures. Also, the the concentration distribution transitions from a short-tailed to a long-tailed one.
b Concentration quantification of the PAGE gel. While $T_{\mathrm{melt}} = 50$ °C has a steep about exponentially decreasing concentration distribution, $T_{\mathrm{melt}} = 58$ °C has a long-tailed shallow exponential decreasing distribution. In between, the "feature" of a local minimum and maximum seen in Figure 3.39 develops and fades away again.

PAGE gel the transition feature is clearly visible, even without the CQ analysis in Figure 3.40 b. In the section above (Section 3.8.1), the influence of double-stranded "elongator"-type complexes was first mentioned. Those complexes are likely responsible for the feature of local minimum-maximum seen here (in Section 3.9 this is discussed in more detail). Basically, the T_{melt} temperature range of 52 to 55 °C allows double-stranded complexes with blunt ends and with dangling ends to stay in a hybridized conformation. The blunt-end complexes are called protectors and are responsible for most of the maximum, while the dangling-end complexes facilitate the rapid growth of long oligomer products.

For the x64 "double bases" AT-random sample in Figure 3.39 c the length distribution for $T_{\mathrm{melt}} <$ 70 °C simply becomes short tailed without the emergence of the minimum-maximum feature.

The `NN` sample with a sequence space of one has a very short-tailed lengths distribution for high T_{melt}. For 50, 55 and 60 °C almost all 12mer monomers are ligated to at least 24mers, with a long-tailed oligomer length distribution. For low T_{melt} there is almost no product other than 24mers again. In the lanes with $T_{\mathrm{melt}} = 55$ °C and 60 °C the already known pattern of alternating

high and low concentrations in oligomer products due to hairpin formation is clearly seen (see Section 3.7.3). Here, a high T_{melt} also leads to the dissociation of complexes with dangling ends that are assumed to be responsible for a majority of oligomer products in long-tailed distribution (see Section 3.9). The `NN` sample has the unique ability to always form hairpins in all strands and especially for all oligomer product strands. The formation of multi-strand double-stranded complexes is rare, due to the high local concentration of reverse complement sequences (typical for hairpins, see Section 5.2). Still, those structures are present in the sample, simply because of the small sequence space. For high T_{melt} those complexes are dissociated in every cycle and long oligomer products, due to the dependence of the dissociation energy on the binding length, are more likely to form hairpins in the following temperature cycles. Thus, high T_{melt} facilitate short-tailed distributions.

The local minimum-maximum feature in the AT-random sample present at 55 °C might be seen in the `NN`-sample as well: because blunt end and dangling end complexes are not dissociated, a long-tailed distribution emerges. And due to the high concentration of templates and the minimal sequence space, almost all 12mers are attached to other 12mers or oligomer products. At the same time, the dissociation temperature is high enough to dissociate the 24mer "dimer" oligomer products from their template, which in turn are likely to form hairpin structures in the subsequent temperature steps. This dissociation is necessary for a high oligomer product concentration, because otherwise neither 24mer hairpins nor multi-strand complexes are dissociated. Consequently, this leaves no complexes for templated ligation reactions. Additionally, the numerical simulation suggest that a low T_{melt} also leads to a slower rate of oligomer production (see Section 3.9).

3.8.3 Ligation time and cycle frequency

With Figure 3.40 d suggesting the dependence of the ligation time also affecting the oligomer product concentration, this section analyzes the variation of the ligation time per cycle t_{lig}.

In Figure 3.41 a the AT-random sample is subjected to 1000 temperature cycles with $t_{\text{lig}} = 10$, 20, 30, 40, 60, 90 and 120 s. Figure 3.41 b does the same for the x64-AT sample. For the short t_{lig} up until 40 s there is no product visible at all. At longer t_{lig} more oligomer product emerges with a long-tailed concentration distribution. In contrast, the x64 sample which is basically also an ensemble of random sequences, but with a sequence space of only 64, has almost similarly looking lanes for all t_{lig}. The AT-random sample has a factor of 64 greater sequence space compared to the x64 "double bases" sample. And the resulting conformation space for the assembly of double-stranded and ligatable complexes is smaller as well. The hybridization into double-stranded complexes is driven by diffusion (see Section 5.3.2) and therefore depending on the t_{lig}-time step: long t_{lig} allow for more complexes and optimal conformations, before said complexes are dissociated by T_{melt} in the dissociation step of each temperature cycle. For short t_{lig} the probability of two or more strands to hybridize to a ligatable complex is comparably lower. Consequently, a longer t_{lig} leads to more oligomer product strands. Though, at some point, which depends on the DNA concentration, the viscosity of the fluid and the sequence space, an increase in t_{lig} will only have a minor change, because the majority of strands that can hybridize and form complexes is already bound.

In Figure 5.3 the $k_{\text{on}}^{\text{hyb}}$ hybridization on-rate is estimated for different complex-configurations and sequence spaces. Based on studies by Wetmur, Kinjo, Schön, and Cisse [26, 71, 108, 129], the upper bound for the on-rate is assumed to be in the range of 1/(µM s). For the AT-only random sequence 12mer pools with $c_{\text{total}} = 10\,\mu\text{M}$, $k_{\text{on}}^{\text{hyb}} \approx 6$ s. Figure 3.43a shows that for $t_{\text{lig}} = 10$ s no

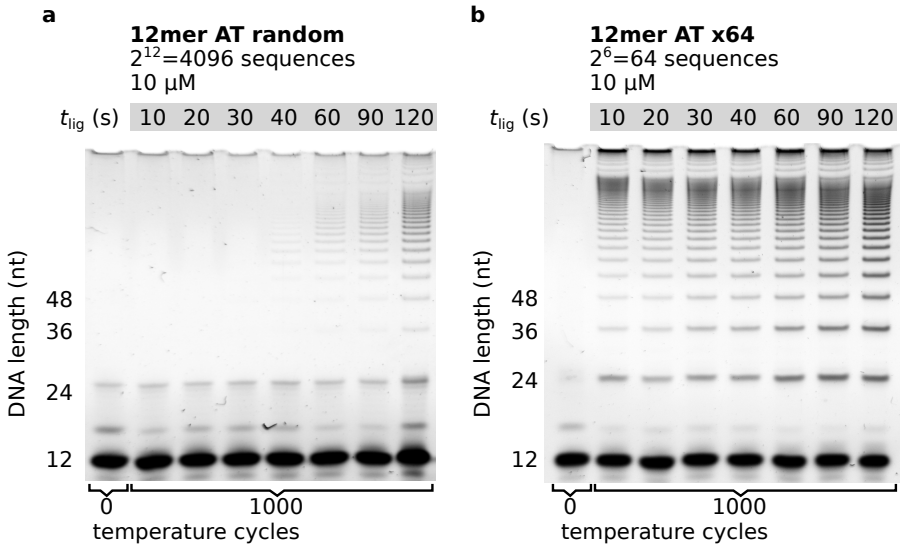

Figure 3.41 t_{lig} sweep for AT-random and AT x64 "double bases" samples:
The ligation time step t_{lig} is varied for all lanes. With a cycle count of 1000, the total experimental time is longer or shorter for the respectively t_{lig}. The temperature is similar for all cycles: T_{lig} = 33 °C, T_{melt} = 75 °C.
a AT-random: for short t_{lig} = 10 s to 30 s no oligomer products can be observed. From the first visible bands at lengths of about 72 nt at t_{lig} = 40 s and longer bands develop for longer t_{lig} with a higher total product concentration.
b In stark contrast to the AT-random sample, oligomers emerge in the x64 "double bases" sample for all t_{lig}. For longer t_{lig} the concentration of especially short oligomers is larger than for short t_{lig}.

product strands emerge after about 66 hours of total experimental time. At the same total experimental time but for t_{lig} = 30 s a significant amount of product strands emerged. For the reduced sequence space of the x64 sample t_{lig} = 10 s is already sufficient for extensive emergence of product strands. Here, k_{on}^{hyb} is assumed to be about 0.8 s.

After 1000 temperature cycles the samples with the longest t_{lig} per cycle have the highest oligomer concentration. But at the same time, the system with short t_{lig} also has the shortest total time that can facilitate ligation reactions (because each cycle consists of heating, T_{melt}, and cooling back down to T_{lig}, therefore, more cycles in a similar total experimental time result in a short accumulative time at T_{lig}). As described above, the total oligomer product concentration depends on t_{lig} and as seen in Figure 3.42 the amount of temperature cycles does also affect the oligomer concentration. The amount of temperature cycles increases the amount of oligomer product. This is not surprising for a closed system without a (significant, compare stability of DNA in water in Section 5.1) denaturation process of the oligomers. But there seemingly is only a small difference between 800 and 1000 temperature cycles, which might indicate a slowing down of the reaction. In Figure 3.53 the experiment is run for up to 2000 temperature cycles. The difference for medium-length strands becomes small and there is a higher amount of products strands only for very long oligomers. This means that even after 1500 temperature cycles, the ligase is still active and the system still has enough monomer strands with the correct sequence pattern and A:T ratio to elongate existing and new complexes.

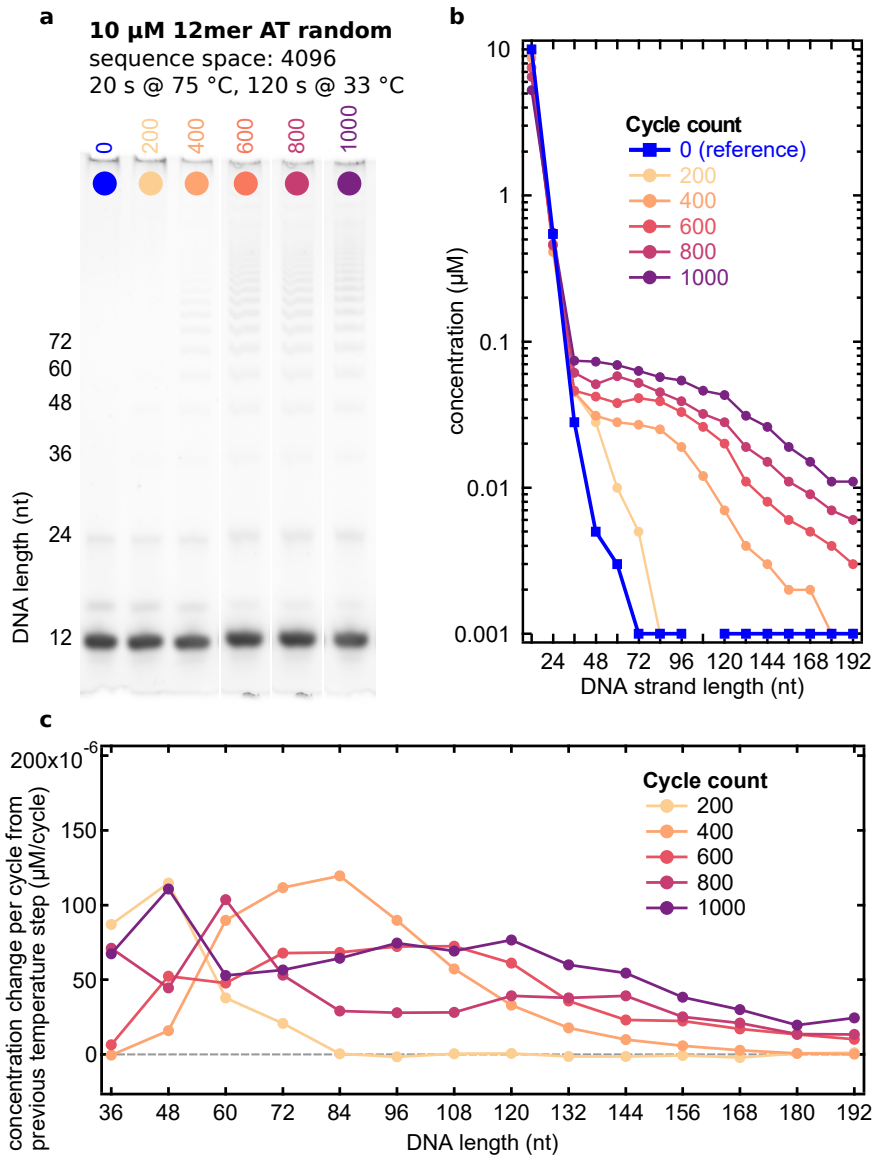

Figure 3.42 Time resolved AT-random experiment:

a PAGE image of the time resolved AT-random experiment.

b Concentration quantification of panel a. Although there are no bands visible by eye for 200 temperature cycles, the quantification tool can identify low concentrations of up to 72mer oligomer product strands. For temperature cycles the concentration increases steadily and the slope of concentration over length becomes flatter and more long-tailed.

c Production rate as the difference to previous cycle-step (graph) in μM/temperature cycle. Taking each lane as a snapshot in time shows different strand concentrations as the foundation of subsequence templated ligation reactions. Calculating the length-dependent production rate in μM/temperature cycle reveals different length regions that are predominantly emerging due to the templated ligation strand extension. Initially, the emerging oligomer strands are short and long ones are rare. After 400 temperature cycles, short strands are significantly rarer than medium-length oligomers. After 1000 temperature cycles oligomer lengths of 48mers and 120mers have the largest production rate compared to the previous step.

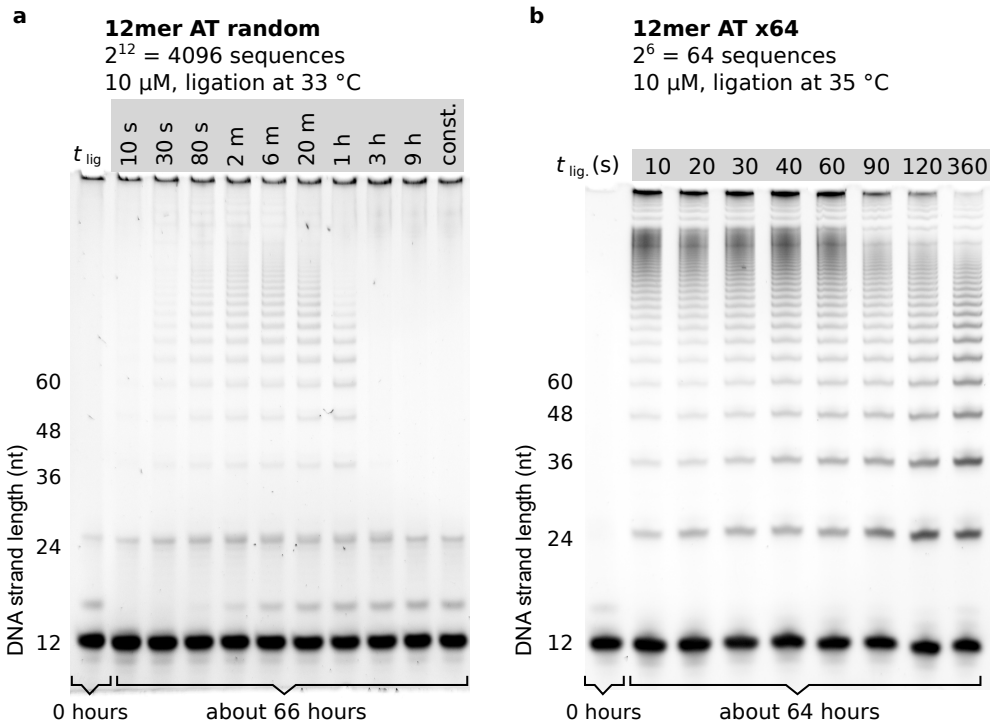

Figure 3.43 t_{lig} sweep with same total cycling time for AT-random and AT-x64 "double bases" samples: In contrast to Figure 3.41 the samples here are temperature cycled for the same amount of time (66 hours) instead of the same amount of cycles. **a** AT-random: with the same total cycle time even t_{lig} = 20 s produces oligomer product strands. With roughly 3500 temperature cycles the sample still has a lower total concentration than the sample with t_{lig} = 360 s cycled for 66 h (about 500 temperature cycles).
b The AT-x64 "double bases" sample has a 64 times smaller sequence space than the AT-random sample. The lanes for t_{lig} = 10, 20, 30, 40, and 60 s are almost similar, while the oligomer distribution becomes less long-tailed for longer t_{lig}.

To explain the dynamics here, a simple thought experiment is useful: Imagine starting the experiment with the similar pool, but additionally a small amount of varying concentrations and length distributions of already existing oligomers that also act as templates and substrates. This system with a "head start" of template strands will have different resulting rates depending on the oligomer length of the spiked oligomers and depending on the before mentioned parameters. In the random templated ligation experiment this thought experiment is still relevant, because the 12mer pool presumably, at least as quantified with the CQ-tool, visibly doesn't significantly and the A:T-ratio analysis in Figure 3.5 suggests that the pool is not substantially depleted of 12mer building blocks even after 1000 temperature cycles. At the same time the oligomer concentration distribution changes significantly between the different lanes (meaning after 200 more temperature cycles each). This hypothesis can be tested with the concentration data obtained from the CQ-tool. Dividing the concentration difference per oligomer length by the difference in cycle counts gives the mean concentration change per cycle between two time-/cycle-steps. Figure 3.42 c shows, that initially for 200 temperature cycles only short oligomers emerge. In contrast, from 200 to 400 temperature cycles especially medium-lengths and between 800 and

1000 cycles especially long oligomers emerge [i]. Overall, the total oligomer production rate is comparable in all steps and suggests, that the reaction is still ongoing after 1000 temperature cycles.

Because Figure 3.42 shows that more temperature cycles increase the amount of oligomers, and Figure 3.41 shows, that an increase in t_{lig} also increases the oligomer concentration for the same amount of temperature cycles, the experiment is repeated with the same total experimental time. Now, for $t_{\mathrm{lig}} = 20$ s Figure 3.41 a shows the emergence of medium-length oligomers. But in comparison, the oligomer concentration for $t_{\mathrm{lig}} = 360$ s is still higher. Presumably, the parameters cycle count and t_{lig} are interchangeable to a certain degree: as described in Section 5.3.2 the hybridization of strands forming double-stranded complexes is primarily driven by diffusion and can be estimated as a function of sequence space and strand concentration (also complex-alignment space and physical properties, that aren't varied here, like viscosity, salt concentration, surfactants, *etc*.). Therefore, the complex formation can be described as a function of reaction time at T_{lig}. Anyhow, the probability of three strands forming a complex is not linear in as a function of time, but converges towards a certain value for long times, as the system comes to an equilibrium hybridization configuration. This is the reason for the higher concentration for longer cycle times in Figure 3.43: a large amount of short temperature cycles has a lower total probability for complex formation than a lower cycle count with long t_{lig}. The lack of temperature cycles for long t_{lig} prevents the dsDNA complexes from dissociating and rehybridization in order to template new ligation reactions.

For the x64 "double bases" sample the distribution of oligomer concentration is long-tailed, already for $t_{\mathrm{lig}} = 10$ s, and stays almost similar for all steps until $t_{\mathrm{lig}} = 60$ s. For longer t_{lig} the distribution becomes more short tailed and has a distinct lack of concentration for long oligomers, while the bands for short oligomers of 24mers, 36mers, and 48mers show an increase in concentration, especially for $t_{\mathrm{lig}} = 360$ s. As argued above, an increase in total-t_{lig} in the experiment increases the concentration, and an increase in temperature cycles also increases the total oligomer concentration. Here, the small sequence space of the x64 "double bases" sample together with the same 12mer stand concentration of 10 µM allows for a faster hybridization of strands into complexes, effectively requiring a shorter t_{lig} for similar results as for the AT-random sample. Even the shortest t_{lig}-timestep of 10 s has a very similar appearance to the $t_{\mathrm{lig}} = 60$ s experiment on the PAGE gel in Figure 3.43 b suggesting, that the majority of ligation reactions that could happen did happen (for every timestep, as the resulting bands are very similar). Therefore, the lower cycle count for $t_{\mathrm{lig}} = 90$, 120, and 360 s results in a lower concentration of long oligomers.

Hypothetically, with a decrease of the cycle count and an increase of t_{lig} at some point the AT-random system should not show any product anymore, as hybridized complexes are reluctant to dissociate at $T_{\mathrm{lig}} = 33$ °C. In Figure 3.44 a the sample is incubated at a constant T_{lig} of 33 °C for about 60 hours and no oligomer products emerge. The system apparently does not dissociate double-stranded complexes at low to medium T_{lig} and thus prevents longer strands to act as templates for subsequent templated ligation reactions. In the case of AT-random temperature cycling seems to be necessary to at least partially "reset" the double-stranded complexes to single-stranded ones.

In contrast, the `NN`-sample shows the highest oligomer concentration for the constant t_{lig} without temperature cycling. Anyhow, the `NN`-sample is suspected to grow in a hairpin mediated way, as discussed in Section 3.7.3 before. In this particular case, the temperature cycling is

[i] Due to the comparably large error for short oligomer concentrations the graph starts at 36mers. Still, some artifacts that stem from the detection errors and the division of two small values are visible for 800 and 1000 temperature cycles at 60mer and 48mer lengths.

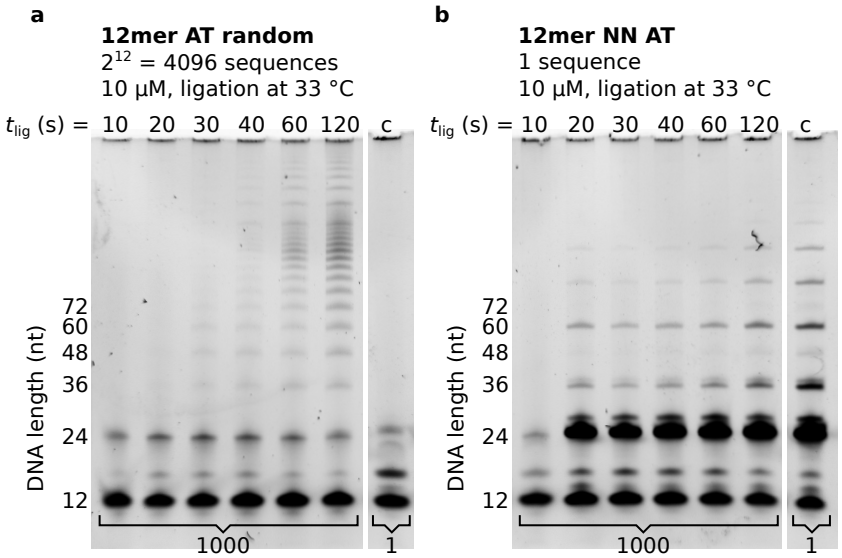

Figure 3.44 t_{lig} parameter sweep and comparison to constant temperature experiment:
The first six lanes show a t_{lig}-parameter sweep, as in Figure 3.41. The last lane marked "c" shows the same sample but stored at 33 °C for the duration of the experiment of about 60 hours.
a The AT-random sample does not show any oligomer product for the constant T_{lig} experiment without temperature cycling.
b In contrast, in the AT-`NN` sample with just one self-complementary sequence, oligomer strands emerge and show a high concentration.

actually preventing further ligation reactions, as the necessary hairpins and complexes made from several strands are dissociated.

3.8.4 Sequence space

The details in the dynamics of the random templated ligation reaction are dependent on the interplay of the experimental parameters. In the section above discussing the ligation time per cycle and the cycle times, the description and the interpretation cannot be made without mentioning the sequence space. This rather abstract parameter describes the set of strands with all possible permutations of the bases in each strand, that make up the sample. The sequence space is a function of the amount of bases and the length of the strand:

$$\text{base types}^{\text{strand length}}. \tag{3.5}$$

For a full random DNA with length 5 nt, there are already $4^5 = 1024$ possible strands. A large sequence space is sub-optimal for the experiment based on templated ligation. The underlying assumption is, that a large sequence space makes it more difficult to find reverse complement sequences to hybridize on the template strand, per concentration. For an example system of full ATGC-random 5mers, only 1 out of every 1024 strands is the correct perfect reverse complement fit. In comparison, the same 5mer with only bases `A` and `T` gives a sequence space of $2^5 = 32$, thus 1 out of 32 strands is the correct binding partner.

But not only the concentration of correctly paired strands is low for a large sequence space, strands with a partial reverse complement sequence can still hybridize and block the ligation- and hybridization site for the correct strand. Especially with a lot of different base-types, the amount of partial reverse complements is overwhelming. In the before mentioned example system with ATGC-random 5mers: the correct reverse complement for the sequence `AAAAA` in a full random 5mer pool is the sequence `TTTTT`, so again, 1 out of 1024 sequences binds ideally. But the amount of sequences with a single missmatch at the 5'- or 3'-end of `AAAAA` is six:

- `ATTTT`
- `GTTTT`
- `CTTTT`
- `TTTTA`
- `TTTTG`
- `TTTTC`.

In our experiment, the presumed first ligatable complexes are three 12mers, with two 6 nt long hybridized sections. In a ATGC-random case this would mean, that only 1 out of 4096 strands could ideally bind onto an already formed two-strand-complex.

Therefore, the easiest way to increase the likelihood of complex-formation is to reduce the sequence space. In our system we only include bases `A` and `T`, decreasing the sequence space of 12mers from $4^{12} \approx 16.8 + 10^6$ sequences to 2^{12}=4096 sequences. For the complex of three 12mers, this reduces the chance to find the ideal binding partner from 1 in 4096 to 1 in 64. The second way to reduce the sequence space, is to reduce the length of the monomer strands - but in this experimental system, the shortest AT-only strands for which the reaction works, are 12mers already. In Section 3.7.1 the sequence space is reduced by introducing "double-bases". Here, the sequence space for one strand to bind onto an already existing two-strand-complex is only 1 in 8. This difference in sequence space can be seen in Figure 3.41. After 1000 temperature cycles with t_{lig} of 10 s there are no oligomer products for the AT-random sample, while the x64 "double-bases" sample has ligated as much as for longer t_{lig}. Here, the 12mer pool concentration is similar for both experiments, meaning the amount of strands per volume is equal. The smaller sequence space enables the fast formation of oligomers for the x64 "double bases" sample. This means, that the timescale of oligomer emergence is directly proportional to the ratio of sample concentration and sequence space.

3.9 Numerical modeling of a templated ligation system

In physics, more than in any other field of science, there are two basic directions or research-methods that almost all studies can be classified by. Either experimental results are obtained and a subsequent model is build to explain the observed phenomenons, or the other way around, theoretical systems are designed with already known physics in mind and confirmed by subsequent experiments. Both methods can facilitate the understanding of complicated interactions, but the lack of experimental systems often limit the conformations of theoretical studies. At the same time, simple models might not explain all effects in experimental systems and real-world features might be neglected do to a lack of explanation. Up until here, the experiment carried the study and small models could explain or theoretically show comparable results, like the emergence of a bimodal A:T-ratio distribution in 24mers (Figure 3.3) or the emergence of the ligation-site `ATAT` sequence motif (Figure 3.17). In all cases, only the initial and the final properties could be observed and compared. But as seen in the extensive sweep of the parameter space in Section 3.8 above, the dynamics are a blurred idea of reality at best.

For insights into the hybridization, ligation and dehybridization dynamics of a random-sequence pool templated ligation reaction we modeled the experiment *in silico*. Modern days numerical computer simulations can substantially aid model-systems by introducing methods like the monte-carlo-simulation [13] or the Gillespie-algorithm ([47, 49, 50]). This simulation is based on ref. [104] and was created in cooperation with J. Rosenberger, T. Göppel and B. Altaner from the Gerland chair "Physics of Complex Biosystems" of technical university Munich (TUM). The design and implementation of the numerical simulation was done by Rosenberger, Göppel and Altaner, while discussion, verification, model improvement and study design were done in cooperation.

While the number of operations for the computer is only caped in the context of time, the limit for such models is usually found in the computer-memory. The sequence space of 4096 different species in the experiment results in a 4096x4096 matrix to store the interaction of 12mers already, and the RAM usage grows exponentially for oligomer products. Therefore, the model-system also uses a model DNA sequence: the base `N` is self-complementary and can therefore build double-stranded complexes `N-N`.

The theoretical foundation of the simulation is the Gillespie algorithm. As conceptualizing and programming of the simulation are part of the PhD thesis of J. Rosenberger and T. Göppel, only a short summary of the technicalities will follow, while the focus of this section lays on the connection of the results to the experiment and possible conclusions that improve the understanding of the experimental setup and observation.

3.9.1 Gillespie algorithm and simulation basics

During its introduction and first publication in the 1970s, the novel algorithm to describe reactions in well mixed gases by Dan Gillespie caused mixed feelings in the scientific world, as the author describes himself [51]. Nowadays, the Gillespie algorithm is the basis for statistical studies, not only on gaseous systems, but fluid systems and even SIR-systems simulating infectious diseases (*S*usceptible, *I*nfectious, *R*ecovered)). The advantage of this algorithm is its stark similarity to "real world scenarios". Instead of relying on bulk materials, the statistical approach simulates single strand species and their diffusion driven collisions. This is done by randomly selecting possible reactions with their individual reaction rates weighted by the probability that the reaction occurs (that is in turn based on the abundance of the reaction species). After a

reaction a waiting time-step is implemented that simulates the time until the next collision (exponentially distributed).

In ref. [104] the Gillespie algorithm is adapted to simulate the ensemble of DNA strands and the product strand length distribution characteristics that emerges due to the templated ligation of two strands. The relevant parameters of the model are only the reversible hybridization and dehybridization rates k_{on} and k_{off}, the mean binding energy per hybridized basepair γ, the systems influx of short oligomers, so that the concentration is constant c_μ, outflux k_{out}, and the rate of ligation of two strands k_{lig}. Additionally, the model can simulate temperature cycling conditions by introducing a cutoff $k_{cut} = k_{off}$ rate that implements a dissociation of all double-stranded complexes with a frequency $\tau \approx (k_{cut})^{-1}$. This variant is called "bounded model".

Figure 3.45 schematically describes the possible elementary processes in the simulation model. The system is initialized with a set concentration c_2 of dimer `NN` strands. All strands are di-

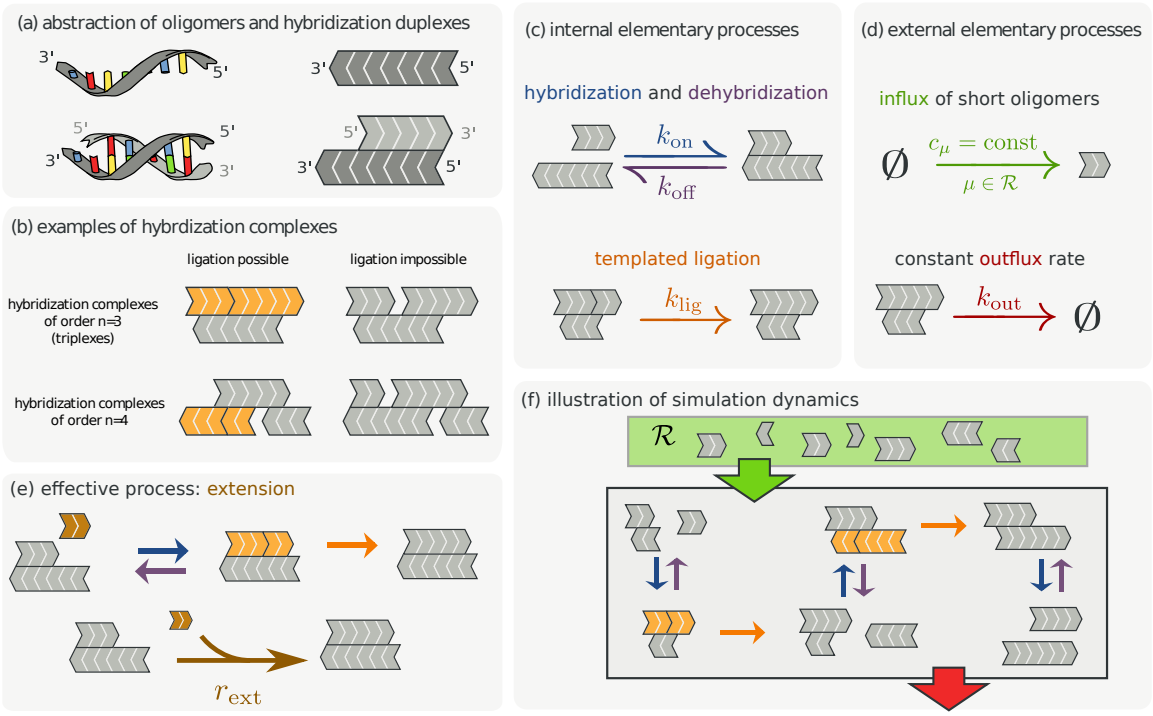

Figure 3.45 Elementary processes in the numerical model:
This figure is adapted from ref. [104].
a Abstraction of the bendable and helical DNA structure as stiff blocks, analogous to the rigid polymer assumption in the model.
b Examples for double-stranded hybridized complexes made from three or more strands.
c There are three elementary internal processes: the complex formation rate k_{on}, the complex dissociation rate k_{off}, and the ligation rate k_{lig}.
d The model is coupled to a reservoir of short oligonucleotides with an influx, so that their concentration is constant c_μ. The system outflux rate k_{out} is length-independent.
e The extension of a strand is an effective process with the rate r_{ext}that consists of the hybridization of at least three complexes and the ligation of the two substrates with the rates described in panel c.
f All processes together are the basis of the simulation and describe the strand elongation by templated ligation: short strands enter the system from the reservoir because of the chemostated boundary conditions. Multiple strands can then hybridize, ligate, dissociate, and thus develop a steady state length distribution.

rected with a distinct 5'-start and 3'-end. Two strands can form a complex by hybridizing with the rate k_{on}. If another strand can hybridize on the same complex to form at least a three-strand templated double-stranded complex, the two substrate strands are ligated with the rate k_{lig}, and dissociated with the rate k_{off}. The simulation system is open, meaning that the concentration of dimers is constant (coupled to a reservoir). At the same time every complex can leave the system with a rate of k_{out} effectively simulating a degradation process. Similar to the ligase enzyme in the experiment, double-stranded complexes with gaps between hybridized substrate strands are not ligated, as shown in Figure 3.45 b. But in contrast to the experiment, the ligase is not explicitly modeled as a particle in the system. Ligation is modeled by an effective ligation rate as an elementary simulation process k_{lig}. Also, as mentioned above, the simulation does not contain an explicit temperature cycling process, but can include an effective temperature cycling by the cutoff k_{off}-rate, k_{cut}. All elementary processes combined are shown in subpanel f. Importantly, the on- and off-rates are connected to the binding free energy of a hybridization cite:

$$\frac{k_{off}}{k_{on}} = c^0 e^{\beta \Delta G^0}, \text{with } \beta = (k_B T)^{-1}. \tag{3.6}$$

The exact timescale of the simulation is not important, only the ratio of the different reaction rates matter. Therefore, all rates are in units of 1/time and thus given without unit.

3.9.2 Simulation results - non trivial length distributions in oligomer products

If not stated otherwise, the simulation is run until a steady state length distribution is reached. Figure 3.46 shows the results for a parameter sweep of the outflux rate k_{out} in the bounded and the unbounded variant of the model. Limiting k_{off} to k_{cut}=0.005, the oligomer length distribution transitions from a short-tailed distribution for high k_{out}=10^{-4} to a long-tailed distribution for k_{out}=10^{-8}. For the unlimited dehybridization rate model, meaning k_{off} gets exponentially lower for large strands, a local minimum-maximum feature appears for k_{out}<3x10^{-7}. The outflux rate changes the height of the feature and its absolute position. For higher k_{out} the length distribution is very similar to the bounded variant of the model. Because the effective rate of complex formation r_{ext} is then significantly lower than k_{out}, emerging oligomer products have a too low half life time to sustain their population. In the mean strand length distribution in Figure 3.46c, d the bounded model shows a slight deviation from the exponential function of mean length over k_{out} around the critical k_{out}-rate.

The difference in both models can be visualized by comparing the elementary reactions rates to the oligomer strand length, as shown in Figure 3.46 e, f: The minimum k_{off} is the cutoff k_{cut} which is too large for the off-rate to intersect with the values of r_{ext} and k_{out} in this graph. The intersection of those rates in the diagram (as in panel f) mark the positions of the two critical parameters L^* and $L^\dagger$. The smallest oligomer product length L_0 for which so-called "extension cascades" become possible is called L^* with

$$1 = 2(2L^* - 1)c_2 k_{lig} e^{-\gamma L^*}, \tag{3.7}$$

the binding energy per nucleotide γ, and the dimer concentration c_2. The transition, where the dehybridization-rate becomes smaller than the outflux rate is defined at the point $L^\dagger$, with

$$k_{off}(L^\dagger, \gamma) = \frac{e^{\gamma L^\dagger}}{2L^\dagger - 1} = k_{out}. \tag{3.8}$$

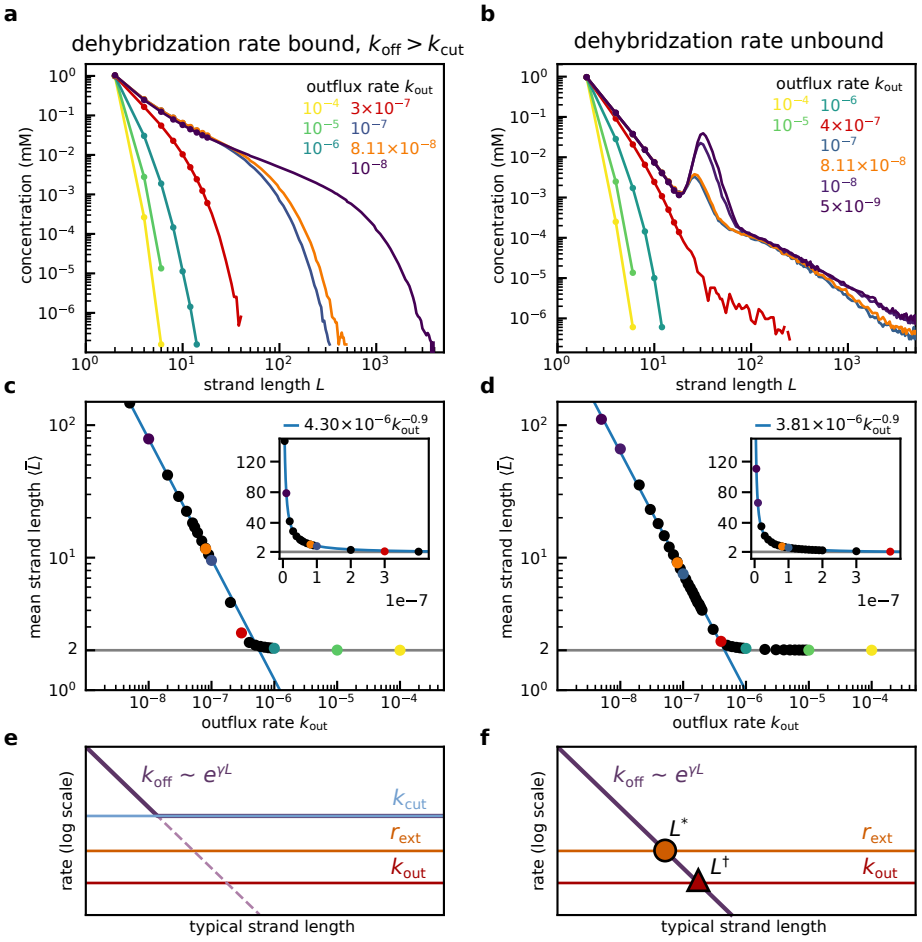

Figure 3.46 Oligomer length distribution for different outflux rates:
This figure is adapted from ref. [104]. The left column shows the bound model variant with $k_{off} > k_{cut}$, while the right column depicts the unbound model results.
a, b For outflux rates larger than the critical value (red graph) both systems have a short-tailed length distribution and very similar concentrations. For smaller outflux rates the unbound model develops a distinct non-monotonous oligomer length distribution with a local minimum (L_{min}) and maximum (L_{max}). A further decrease in k_{out} increases the maximum concentration and also shifts the maximum position towards longer strands, while the minimum is unchanged.
c, d The mean strand length of the emerging oligomers is constant for high outflux rates at the length of the elementary building blocks, $\langle\bar{L}\rangle = 2$.
e, f For small k_{out} the system is in a competition of the length-dependent dehybridization rate with either the effective rate of extension r_{ext} or the outflux rate. In the bounded model the dehybridization is the fastest rate and therefore, there is no intersection of k_{off} with r_{ext} or k_{out} in the graph.

Both equations (3.7) and (3.8) can be simplified for strands with strong binding energies $\gamma < -1$:

$$L^* \approx \ln\left(c_2 \frac{k_{\text{lig}}}{k_{\text{coll}}}\right) \gamma^{-1}, \tag{3.9}$$

$$L^\dagger \approx \ln\left(\frac{k_{\text{out}}}{k_{\text{coll}}}\right) \gamma^{-1}. \tag{3.10}$$

The above mentioned "extension cascade" is the most relevant dynamical feature in the simulation. An extension cascade can start on a double-stranded complex with at least one non-blunt-end. At a certain critical length in the unbounded model, the effective extension rate of the complex is comparable to the rate of dissociation which scales with the amount of hybridized bases in the complex. With every additional ligation event k_{off} decreases as the total binding energy of the complex increases, which makes further ligation events more likely. One of the major differences between the experiment and the simulation is the possibility to analyze the type of complex that can be found in the system at all times, and even analyze the typical complex structure as a function of strand length. In Figure 3.47 all relevant types of sequence structures are shown. For short oligomer strand lengths single-stranded complex (so simply single, non hybridized strands) dominate the system. At the maximum of the non-monotonous length-distribution double-stranded complexes are the most common, especially fully hybridized complexes with two blunt ends.

Other common types are odd-parity double-stranded complexes and higher-order complexes which are not ligateable in their current form (usually because of gaps between substrate strands). At around the minimum L_{min} even-parity complexes are very common. Those complexes grow by the process of extension cascades until the complex has two blunt ends, at which point further growth without a dissociation of the complex is not possible. The odd-parity double-stranded complexes on the other hand can grow in an almost uninhibited way until they are the most common complex in the system for very long oligomers. In a system with only even length building blocks odd-parity complexes cannot become blunt ended without dissociation. In Figure 3.47 b the fraction of complexes is shown instead of the log-concentration. Only around L_{min}, the point where extension cascades become relevant on the oligomer length scale, the system has a substantial part of complexes that are not single strands (for short lengths), fully hybridized strands (medium-length strands) or odd-parity strands (long strand lengths). How those complexes and the critical values L^* and $L^\dagger$ influence the understanding of the experiment is detailed in the discussion section Section 3.9.3.

The simulation standard conditions are selected to be comparable to the real world experimental conditions in order to allow for comparison. However, such selected values can be arbitrary to a certain degree, especially when applied in model system which might not include all processes. In such scenarios a parameter sweep of the utilized parameters is helpful in understanding the simulation steady state, especially whether unexpected features like the local minimum-maximum are artifacts or not. In Figure 3.48 a parameter sweep for the the outflux rate, the binding energy per nucleotide, the ligation rate, and the chemostated concentration of the "monomer" building blocks is performed. With the critical parameters identified in equations (3.9) and (3.10) above, the expected behavior can already be predicted: A decrease in the outflux rate changes the position of the maximum towards longer lengths. Additionally, the relative concentration of oligomers grows. In the limit of large outflux rates, as already seen before in Figure 3.46b, the minimum-maximum feature disappears. The position of $L^\dagger$ shifts towards longer strands, but in contrast to the analytical solution which is calculated for a continuous length, the building blocks in the simulation are dimer `NN` strands and the position therefore shifts in steps of length two.

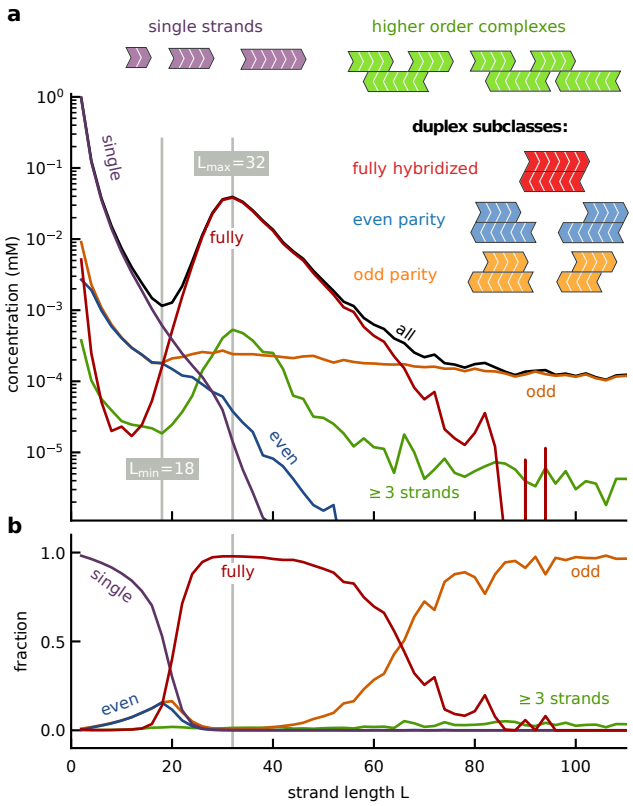

Figure 3.47 Abundant complex-types depend on the oligomer length:
This figure is adapted from ref. [104].
a The oligomer length distribution (black line) consists of different complexes depending on the length of the DNA strands. Until the local minimum $L_{\min}$ single-stranded DNA dominates. For the local maximum at $L_{\max}$ and similarly sized DNA strands fully hybridized double-stranded complexes with two blunt ends are the most common sequence type. At long lengths sequences in double-stranded complexes with uneven overhangs dominate.
b Plotting the graph in panel a as a fraction of all strands reveals the extend of the length-dependent complex abundance. Only close to lengths of $L_{\min}$ strands other than single, fully or uneven hybridized strands can be found in a non-negligible concentration.

The binding energy per base is chosen in agreement with the SantaLucia library for DNA hybridization energies [107] for strands made form `A` and `T` only. But the simulation can also implement an effective binding energy that might be altered by salts or surfactants in an experiment by varying the binding energy γ. An increase in the binding energy decreases both characteristic lengths L^* and $L^\dagger$ exponentially. For large binding energies, the peak width is narrow and the peak height large, while for low binding energies, the relative peak height is low, but the peak width almost 100 nt.

The ligation rate in the experiment is limited by the activity of the ligase, which depends on the temperature of the system. The exact activity, even of such an evolved enzyme, is difficult to describe, as it will depend on the concentration of ligatable substrates, their length, and their structure. For the Taq DNA ligase from *New England Biolabs* the activity is only given in units/ml. The manufacturer states: "one unit is defined as the amount of enzyme required to give 50 % ligation of the 12-base pair cohesive ends of 1 µg of BstEII-digested λ DNA in a total reaction volume of 50 µl in 15 minutes at 45 °C" (for a more detailed description see Section 5.5). The model can

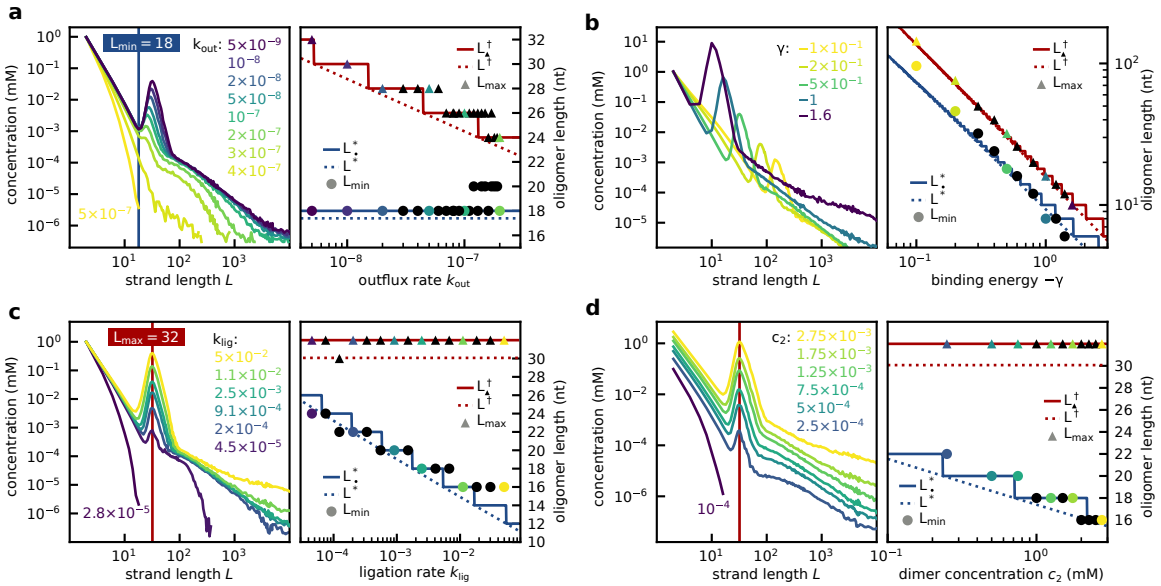

Figure 3.48 Parameter sweep for outflux rate, binding energy, ligation rate, and chemostated monomer concentration:
This figure is adapted from ref. [104]. The left graph in each subpanel shows the steady state length distribution.
a With a decrease of the outflux rate the concentration of oligomer products increases and the peak maximum shifts to longer oligomer lengths (as seen in Figure 3.46). The variation of k_{out} only has impact on the characteristic length $L^\dagger$.
b An increase in the binding energy γ shifts the peak towards shorter oligomer lengths and increases the relative peak height. In the simulation, both L^* and $L^\dagger$ are decreasing exponentially with γ, as expected.
c A larger ligation rate k_{lig} increases the peak height of the double-stranded complexes, but does not shift the peak position.
d Similarly, varying the monomer concentration c_2 changes the relative peak height, and also the total amount of oligomers in the simulation, as described in equation (3.9).

simply adjust the ligation rate k_{lig} to a value comparable to other system rates like k_{on} or r_{ext}. A small ligation rate has an overall comparable effect to an high k_{out}. Below a certain threshold the outflux rate is so large that the total length distribution remains short-tailed. In contrast to k_{out}, the variation of k_{lig} changes the position of the minimum L_{min} but not the position of L_{max} (described by L^* and $L^\dagger$). The concentration of the dimer-`NN` building blocks is chosen according to the Taq DNA ligase protocol and the experiments, 10 µM. A decrease in concentration leads to a smaller relative peak height and a more narrow peak, while L^* shifts towards longer lengths, see Figure 3.48c. For all varied parameters, the distinct local minimum-maximum feature appears above a certain parameter threshold and with well described positions for L_{min} and L_{max}. The onset of extension cascades is clearly the reason for this feature which leads to the emergence of persistent double-stranded complexes in the system.

Another mechanism that is inaccessible for analysis in the experimental setting, is the extension mode. Denaturing PAGE analysis and the sequencing of ssDNA strands cannot give detailed information about the complex types and how strands elongate exactly. The AT-only sample NGS analysis revealed sets of mutually complementary sequences in the A-type and T-type groups, as shown in Figure 3.28. But still, it is unclear if those strands grow by additions of single 12mers or by combinations of longer strands. Here, the simulation provides information about the com-

plex types by design and trajectories in complex-space reveal the extension modes by which the strands grow. Figure 3.49 shows the three distinct modes **primer extension**, **primer-template switching**, and **template extension**. Although Figure 3.49 shows the extension on only one side

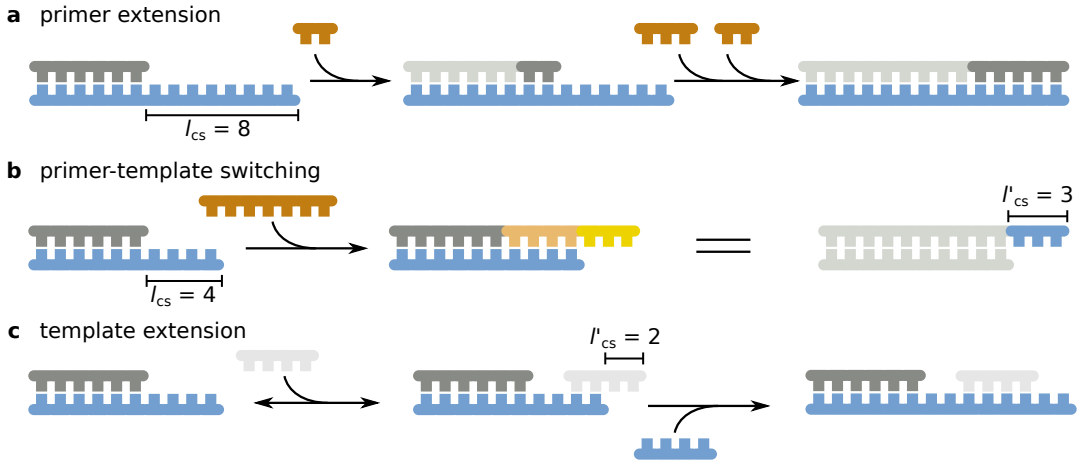

Figure 3.49 Strand extension modes discovered in the simulation:
This figure is adapted from ref. [104].
a, primer extension The substrate site is extended by the addition of short ssDNA substrate strands that hybridize on the single-stranded, dangling template strand. With each ligation reaction the iteite shrinks. The extension continues until the strand is either completely double-stranded with two blunt ends, or until the last attached substrate produces an overhang on the substrate site.
b, primer-template switching Extending the substrate site with a strand longer than the copy site l_{cs} effectively makes the original template site the new substrate site.
c, template extension With a "helper strand" the template side can be extended as well. After the ligation reaction, the helper strand can dehybridize resulting in a longer copy site of the template.

of the strands, these extension modes are symmetrical and occur on both ends of each complex. In primer extension mode a long template binds a shorter substrate. The dangling end on the template site of the double strand is called the copy site and enables the hybridization of a second substrate strand. After ligation, the two substrate strands act as the single, new, already hybridized substrate. If the copy site l_{cs} was longer than the added substrate strand, the copy site is still present, but with $l'_{cs} = l_{cs} - l_{as}$ ("as": added strand). The substrate site can then undergo subsequent extension reactions by templated ligation, until the complex either ends in a blunt end, or the last added substrate forms a new dangling overhang, but on the substrate side.

The latter case is similar to the second growth mode primer-template switching. If the extending substrate strand is longer than the copy site, the substrate side effectively becomes the new template in the complex.

In a third case, the template is extended, not the substrate. With the hybridization of a temporary template "helper strand" the original template strand of a complex can be ligated to another complex (single-stranded or double-stranded). This ligation reaction elongates the original copy site.

Tracking the initial complex sizes C_{initial} and final complex sizes C_{final} of all substrate strands in the simulation reveals an interesting effect, as shown in Figure 3.50a: Strands shorter than $L_{\max}$ predominantly result in final complexes of a similar length. Importantly, the simulation uses monomers and dimers as building blocks. For template strands of about the length $L_{\max}$

the final length is likely to be a complex longer than C_{initial}. The details of the calculations and

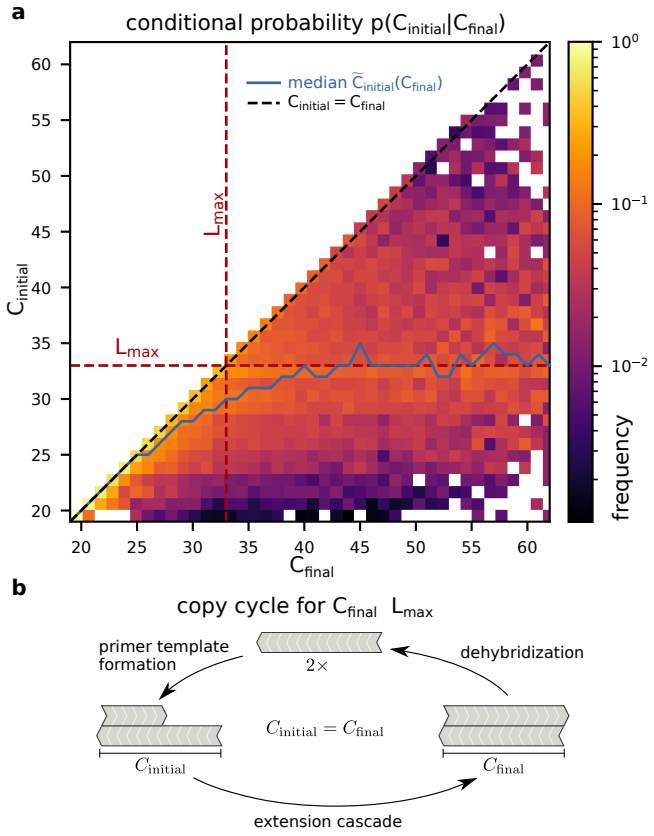

Figure 3.50 Dominant strand extension mode:
This figure is adapted from ref. [104].
a Comparing the initial complex length C_{initial} to the final complex length C_{final} shows that short initial complexes tend to result in complexes of similar size, as sketched in panel b. For strands with a length of approximately L_{max} or longer, C_{final} is likely to result in longer final complexes, probably du to the primer-template switching extension mode.
b A template stand with length C_{initial} binds a substrate strand, that is in turn extended by the ligation of another substrate strand. In the most likely case, the resulting strand has a final length of $C_{\text{final}} = C_{\text{initial}}$.

more intricate analysis of the simulation results is shown in ref. [104]. In the next section the implications drawn from the simulation that affect the experimental system with 12 nt monomers are discussed.

3.9.3 Simulation discussion - understanding the experiment

The simulation uncovered several insights, that were not clear or at all accessible in the experiment, such as the segmentation of complexes into several length-dependent sections dominated by different complex-types. A specialty and often purpose of simulations is the prediction of experimental data. Here, the characteristic lengths L^* and $L^\dagger$ were found in the simulation and verified as not being artifacts. They are used in the following to estimate L_{min} and L_{max} in the experimental setup.

With the hybridization of substrate strands onto template strands being the underlying mechanism for all extensions in this system, the effective mean binding energy per nucleotide will

influence the form of the resulting oligomer concentration distribution significantly. In Figure 3.51 a the effective melting curve for a system of short AT-only strands is plotted as the fraction of unbound bases *versus* the system temperature. The melting temperature of 50 °C and a transition width of about 10 K [64, 84] can be approximated by a Fermi-like function, where the probability p of nucleotide to be unpaired at a temperature T is given by

$$p(T) = \left(\mathrm{e}^{\frac{T - T_{\mathrm{melt}}}{\sigma}} + 1 \right) \tag{3.11}$$

with the melting temperature T_{melt}. The temperature dependence of the binding energy γ can then be approximated by the exponential factor and yields the off-rate

$$k_{\mathrm{off}} \approx k_{\mathrm{coll}} \, \mathrm{e}^{\gamma T}. \tag{3.12}$$

In Figure 3.51 c the effective dehybridization rate k_{off} is plotted against the length of oligomer strands. The horizontal lines mark the effective extension rate r_{ext} (top) and the inverse observation time $(\tau_{\mathrm{obs}})^{-1}$ (bottom) which replaces k_{out} in determining $L^{\dagger}$ (and thus L_{max}) in the transient closed simulation. Here, the rates are given in s^{-1} in order to enable the comparison with the experiment: the time per cycle is $\tau_{\mathrm{cycle}} = 180$ s, the effective collision rate of complexes $k_{\mathrm{coll}} = 1\mathrm{x}10^{5}\ \mathrm{s}^{-1}$, the extension rate $r_{\mathrm{ext}}=(\tau_{\mathrm{cycle}})^{-1} = 5.56\mathrm{x}10^{-3}\ \mathrm{s}^{-1}$, and

$$\tau_{\mathrm{obs}} = N_{\mathrm{cycle}} \, \mathrm{x} \, \tau_{\mathrm{cycle}} = 1.8\mathrm{x}10^{5}\ \mathrm{s}. \tag{3.13}$$

For low temperatures and no temperature cycling, the slope of the dehybridization rate over the strand length is steep and the characteristic values for L^{*} and $L^{\dagger}$ that determine L_{min} and L_{max} are smaller than the 12 nt monomer length of the experimental system. For an experiment with $T_{\mathrm{lig}} = 33$ °C and $T_{\mathrm{melt}} = 50$ °C the slope is still steep, but intersects with the rates of extension and inverse observation time. Anyhow, the rounded %12 values of the continuous model coincide at the 24 nt length - minimum and maximum are at the same position and thus a short-tailed length distribution is expected. Only starting for higher $T_{\mathrm{melt}} = 54$ °C the slope becomes flatter and minimum as well as maximum fall on the lengths of $L_{\mathrm{min}} = 36$ nt and $L_{\mathrm{max}} = 48$ nt. For this temperature the local minimum-maximum feature would be expected in the experiment. For even higher T_{melt} close to the melting temperature of short AT-only strands the slope of effective dehybridization over oligomer length becomes so shallow, that the intersections with the critical rates of r_{ext} and the inverse observation time only happen for very long oligomers. While it might be possible to identify those regions in the simulation, like in Figure 3.48 b, the analysis based on post-stained PAGE for experiments reaches its limits above 120-144 nt.

Figure 3.51d, e show the PAGE gel and the CQ-analysis for the corresponding experiment: 10 µM 12mer AT-only random DNA, with 1000 temperature oscillations between T_{lig} and T_{melt}. For low $T_{\mathrm{melt}} = 50$ °C the length distribution is short tailed and no clearly distinguished local minimum-maximum feature can be seen. For high $T_{\mathrm{melt}} = 58$ °C the concentration distribution is long-tailed and monotonously falling for long strands. The transition in between shows the local minimum-maximum feature in a temperature range of several K. The expected shift of L_{max} towards longer strands is hinted by the data, but the overall peak height is small and the detection limit cannot resolve further detail.

Although the experimental data shows the same behavior as the simulation, there is a noticeable variation in the overall appearance of the oligomer length distribution. The numerical model is for sure a simplification of the experimental setup, but was still able to predict the systems behavior. The distinctions of the experimental and the numerical simulation are responsible for differences in the oligomer distributions:

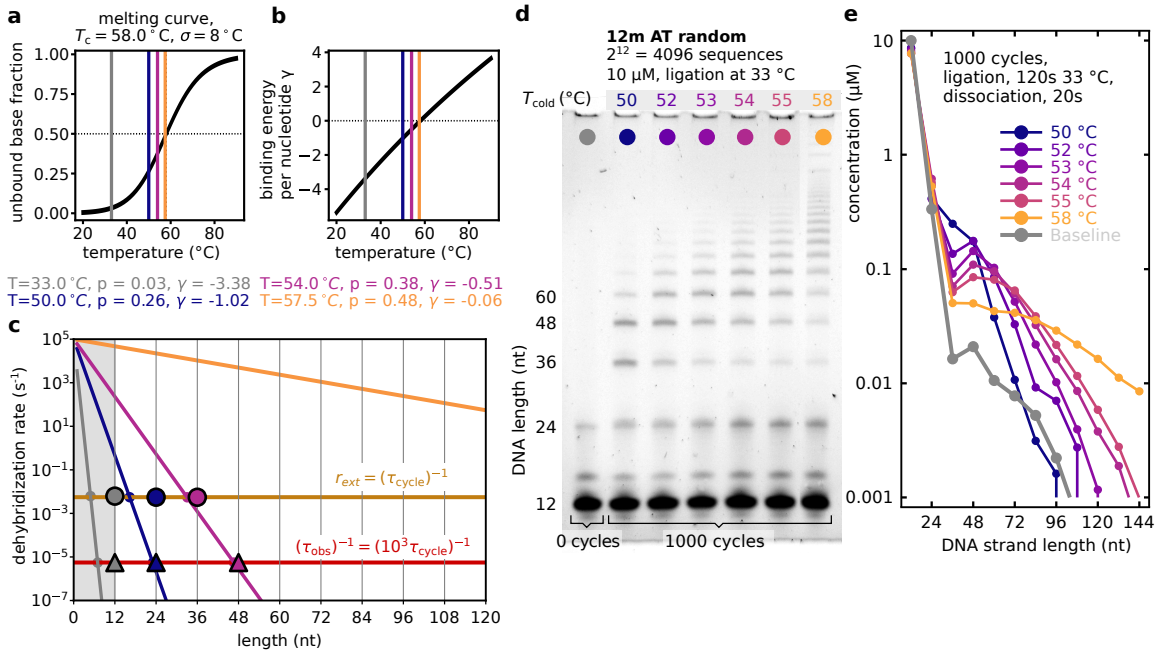

Figure 3.51 Prediction for the behavior of the experiment based in the binding energy:
This figure is adapted from ref. [104]. Utilized parameters: $k_{koll} = 1x10^5\ s^{-1}$, $\tau_{cycle} = 180$ s, $r_{ext}=(\tau_{cycle})^{-1} = 5.56x10^{-3}\ s^{-1}$, $\tau_{obs} = N_{cycle} \times \tau_{cycle} = 1.8x10^5$ s.
a, b Effective melting curve for AT-only short DNA oligomers with the melting temperature $T_{melt} = 58.0\,^\circ C$. The ligation temperature T_{lig} is marked in grey, the dissociation temperature for the experiment in purple to orange. The horizontal line in panel b marks the minimum effective binding energy per nucleotide, which also marks T_{melt}.
c Over the oligomer length, the effective dehybridization rate decreases exponentially. For low T_{lig} its slope is so steep, that the characteristic lengths L^* and $L^\dagger$ are smaller than the 12mer "monomer" building block in the experiment. Without temperature cycling the system is too cold and almost no ligation reactions occur. For temperature cycling between $T_{melt} = 50\,^\circ C$ and $T_{lig} = 33\,^\circ C$ the rounded values for the characteristic lengths fall onto the same length of 24mers and the resulting oligomer length distribution doesn't show the local minimum-maximum feature, but long strands do exist. Close to the critical temperature the slope of the dehybridization rate becomes more shallow and the feature should appear in the experiment. For even warmer T_{melt} the slope is so shallow, that the characteristic lengths are well above the detection limit of the concentration quantification and PAGE analysis and a long-tailed concentration distribution follows.
d, e PAGE gel and CQ-analysis of the 12mer AT-random sample with 1000 temperature cycles between T_{lig} and T_{melt}. In panel e the transition from a short-tailed length distribution to a long-tailed length distribution is clearly visible. The transition shows the local minimum-maximum feature described in detail in the numerical model. Temperatures are color-coded similar to panels a, b, and c.

- **Sequence space**
  A significant difference between the simulation and the experiment is the sequence space. In the experiment strands have a binary random sequence and thus has a sequence space of $2^{12} = 4096$ while the simulation employs a self-complementary model base `N` with a sequence space of 1. In the simulation the on-rate of complexes is solely governed by the kinetics of the strands, their concentration, and the amount of hybridization sites onto another complex. In the experiment, the substrate sequence has to be the reverse complement of the template, which decreases k_{on} in comparison. As already known from the sequence analysis in Section 3.3, Section 3.4, and Section 3.5, there is a strong A:T-fraction and sequence selection in oligomer product strands. Thus, on the timescale of the experiment, a significant amount of 12mer building blocks does not take part in the templated ligation reaction, neither as template nor substrate. In contrast, the actual reactive components in the simulation are all building blocks. This results in a "concentration offset" in the experiment, because unreactive strands still contribute to oligomer length distribution. Overall, the experiment has a time- and oligomer-length-dependent bias of the reaction material, that is difficult to estimate precisely.

- **Hairpin formation**
  There is no mechanism for self-folding implemented in the simulation. However, the sequence analysis in Section 3.3 and Section 3.7.3 strongly suggest that hairpin-formation is an essential mechanism in the experimental system. The simulation is still comparable though, because in the experiment oligomer strands segregate into A-type and T-type oligomers that inhibit hairpins. But the sequence analysis, especially Figure 3.4 and Figure 3.9 show, that the ligation of A-type and T-type oligomers does happen on a small scale. The resulting strands are then capable of forming hairpins which will significantly lower their extension rate r_{ext} as well as their templation capabilities. In terms of the simulation, this can be described as effectively reducing the ligation rate of that specific complex. The simulation does not account for a statistical length-dependent k_{lig}-decrease which thus results in a higher concentration of long oligomers.

- **Ligation rate and effective extension rate**
  The ligation rate of the ligase in the experiment is not easily accessible and even the manufacturer gives a rather impractical definition for the activity (see Section 5.5). While it is possible to estimate the ligation rate with the variation of the system sequence space (see Section 3.7), a possible variation for specific sequences or a length dependence is difficult to specify, as argued in the discussion section Section 4.6.

  In the simulation the ligation process is simply implemented as a length-independent rate k_{lig}. Consequently, all processes different to an effective ligation rate are neglected. As the experimental system has a very low degradation rate (because of the high stability of DNA, Section 5.1) the corresponding rate k_{out} is low. With the other two simulation parameters binding energy γ and building block concentration c_2 being well known, the ligation rate might be responsible for the arguably differently sized features of the oligomer length distribution in the experiment and the simulation (like in Figure 3.46). Figure 3.48 c shows the impact of different ligation rates: a small k_{lig} results in a small peak and steep slope for long oligomer products. A weakly established local minimum-maximum feature in the experiment can thus be due to a low effective ligation rate.

- **Open system vs closed system**
  The most obvious feature of the simulation system is its coupling to a building block reser-

voir. Together with the degradation or outflux rate k_{out} the system will at some point reach a steady state. In the experiment the system is closed and there is no exchange of particles with the environment. Thus, the systems state will change until all possible reactions occurred. With a substantial amount of comparably unreactive building block species (meaning, they might still ligate on a significantly longer timescale, compare [119]) the closed experimental system will not reach a steady, unreactive state in the observation time. The simulation can be adapted to reflect this behavior. Setting the outflux rate k_{out} to a value smaller than the observation time τ_{obs} inhibits long strands from leaving the system. And choosing to not set the building block concentration c_2 as a chemostated value, but only giving the initial amount of strands, simulates the lack of strand influx. In the

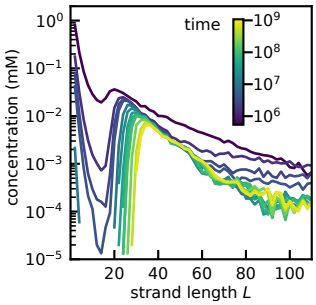

Figure 3.52 Transient evolution of the oligomer concentration in a closed system:
This figure is adapted from ref. [104].
In the closed system variant of the simulation model, there is no influx of new building block strands. With an increasing simulation time, the length distribution shifts to longer strands and the local minimum-maximum feature becomes more prominent. As the only base in the simulation is the self-complementary `N`, all building blocks can and will be incorporated into longer oligomers by the templated ligation reaction. The first plotted line for $\tau_{\text{obs}} \approx 10^6$ s has a shape similar to the experiment in Figure 3.51 e.

simulation shown in Figure 3.52 the existing monomers are all ligated to other complexes over time, as the concentration of short strands depletes completely. The local minimum-maximum feature remains, but L_{max} shifts towards longer oligomer lengths, as expected from the open system. This means, that the underlying processes of blunt-end fully hybridized complexes is still the driving factor for the emergence of the maximum. Comparing the closed system simulation to the experimental graphs in Figure 3.51e, the slope of the first plotted timepoint at $\approx 10^6$ s looks remarkably similar. The estimated τ_{obs} for 1000 temperature cycles is $\tau_{\text{obs}} = N_{\text{cycle}} \times \tau_{\text{cycle}}$=1.8x10^5 s as used in the prediction for the experiments in Figure 3.51, and therefore close to the $\approx 10^6$ s of the first plotted graph. The weakly developed local minimum-maximum feature might therefore suggest a short experimental time frame, with the feature not completely developed.

The difference between the experiment and the numerical simulation is presumably a superposition of multiple of the above discussed reasons. While the similarity of the closed system simulation and the experiment are remarkable and could explain most differences, the lack of very long strands in the experiment might be explained by the hairpin formation. Overall, the assumptions made in the simulation, especially

- the effective ligation rate k_{lig},
- neglecting sequence information and implementing hybridization by a mean binding energy γ and a factor Φ for the possible binding sites for a single self complementary base `N`,

- no explicit temperature cycling but an effective off-rate k_{off}

seem to be reasonable and match the experimental data well. The main focus of the numerical model is the analysis of the oligomer length distribution in random templated ligation. Thus its design, implementation and simplifications are chosen accordingly.

On one hand, the simulation lacks sequence information and can't model processes depending on the oligomers sequence, like the emergence of A-type and T-type sequence groups. Still, the chosen parameters can account for some of the implicated processes. On the other hand, the simulations provide insights into strand configurations and the elongation process, which is not at all accessible in the experiment and can only be estimated by reverse-engineering the sequence information. In Section 3.7.2 the de Bruijn network reveals that the A-type and T-type sequence groups (Section 3.3) are largely the reverse complement of each other. As templated ligation reactions need their substrates to be the reverse complement sequence of the template, this was no surprise. But extrapolating these findings to estimate complex formations is an educated guess at best. However, the abundance of fully hybridized double-stranded complexes found in the simulation (Figure 3.47b), either with blunt ends or with overhangs, explains the close correlation of the A-type and T-type networks. Even with the moderate temperature cycling conditions in the experiment longer strands apparently do not dissociate and in turn hybridized complexes of two strands of about similar length are abundant.

The simulation also identified critical parameters L^* and $L^\dagger$ that specify the onset of the extension cascades and the transition to the outflux-dominated regime (or the observation time, as shown above). These parameters internally depend on the inverse binding energy γ and thus on the system temperature. They can therefore be used to estimate the range of the double-stranded oligomer complex regime in the standard experiment as well. First measurements showing the local minimum-maximum feature might have been interpreted as incorrectly analyzed PAGE images and CQ. Only the simulation could verify that this feature was real and also explain the underlying mechanism. The parameters predicted in Figure 3.51c produced remarkably similar data as seen in the panel d, e of the same figure. But the simulation also predicts the experimental behavior for untested samples: with a monomer length of only 2 nt (`NN`) describing the intricate transition of complex types over the complex length (see Figure 3.47) and the detailed shifts in peak positions (see Figure 3.48), the simulation models all length %2 monomer strands, like the 12 nt AT-random monomers (see Figure 3.51). The values of the critical parameters L^* and $L^\dagger$ are simply rounded and still capable of describing the experimental system and its transitions regions.

The ligase seems to be still active in the experiment even after 1000 temperature cycles and about 50 hours experimental time, although it is difficult to quantify a possible degradation. Therefore, in light of the simulation in Figure 3.52, the experimental time was prolonged by another 1000 temperature cycles for a total experimental time of about 100 hours. For the three conditions with $T_{\text{melt}} = 52, 54, 56\,°\text{C}$ Figure 3.53 shows the PAGE gel and the CQ-analysis for the experiments. Already from the gel image in Figure 3.53 a the difference in the length distributions is visible. In their shape and slope the graphs are very similar to Figure 3.51 e, as expected, but the transient resolution with 0, 500, 100, 1500, and 2000 temperature cycles reveals the progression over time. Especially for the lower temperature, it takes more than 500 cycles for the local minimum-maximum to emerge, while for $T_{\text{melt}} = 56\,°\text{C}$ the feature is fully formed after the same time. In all of the shown experimental settings more temperature cycles correlate with a more shallow oligomer length distribution. For the very long-tailed distribution it becomes difficult to analyze the longest of strands, like in the second to last lane on the gel image, which

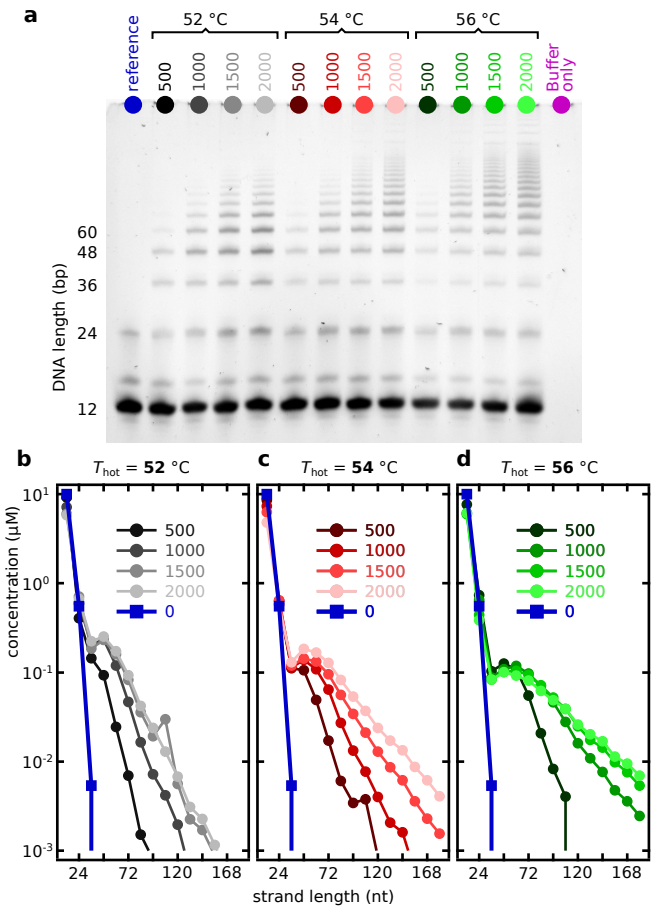

Figure 3.53 Extended experiment with 2000 temperature cycles:
This figure is adapted from ref. [104].
a PAGE gel for the three experiments. On the gel it's already possible to see a difference in concentration and length of the oligomer length distributions.
b CQ-analysis for $T_{melt} = 52\,°C$: the distribution is short-tailed overall but develops the local minimum-maximum feature for more than 500 temperature cycles. For the band at a length of 108 nt at 1500 temperature cycles an impurity on the gel inhibits the concentration quantification. Even though the difference between the graphs for the first three time points is seemingly larger, there is still a measurable increase in concentration for almost all oligomer lengths from 1500 to 2000 temperature cycles.
c CQ-analysis for $T_{melt} = 54\,°C$: here, the slopes of the oligomer length distribution become more shallow for an increased number of temperature cycles.
d CQ-analysis for $T_{melt} = 56\,°C$: has the most shallow slope with the most long-tailed distribution. For very long strand the CQ-analysis becomes difficult and results might introduce errors.

shifts the datapoints for the lane (compare Section 5.6). The analysis in the simulation is done for up to 10^9 s $\approx$ 32 years, which is obviously difficult to realize. Even an experimental time of over 100 hours is a stretch for the experimental setup. But for this case the already described reduced complexity sequence pools (Section 3.7) might be analyzed.

The end of Section 3.9.2 showed the predominant growth modes of all complexes in the simulation. L_{max} marks the transition length from a blunt-end to blunt-end primer extension growth mode to a primer-switching and template extension dominated growth resulting in similarly

sized new strands in the first, and longer strands in the latter case. The transition at L_{max} is at about 48 nt in this experimental system (see Figure 3.51). For $C_{\text{initial}} < L_{\text{max}}$ the final strand length is expected to be similar. Therefore, 24mer "dimer" reaction products are expected to prevalently facilitate the formation of another 24mer.

Because the Taq DNA ligase does not ligate 4 nt overhangs and shorter (see Section 5.5), the hybridization position of the substrate strands can be ± 4 nt from the 5'- or 3'-end of the template strand. A complex in the experimental system with such a shift is technically also a blunt-ended complex, as further extension is only possible by de- and re-hybridization of the two strands in the complex. Consequently, the sequence specificity is slightly reduced and allows for small shifts in the templated motif, which is also described in Figure 3.15 and Figure 3.16.

The uncovered basics of length-dependent growth modes is the basis of the random templated ligation simulation in Figure 3.17. At the ligation site a sequence pattern of `ATAT` is dominating, due to an abundance of the 3'-end sequence motif `AT` in 12mer monomers (see also Figure 3.12). By templated ligation, this self-complementary pattern is amplified and induced on the 5'-start of the downstream substrate, as long as the ligation site is somewhat located. This assumption is supported by the simulation here, that suggests that most product strands in the experiment (24mers, 36mers and 48mers, all with $C_{\text{initial}} \leq L_{\text{max}}$) are formed by the primer-extension mode. The resulting 2 nt motif analysis reproduced the experimental data, but lacks patterns, that were not simulated by design, like the before mentioned small motif shifts or the inhibition of self-complementary sequences (see Section 3.3).

A preference for a blunt-end to blunt-end extension might seem logical simply by analyzing the most common sequence motifs, as done in Figure 3.19 and Figure 3.28. But this is merely the result and only an indicator to the dominant growth mode. In the experimental system only the analysis of the initial state and the final state is possible. The basic reactions of the particle- or complex-based simulation outlined in Figure 3.45 are inherently based on the formation and dissociation of complexes.

4 Discussion

Some conclusions in the results-Section 3.9.3 could be drawn by experiment or simulation only. However, especially the last section pointed out, that other conclusions can only be drawn by analyzing theory and experiment simultaneously. As seen from a chronological view on this study, the simulation found new insights that provided more conclusions for previous experiments discussed in the prior sections and arguments. This discussion takes up points that were not discussed in detail in the results and combines them with findings accessible up until here.

4.1 A-type and T-type sequence entropy

The first major point, and a prerequisite for the whole initial argumentation, is the reduction of entropy of product oligomer strands. Importantly, this reduction must be caused by an underlying selection mechanism and not simply be the result of less, but equally random sequence strands. In Section 3.2 the reduction of entropy is shown for 60mer oligomers. With the later results of the A:T-composition and the high similarity of strands, the relative entropy reduction in Figure 3.1 appears smaller than expected.

As quickly recognized in Section 3.3, the analyzed 60mers are predominantly made from clearly distinguishable A-type and T-type strands. The known A:T-ratio can be used as an input to calculate the entropy of a set of random binary sequence strands. In contrast to the results section, where the maximum entropy was the entropy of the 12mer "monomer" pool as the upper limit, the maximum entropy of a strand with a given A:T-ratio can simply be calculated by:

$$E_{\max} = (1 - \text{ A:T-ratio }) * \ln(1 - \text{ A:T-ratio }) + (\text{ A:T-ratio }) * \ln(\text{ A:T-ratio }). \tag{4.1}$$

Figure 3.4 indicates the centers of the bimodal distribution to be at about 0.72. Comparing the A-type and T-type strands to a set of random strands with a set base ratio implicates the already shown entropy reduction by selection. But Figure 4.1 shows, that dependent on the analyzed position the single-bases in the strands have a higher or lower entropy than expected. However, setting the maximum relative entropy to the value describing the randomness of a strand with a set base-ratio (meaning normalizing the data to the entropy expected for the set A:T-ratio) shifts the data on the y-axis of Figure 4.1 and highlights the location-dependence of the entropy reduction in oligomer strands. On a strand with a given A:T-ratio the frequency of sequence motifs is location-independent. But the abundant ligation site sequence pattern `ATAT` in the emerging oligomers is located and occurs significantly more often than expected for strands with 70:30 A:T-ratio. Similarly for the inter-ligation site regions: poly-`A` and poly-`T` motifs are more abundant than expected. Here, the entropy in the inter-ligation site regions drops to 0.3, describing a low variance of local sequence motifs. The T-type strands have a less pronounced entropy reduction in the same region. The analysis of the 2 nt sequence motifs in the initial 12mer pool and longer oligomers (see Section 3.3, Figure 3.17, and Figure 3.7) already reveal the lack of the specific motif `TT`. As the inter-ligation site regions are predominantly made of poly-base motifs, there are significantly less fitting 12mer T-type building blocks and thus less assembled T-type oligomers (see below).

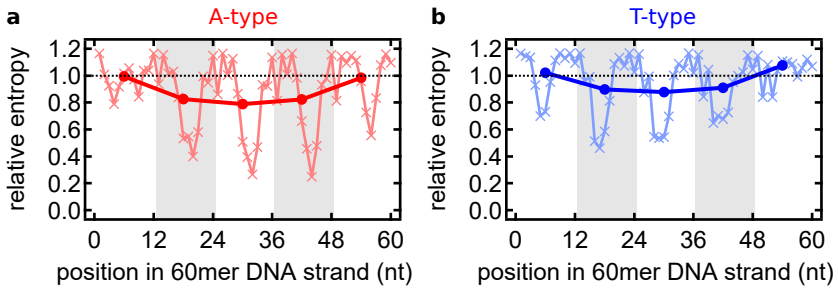

Figure 4.1 Reduction in sequence entropy for A-type and T-type oligomer products:
For both groups the maximum entropy was calculated for an A:T-ratio of 0.72 (black horizontal line). Analyzed strands include only strands with more than 60 % of the analyzed base-type group (so strands with A:T 0.4 to 0.6 were excluded). Grey vertical bars are guides to the eye and mark individual 12mer subsequences.
a A-type 12mer subsequences in the center of 60mer strands have a lower entropy than random sequences with an A:T ratio of 0.72. The single positions show, that the ligation sites are significantly more random, while the inter-ligation site locations are less random than expected for the given A:T-ratio.
b T-type strands show a comparable pattern: the start and end sequences are more random than expected, while the center subsequences have a lower entropy. Poly-`T` motifs are less abundant compared to the poly-`A` motifs in A-type strands though, explaining the overall higher entropy.

The larger entropy of the 5'- and 3'-end subsequences can partially be explained by the analysis done in Figure 3.19. Subsequences at the start and end of an oligomer are different compared to the center subsequences, but very similar to themselves. Those motifs are less structured, as they lack the specific selection for poly-`A` poly-`T` as well as the alternating pattern on the upstream (5'-end subsequences) or downstream (3'-end subsequences) end.

4.2 A:T-ratio of oligomer products

The most dominant effect causing the reduction in sequence entropy is the subdivision of strands in either A-type or T-type strands. The A:T-ratio is not only overall at about 70:30 and 30:70 in the respective groups (see Figure 3.4), but the 12mer subsequences are also very likely to be of the same base-type themselves (see Figure 3.9). In the results section it is argued, that this is mainly due to the inhibition of hairpin formation of emerging oligomer strands. A strand with about 50 % of each base has a high probability to include reverse complement sequence motifs that facilitate the hybridization of the strand to itself. But a folded strand does not take part in the templated ligation reaction, not as a substrate and not as a template.

By extending the templated ligation simulation of Tkachenko and Maslov [119] to include an energy based self-folding mechanism, the A:T-ratio of emerging strands was analyzed (see Figure 3.3). Even though the simulation was started with an initial feeding of 24mer strands with a binomial A:T-distribution as templates, the emerging 24mers show a clear segregation into A-type and T-type strands. The simulation model only assumes a single paired base per substrate strand for minimal interference of the templated sequence motifs. The resulting bimodal distribution is remarkably similar to the experimental data. This strongly suggests, that the inhibition of hairpin strands is indeed the driving factor of this subdivision.

For longer oligomers the experimental data finds an abundance of A-type strand in comparison to T-type strands (see Figure 3.4). As argued in Section 4.1 the reason is likely the lack of the 2 nt sequence motif `TT` in the building block "monomer" strands. The solid state synthesis of the random sequence strands seems to be biased against `T`-`T` connections. Later analysis of the predominant growth mode (see Section 4.3) and the subsequent structure of emerging oligomers (see Section 3.7.2) emphasizes the similarity of the A-type and T-type groups as the respective hybridization partners of each other. The network stresses, that abundant T-type strands are the exact reverse complement of abundant A-type strands. A-type strands include a dominant poly-`A` motif at inter-ligation sites. Thus, the lack of the `TT` motif in the underlying building blocks inhibits the formation of "correct" T-type strands. The overall lack of T-type strands in the initial 12mers is additionally impairing the elongation by ligation. The abundance of A-type strands is becoming stronger starting from 48mers and longer. The kinetic simulation in Section 3.9 describing the advances of ref. [104] suggests, that double-stranded complexes of length $L_{\max}$ and longer are unlikely to dissociate. Therefore, long A-type oligomers bind the less abundant T-type strands in persistent double-stranded complexes.

The x1-sequence space sample in Section 3.7.3 is a sample without a possibility to suppress the hairpin formation by shifting the A:T-ratio in oligomer strands. The entire pool is made from the same strand with a self-complementary motif: `AAATTTAAATTT`. While 12mers seem too short for the formation of hairpin structures, the emerging longer strands are wholly capable of folding on themselves. Actually, those strands have no choice for complex formation, other than hybridizing to another single strand. The high local concentration due to the hairpin mechanism (see Section 5.2) clearly favors the hairpin-case. The resulting length distribution in Figure 3.35 shows a short tailed length distribution and a bias towards odd-multiples of 12mers (36mers, 60mers, 84mers, *etc.*). The hairpins clearly inhibit the formation of long oligomers by blocking template and longer substrate strands from taking part in the reaction. 12mer strands incapably of forming hairpins are therefore the most active substrate and template species, but elongation typically stops for 24mers already. This effect is even stronger for the GC-`NN` sample that results in even more-compact length distributions and almost all 12mers are ligated at 24mers due to the higher binding energy of the `G`-`C` pair.

The sequence structure of oligomers implies the incorporation of specific strands by the templated ligation reaction. In Figure 3.5 the 12mer building block "monomer" pool is analyzed before and after 1000 temperature cycles. The normalized A:T-ratio graphs in yellow (0 cycles) and black (1000 cycles) mainly differ in three regions: A-type and T-type strands with A:T-ratios of roughly 70:30 and 30:70 are less abundant after 1000 cycles, while strands with 5:7, 6:6, and 7:5 bases A:T are significantly more abundant. Consequently, A-type and T-type strands are consumed in the reaction to build oligomers, and 50:50 strands remain in the 12mer pool. Although this is expected in a closed system, the sequence analysis provides further confirmation.

Tkachenko and Maslov showed before, that a system with an extended set of sequences undergoes a non-trivial selection mechanism and results in a reduction of the sequence space of N^2 to simply $2N$, with N monomer-types [119]. Figure 2 therein features an entropy over log-time graph, that describes a two-stage entropy reduction. While the second stage is due to the active competition of strands for the correct ligation site sequences, the first stage entropy reduction is caused by slightly different ligation rates λ_{ij}, depending on the ligation site sequence i and j of substrate and template strands. In the present experiment, the formation of oligomers from 50:50 A:T-ratio strands is significantly slower and less efficient than for 70:30 A:T-ratio strands because of the hairpin-inhibition. However, the simulation of Tkachenko and Maslov suggests, that in the absence of a degradation mechanism the system-entropy might increase again for ex-

ponentially longer times. The depletion of A-type and T-type building block strands reduces the effective extension rate for the presented extension mode. And even though unlikely on a scale of tens of hours, the abundant and hairpin-prone 50:50 A:T-ratio 12mer monomers might produce a larger set of 24mers. The second entropy reduction predicted by Tkachenko and Maslov is probably not yet visible in the entropy reduction analysis here. The experiment in this study includes more selection mechanisms than the two-stage entropy reduction simulation, like the autocatalytic emergence of ligation site sequences. This could lead to a different long-time entropy reduction behavior.

4.3 Oligomer growth modes

For longer oligomers the theoretical analysis in Figure 3.8 differs from the experimental data by the width of the peaks. The A:T-ratio in long oligomers becomes less divers and only a small number of different A-type strands with the correct A:T-ratio are incorporated, as shown by the network analysis (Figure 3.28), the entropy reduction (Figure 3.1), and the abundance of self-similar motifs (Figure 3.19). This suggests an additional selection mechanism other than the inhibition of hairpins.

Even by simply plotting the abundance of each base on each position in oligomer strands like in Figure 3.12 two patterns become obvious: Strands seem to be made from alternating `ATAT` motifs separated by poly-`A` or poly-`T` motifs. The poly-base motifs seems to predominantly find their origin in the inhibition of hairpin formation in oligomer strands. The emerging alternating `ATAT` motif can however not be explained by this mechanism.

An analysis of this property is only possible in the combination of the kinetic simulation of random templated ligation described in Section 3.9 and ref. [104], and the sequencing data obtained for the initial 12mer AT-random sequence pool in Figure 3.7. The solid state synthesis of the supposedly random sequence 12mers seems to feature two distinct biases: as described above, `TT` motifs are less abundant than expected. Additionally, Figure 3.7a shows, that the last 2 nt position has a higher frequency for the motif `AT`. Starting for the 24mer oligomer strands, position 11 has a higher abundance of the same motif `AT`. The very first 2 nt motif in 12mer strands has no obvious bias, other than a slight lack of `AA`. Without further knowledge of the exact extension mechanism, there is no explanation for the emergence of the ligation site `ATAT` motif.

The kinetic simulation, despite not including a detailed sequence description of strands, uncovered three distinct growth modes for oligomers (see Figure 3.49 and Figure 4.1). Depending on the length of an oligomer, one growth mode is more likely than the others, as shown in Figure 3.50. For strands smaller than $L_{\max}$, a complex tends to template a complex of similar total length. In the `NN` 2 nt "monomer"-simulation model offers a higher resolution of the length-dependent behavior compared to the 12mer "monomers" experiment. This length selectivity suggests the predominant growth mode in short strands is the "primer-extension"-mode, where a longer template strand binds two 12mers for ligation. In the experiment, this limits the hybridization positions and therefore the location of the ligation site. The simulation discussed in Figure 3.17 is based on this localization of the ligation site on the template strand. Because the motif `AT` is its own reverse complement and because strands hybridize in 5'-3' to 3'-5' opposite directions, template strands with this `ATAT` motif induce the same motif on substrates at the ligation site: strands with a 3'-`AT` end and a 5'-`AT` start motif.

The sequence structure of emerging oligomers originates in the combination of a slight bias in the pool of building blocks and the inhibition of the mechanism that is harmful to the ex-

tension mechanism (template inhibition caused by hairpins in templated ligation) without any other input. For a critical review of the extension of random strands it is essential to also analyze a potential influence of the ligation process itself. As briefly explained in the introduction Section 2.2 Taq DNA ligase, an evolved enzyme that repairs nicked dsDNA in its original host the bacterium *Thermus thermophilus* (see Section 5.5), is chosen as a "model ligation mechanism" to mimic a non-enzymatic chemical ligation. With the ligase's native application in mind, it seems likely that the ligase should only ligate correctly paired bases, especially at the ligation site. In Section 5.5 this is shown to be true. The Taq DNA ligase is very sensitive to incorrectly paired bases at the ligation site and the ± 1 position. With a similar line of reason it seems unlikely, that there would be a pronounced bias for the ligation of some sequence motifs. However, studies from Lohman *et al.* [78] and Kim and Mrksich [70] suggest, that a slight bias might stall the ligation of templated by `AA`. But the results of the simulation in Figure 3.17h do not suggest the collapse of the ligation site motif abundance even in this worst case scenario.

Strands can be extended in both, the 3'-5' downstream and 5'-3' upstream direction. Together with the strong localization of the ligation site on the template strands, center subsequences are more restricted in their sequence motifs (lower entropy, see Figure 3.1) and the ligation site shifts (lower probability for ligation site poly-base motifs in center regions, see Figure 3.16). The sPCC matrix of 84mers vs 72mers (see Figure 3.19) suggests, that start- and end-subsequences do not experience a similar selection pressure. Comparing 36mer A-type strands to 36mer T-type strands revealed, as expected, no correlation. But the comparison of the A-types with the inverse complements of the T-types correlates significantly, as seen in Figure 3.20. Therefore, start-subsequences in A-type strands (of all oligomer lengths) have a distribution with a common set of motifs. This set of subsequences is found as the ideal binding partner with the reverse complement sequence in T-types strands (of all oligomer lengths). But other than predominantly being a subsequence with an A:T-ratio similar to the full strand, there is no clear motif-selection for start- and end-subsequences, as shown by the lack of sequence entropy reduction in Figure 4.1.

For long strands, again starting for length $\leq L_{\text{max}}$, long hairpin sections become more abundant, despite the selection pressure for shorter strands inhibiting self-folding. Here, the predominant extension modes are the "primer-template switching" and "template extension" modes due to the complex length. While the "primer extension" mode and the high probability of $C_{\text{initial}} = C_{\text{final}}$ for short oligomers effectively requires that most emerging short complexes were (almost) entirely hybridized to the template, the extension modes for longer oligomers include single-stranded sections. The unbound sections do not experience the selection pressure of templated ligation as before. A single-stranded section in a complex can have an arbitrary sequence. By that mechanism long strands with a combination of A-type and T-type sections can emerge due to "primer-template switching" or the "template extension" that is assisted by a "helper-template" strand (see Figure 3.49). The analysis in Figure 3.9 shows that long oligomers might feature an internal change in base-type, while switching back and forth between A-type and T-type subsequences is very rare. As discussed in the introduction, the mechanism of oligomer self-folding is deemed essential on the way to catalytically active complexes. And even though the templated ligation inhibits self-folding in short and even most long oligomer products, there are long, highly specific hairpin structures emerging from the random sequence pool. Those complexes could help in the formation of catalytically active oligomer-self-folding structures.

In Section 3.7.2 the ligation performance of three pools, each made from only eight strands, was compared. While the strands selected by the interconnectivity in long oligomer products (`Network`-sample) yield most oligomer strands especially for short and medium-lengths, the

Replicator-sample seems well optimized for very long strands (see Figure 3.29). The analysis of the resulting subsequence-sequence network by NGS in Figure 3.33 suggests, that the Network-sample grows by a $C_{\text{initial}} = C_{\text{final}}$ blunt-end to blunt-end complex mode, comparable to the AT-random 12mer sample. This is the result of the selection in the AT-random experiment, as the selected sequences in this Network-sample virtually condense all underlying selection-effects down to a set of eight strands. This apparently also holds true for the growth mode.

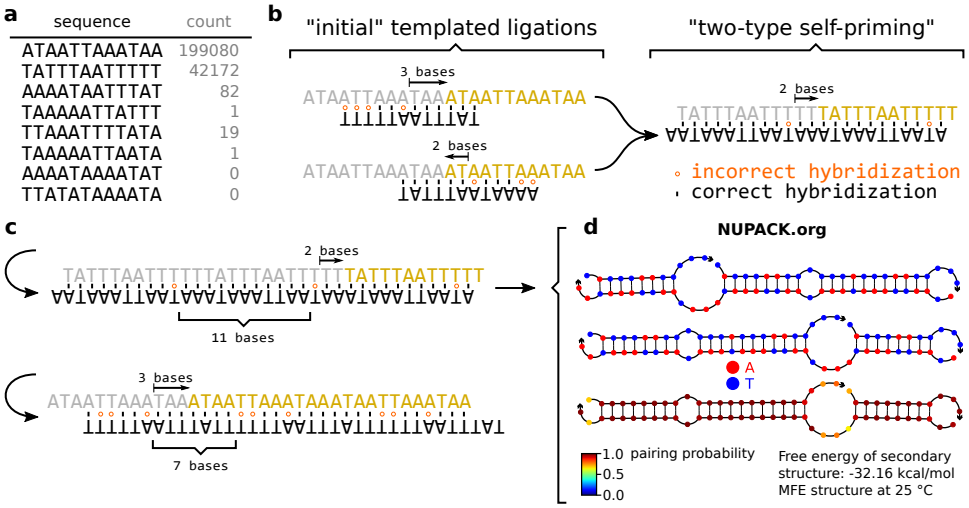

Figure 4.2 Random**-sample hypothetical extension mode:**
a Eight randomly chosen sequences make up the Random-sample. The NGS data shows, that one sequence is significantly more abundant than all others, as already seen in Figure 3.33b.
b Possible double-stranded conformations with strands from panel a. Incorrectly paired bases in the complex are marked in orange. The top complex has three bases from the ligation site until the first incorrect basepair, the bottom complex only two. Although those complexes are rare and not found by the NUPACK.org tool, they will emerge with a small chance regardless. Even though the Taq DNA ligase is very specific for sequence errors on the ligation site (see Section 5.5), an error at ±2 or 3 might still be ligated at a (very) low rate.
c Subsequent elongation in longer complexes also have incorrectly paired bases.
d Due to the length of the complexes, the binding energy is larger and the NUPACK.org analysis tool finds the hypothesized conformations at 25 °C.

On the other hand, the Replicator-sample was originally designed to also facilitate "primer extension" with $C_{\text{initial}} = C_{\text{final}}$ complexes and allows for the assumed initial three-strand complexes with overhangs for each substrate. As odd-parity extension cascades were identified as the primary source of very long complexes in the kinetic simulation (Section 3.9 and Figure 3.47) it is likely that this type of complex is also the predominant driving force in the extension of long Replicator-strands. Due to the subsequence structure of the AT-random sample this mechanism is missing in the Network-sample, which only includes perfect reverse complements with a very low inter-templation capability of the strands. The sequence design of the Replicator-sample enables the formation of templates with an offset of 6 nt. Those complexes are the odd-parity equivalent extension cascade complexes in the x8-samples context and facilitate strong growth of long strands. The comparably more frequent dissociation of the shorter Network-sample complexes enables a better short-complex templation caused by the "primer extension" growth mode with substrate strands mostly resulting in similarly sized 24mer and 36mer strands.

Any Random-sample made from eight random AT-only strands would be unlikely to form ligation product strands. The set of eight strands in the Random-sample here do lead to the emergence

of longer strands though, as seen in Figure 3.29a and Figure 3.33b. Employing NUPACK.org, a common tool for gauging temperature-dependent melting behavior and hybridization conformation in sets of RNA and DNA strands, the algorithm does not find structures that would be ligatable by the Taq ligase. Despite that, there must have been ligatable complexes present in the experiment for the emerging oligomers. The reason why NUPACK does not find suitable complex-conformations (at least in more than 99.99999 % of analyzed complexes at $T_{\mathrm{lig}} = 33\,^{\circ}\mathrm{C}$) might be caused by the algorithm itself. The software analyzes the equilibrium conformations at a set temperature and gives the corresponding probability of each basepair in each complex to be hybridized. However, in the experiment all types of complexes will form just by random chance. And even though the half-life time will be very short, there is a low but non-negligible chance of such a complex being ligated, simply because of the abundance of strands and complexes due to the high relative concentration in the small sequence space. Figure 4.2a lists the eight sequences and marks how often they were sequenced as 12mer subsequences in 84mer oligomer product strands. Panel b gives two examples of (partially) double-stranded complexes that might be ligated. The Taq DNA ligase is very sensitive to mismatches at the ligation site, as discussed in Section 5.5. Further away from the ligation site, at $\pm$ 2 or 3 bases, the ligation rate might still be low, but some complexes will inevitably be ligated. And as for the AT-random sample and the extension cascades in the simulation, longer strands are better suited as template strands for further ligation reactions. Figure 4.2d shows, that those longer complexes are also found by NUPACK due to their higher binding energy, that scales with the length of the complex. The random choice of eight strands might have been "lucky" in this set of eight strands. Different sets of eight strands selected from the set of 4096 possible sequences in AT-only 12mers might show a more or less functional templation behavior than the `Random`-sample. However, in comparison to the **`Replicator`**-sample and `Network`-sample the ligation rate is distinctly worse, which supports the original idea of the experiment: Subsequence motifs common in oligomers emerging from the random sequence unspecific selection experiment are more viable in the given extension scenario than a set of random strands.

As seen in the hypothetical extension mode, complexes might include dangling ends but weren't referred to as odd-parity elongator or extension cascade strands. While the simulation can ligate strands of arbitrary length with the implementation of the ligation process as an elementary system rate, the Taq DNA ligase has limitations. The manufacturer states, that the ligase does not ligate 4 nt overhangs (see Section 5.5). This leads to quasi-blunt-ended complexes in the experiment. Overhangs of 1, 2, 3, and at least 4 nt are not extended by any of the extension modes. The shortest building blocks that resulted in successful ligation reactions are 9mer GC-random strands with the smallest possible complexes including at least a 4 nt *and* a 5 nt overhang. This possible variation in the hybridization position of substrate strands on template strands enables the variation seen in the ligation site sequence motifs and the subsequence networks. It also indicates a smooth transition from one extension mode to another (at least with building blocks with a length of 2 nt). The "template primer switching" mode with a copy site length l_{cs} of 7 nt might result in a "blunt-ended complex" at elevated temperatures. The effective extension mode would be by "template extension" with a final overhang of 5 nt. But the complex might be elongated by both the "template extension" and the "template primer switching" mode at lower temperatures. The result is likely the smooth transition of different complex types shown in the simulation (Figure 3.47). The implemented simplifications in the simulation model are apparently still accurate enough for a detailed analysis of a model that includes sequence information and longer strands, by choosing appropriate system parameters like ligation rate, temperature and binding energy.

4.4 Lower complexity samples as model for whole system

In the introduction to Section 3.7.1 it is argued, that a reduced complexity sequence space might still result in a good approximation of the complete 12mer AT-random system. The reduction of the sequence space down to only eight strands in Section 3.7.2 as well as the simplified numerical simulation included only the very elementary properties offered by the DNA strands: hybridization, dehybridization and ligation (here by a ligase). Together with the data for GC-only samples of different lengths that were also analyzed by NGS data, the overall system dynamics and subsequence correlations are about similar.

- Strands build two groups of reverse complementary sequence types and longer oligomers are predominantly made from very similar subsequences and a homogeneous subsequence base-type.
- The similarity of the subsequence motifs are dependent on the position in the strands, but overall strands of different lengths have similar patterns.
- Ligation site sequences favor self-complementary sequence motifs.
- Strand extension follows three distinct modes that can be recognized in specialized samples (like the x8-samples).
- The system rates are closely related and depend on the size of complexes.

The reduced complexity sequence space samples x64 "double bases", the x8-samples, and the dynamics of the templated ligation system as analyzed by PAGE, CQ-tool, and simulation are good approximations of the AT-random sample. The underlying mechanisms and effects seem to be retained and the measured parameters like the sPCC matrix (see Figure 3.19 and Figure 3.30) are reproduced, even for significantly smaller systems.

However, the introduction describes how the presented system of a random sequence 2-letter alphabet DNA ligated by an evolved enzyme is only a model system. Concerning the main question of "what happened at the Origin of Life" or even more generally "what could happen at *AN* Origin of Life" this experiment and analysis gives new insights. There are discussions and studies about how an early oligomer could have been made from only two bases that allow hybridization [27, 61]. In this case, the model would be an excellent approximation and its implications could be included to a great extend. At the same time, a system made from only two bases is very limited in the amount and structure that complexes can form. And as known from modern evolved life, four bases are necessary for encoding the known amount of amino acids in 3 nt long stretches, which can only be done with a sequence space of at least 4^3. So at some point, the two-bases system must have been extended by another basepair. A second option would be similarly likely: organic chemistry for the synthesis of nucleobases suggests, that a wide variety of comparable structures emerged at the same time with similar yields as the known nucleobases. *THE* oligomer responsible for first selection and function is unknown and could have employed a variety of less specific basepairing between other kinds of nucleobases.

In both cases, this study can draw conclusions for said systems. The segregation into reverse complementary A-type and T-type groups is an effect of hairpin prevention. For the argument of templated chemical ligation at the Origin of Life a more complex system will likely undergo a similar segregation. Longer strands are better templates, can store more information and are more likely to form intricate secondary structure with other strands, while accumulation effects in hydrothermal microenvironemts also favor longer strands. A folded structure is inferior in comparison and would lose in a competitive setting.

The emergence of ligation site `ATAT` sequence patterns are due to small variations in the original pool providing the building blocks for the strand extension. In the experiment, this

is probably due to slightly inhomogeneous sequence-dependent polymerization rates during strand synthesis. However, due to the varying abundance and reactivity of educt substances on Early Earth, some bases and base-derivatives might have been significantly more abundant than others. This implies a random sequence pool that might include a significant sequence- or motif-bias [27].

The ligation process might also be biased and favor specific sequences, again likely depending on the exact chemistry. The model system shows how such biases can sustain other selection mechanisms, help in localizing the ligation sites to optimize for certain ligation site sequences, while not inhibiting the suppression of self-folding or the selection of a reverse complement group of strands. The underlying mechanism uncovered by the model is not exclusive for DNA. Other systems made from RNA, proteins or pre-RNA will likely experience similar selection of hybridization-features and sustain biased sequence pool information for several generations.

4.5 System ligation rate as function of system state

The numerical simulation makes a clear distinction between the ligation rate of a complex k_{lig} and the rate of complex extension r_{ext}. Similarly, the experiment has a strand ligation rate that depends on the ligase and system temperature. In both cases, the actual extension rate of strands depends on more parameters. Figure 3.45 explains, that r_{ext} is an effective rate that depends on the elementary system rates k_{on}, k_{off}, and k_{lig} (see Section 3.9.3 as well). However, the hybridization on- and off-rates depend on the binding energy, which in turn depends on the length of the double-stranded sections in a complex. A longer strand with a significantly lower k_{off} will have a different r_{ext} compared to a short strand. In Figure 3.42, the oligomer product concentration is analyzed in a time series of the AT-random 12mer sample. Panel c shows that the yield of strands with a certain length changes over the course of the experiment. For early points in

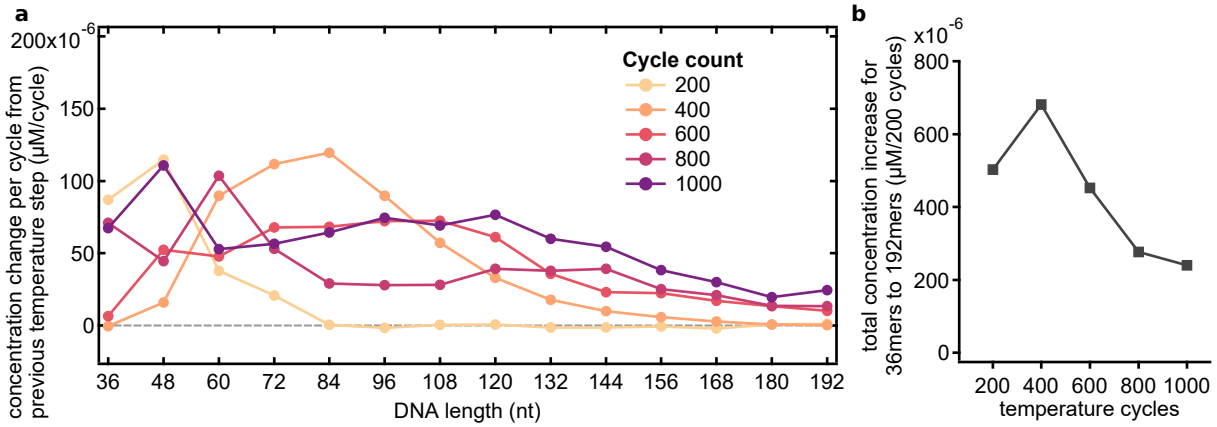

Figure 4.3 AT-random 12mer oligomer production rate:

a Reproduction of Figure 3.42 panel c. For each lane with an additional 200 temperature cycles the concentration was estimated with the CQ-tool. The rate is then the difference between the concentration before that timestep, and the present timestep, binned for 200 cycles.

b Adding all emerging oligomer product strands starting from 36mers to 192mers gives a cycle-count- or time-resolution of the system oligomer production rate with a resolution of 200 temperature cycles. The largest yield occurs between 200 and 400 temperature cycles and then decreases to about a third for 800 to 1000 cycles.

time predominantly short strands emerge, while at later time points more long strands are produced. Figure 4.3b adds all oligomers with lengths between 36mers and 192mers and plots the overall production rate per 200 temperature cycles as a function of experimental time or cycle count. Between 200 and 400 cycles the most amount of strands are emerging and the production rates drops of for more temperature cycles. Also, the initial step from 0 to 200 cycles yields less oligomers than from 200 to 400 cycles. This might be caused by the initial lack of longer template strands that catalyze the formation of more long strands.

The analysis of complex-types as a function of length (see Figure 3.47) suggests that oligomers with different lengths have different effective extension rates. Complexes either enable elongation (single-stranded complexes), inhibit elongation (double-stranded blunt-ended complexes) or catalyze elongation (double-stranded odd-parity type strands and long single-stranded template strands) as so-called extension cascades. These complex types correlate with the complex length C_{initial}. As the dissociation rate of a complex also correlates with C_{initial}, the effective extension rate changes significantly as a function of the complex size. Complexes undergo a transition from a hybridization-dehybridization dominated regime to a dissociation and r_{out} dominated regime when being elongated. With the transition from abundant fully hybridized complexes to abundant odd-parity complexes these double-stranded sections are finally very unlikely to dissociate at the standard $T_{\text{melt}} = 75\,^{\circ}\text{C}$ (see Section 3.8.1). Higher T_{melt} shift the length distribution considerably towards shorter strands. With a constant k_{lig} the change in the oligomer distribution is due to fewer ligatable complexes, despite virtually similar conditions at the ligation step.

The system ligation rate or accumulative complex elongation rate is therefore dependent on the amount and the length-distribution of oligomers, as well as their hybridization conformation. The ligation ligation rate k_{lig} increases with the activity of the ligase with the temperature (to a certain point) while r_{ext} decreases for short strands (dissociation of complexes) and increases for long complexes (no dissociation of double-stranded complexes). The inhomogeneous relations of elementary system processes on different parameters led to a non-linear system behavior. Despite the arguably drastic simplifications of the experimental model system and the numerical simulation, the resulting system behavior is more complicated and intricate than one might expect.

4.6 Implication of the results for the Origin of Life

The field of the Origin of Life evolved over time with new ideas and results emerging. In some parts the discussion about specific circumstances at the Origin of Life are ongoing: research groups try to find realistic chemical pathways for synthesizing the very building blocks of RNA, DNA, and proteins from simple molecules. Other scientist are on the hunt for pre-RNA structures and analyze their hybridization behavior. Another ongoing debate concerns the RNA-world in contrast to other complete descriptions for the emergence of life, such as by the metabolism (metabolism-world), or peptides, *etc.*.

But there is less discussion about the available mechanisms during Early Earth: The foundations of chemical reactions as well as physical and geological properties are comparably well known. There were no enzymes, catalysis by single molecules was very limited and physical mechanisms rarely included more than basic effects other than temperature, pH, and the internal properties of early oligomers like hybridization or stacking interactions.

For oligomers, the most prevalent assumption is that they need to form some kind of binding interactions for their elongation, stability against hydrolysis, function (like catalytic activ-

ity), or information storage. The elongation of those early oligomers might have been due to non-templated polymerization with activated nucleotides, to at least some point. Even in short strands, elongation by the much more efficient templated polymerization and templated ligation did likely take over. And whenever templation happened, there was a selection against strands that cannot hybridize due to their conformation, base-sequence, or overall structure.

In this study, the selection pressure that leads to selected oligomer products is applied by the elongation process. Longer strands did undergo a selection for being a suitable substrate or template strand. The emerging strands showed several effects, that can be extrapolated to a real-world Origin of Life setting. Oligomers typically inhibit self-folding by an abundance of a certain base. Oligomers that fold on themselves are unlikely to be part of subsequent ligation reactions. In the experiment, the two groups A-type and T-type emerge. On Early Earth or in an ATGC-random pool, this effect could be present as well. Strands might predominantly include two bases that do not bind, or even result in more than two groups. The analysis uncovered that the A-type and T-type group, while not binding to themselves, are remarkably accurate reverse complements of each other. A-type strands can template T-type strands and *vice versa*. Thus, this subdivision does not only retain the ability of strands to form double-stranded complexes, but reinforces this process by allowing very specific and located hybridization of strands. The sPCC analysis suggests, that strands also conform to a subsequence-sequence structure that is independent of the strand length. On Early Earth, emerging longer strands are applying a selection pressure on shorter strands for sequence and structure on the scale of the oligomer. The result might be the growth of comparable self-similar oligomers. As argued above, the initial pool is unlikely to be uniformly random with some bases or base-predecessors more abundant than others. This bias must leave a trace in the emerging oligomers. The analysis and simulation in Figure 3.17 shows how even a small bias towards a certain sequence motif can be sustained during elongation and produce clear sequence motifs in long oligomers. In the case of the self-complementary motif `AT` this bias is even amplified and helps with the localization of strands during hybridization. While the exact dynamics of the system depend on the DNA, the ligase, and strand concentrations, the behavior of the studied system could be extrapolated. The elementary rates of k_{lig}, k_{on}, and k_{off} determine the emergence of oligomers and the shape of the length distribution in the experiment. The simulation uncovered that this combination of elementary rates with their non-linear change over temperature and complex-size give rise to different ligation-driven extension modes and oligomer-length distributions.

During the Origin of Life on earth, at some point an elongation of strands has happened. Due to the nature of oligomers and the available elongation process (requiring double-stranded complexes) this study suggests that strands emerged in a non-linear dynamics fashion. Oligomers will form complexes with different properties depending on the length and the hybridized position of the individual strands. The hybridization, driven by reverse complement sequences on both strands, leads to selection of motifs simply by the templation itself. Selected sequences can interact with each other or themselves leading to additional effects, like the subdivision in groups of self-similar strands. By this selection for elongation, oligomers are optimized for the given experimental or Early Earth setting. Oligomers are better templates due to their length than shorter strands and pass their sequence information on to newly formed oligomers, catalyzing the formation of their own binding partners. The important mechanism of hybridization and basepairing is retained and even enforced by the sequence selection in the emerging oligomers. Early oligomers are likely subjected to the mechanism discovered here, which allow an elongation and coupled reduction in sequence space while improving hybridization abilities

of the complexes. In subsequent evolutionary processes, those long, specifically binding sets of strands are a great precursor for the emergence of ribozymes with active sites formed by their secondary structure.

This experimental system and the accompanying simulations pave the way for more detailed analysis and an even better correlation. With optimizations to the ligation temperature, the ligation timesteps (complex on-rate defined by sequence space and concentration of strands) and the ligase buffer it might be possible to ligate ATGC-random sequence 12mers. Deep sequencing can help analyze potential products for their A:T:G:C-composition or GC:AT abundance depending on the location in the oligomers (due to the higher binding energy there might be a higher frequency of bases `G` and `C` at the ligation site). The pool of short building blocks can be extended to an exponentially decreasing length-distribution of strands starting e.g. from 6mers to 16mers. This would improve the analogy to a random sequence and multi-length original pool, possibly made by random non-templated assembly of nucleotides. Oligomers emerging in this setting could have arbitrary lengths. However, initial biases like the increased 3'-end `AT` motif abundance might be involved in the selection of characteristic lengths in the oligomer products. For the hypothesis of a subsequent increase in the system-randomness following the entropy reduction as proposed in ref. [119] could be tested with extended experiments running for several weeks. The A:T-ratio of oligomers in a possible time-resolution graph would then build the bimodal distribution followed by an emerging peak at the 50:50 region.

5 Theory and methods

The results and discussion section refers to several complex and detailed physical processes, methods and sample preparations that are explained in the following chapter.

5.1 Watson-Crick base pairing in DNA

The historic road to the correct DNA structure is described in detail by Tobin [120]: in 1953 it was first suggested, that DNA is predominantly found as a duplex structure by the famous paper from Watson and Crick [127]. Before, different structures with either the salts on the inside of a three-strand intertwined structure [98] or a three-strand complex with the bases on the inside bound by hydrogen bonds were suggested [80]. The break through was provided by Maurice Wilkins at King's College London with his X-ray crystallography which provided the evidence of a helical structure [45].

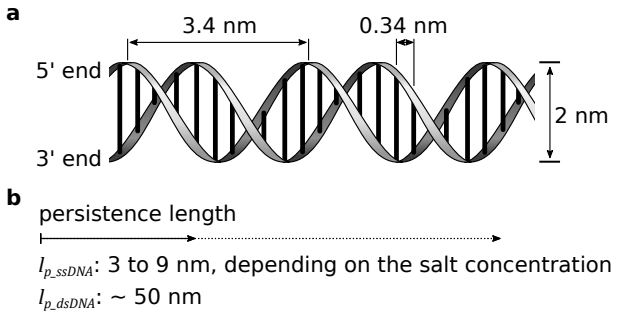

Figure 5.1 Structure of dsDNA:
a DNA forms duplex structures by Watson-Crick basepairing with the sugar-phosphate backbone spiraling in helical fashion around the main axis of the DNA strand. The strands are oriented in opposing directions and bases form distinct pairs of adenine to thymine and cytosine to guanine. For one rotation of the backbone the strands extends for 3.4 nm and spreads across 2 nm. Single basepairs are 0.34 nm apart [99].
b The persistence length gives a scale of length, at which the orientation in different parts of the strand (polymer) lose their correlation. For ssDNA and dsDNA those lengths l_p are very different: from 3 to 9 nm depending on the salt concentration [25] to about 50 nm in helical dsDNA [16].

Watson and Crick first suggested that bases are bound in a paired 1:1 fashion. The bases are perpendicular to the main helix structure at the same z-coordinate, but the two DNA strands run in opposite directions (5'-3' paired to 3'-5'). Pairs are made from a purine and pyrimidine each: adenine (A) to thymine (T) and cytosine (C) to guanine (G). With this strict pairing, the sequence on one strand imposes the sequence on the other strand. Watson and Crick specifically mention, that this "suggests a possible copying mechanism for the genetic material" [127], one of the most important reproductive properties of DNA.

In the context of the Origin of Life, the "RNA world hypothesis" [48] from Walter Gilbert published in 1986 is one of the most widely accepted basic assumptions for the emergence of life. Before, it was assumed, that a combination of information-storing RNA and catalytically active proteins build self-replicating reaction networks. The chemical components that make up RNA and proteins are very different [37] though, which led to the idea of enzymatically active RNA structures (initially found by Westheimer [128] in *Escherichia coli*, where ribonuclease-P cuts the phosphodiester bonds of t-RNA). One core concept is the process of recombination with the help of self-splicing RNA-enzymes that allow the transfer of strands from one RNA strand to another. Overall the "picture of the RNA world is one of replicating molecules [... that] builds and remakes RNA molecules by chunks and also permits the useful distinction between information and function" [48].

When experimenting with oligomers, the substitution of RNA by DNA is common as the binding parameters are very similar [107], which is also useful in the context for sequencing. The theoretical description of the binding parameters of each purine-pyrimidine basepair for RNA [123] and DNA [107] is extracted from extensive experimental data. Web-services and open source software like *mfold* [133], *unafold* [81] and *NUPACK.org* [131] provide an easy way to probe binding-energy for oligomer complexes, design specific complex structures and even compute effective melting curves for bulk systems (details in Section 5.3 and Section 5.3.2).

SantaLucia [107] gives the binding energy per 2 nt long duplex at 37 °C for different sodium concentrations. The energies for `A`- and `T`-only duplexes is significantly lower than for `C` and `G`. This publication and the underlying experimental data is the basis for the assumption of the mean binding energy in the Gillespie-based simulation in Section 3.9.

It is widely known, that the phosphate diester bonds of the DNA backbone are stable in aqueous solutions in order to protect the encoded genetic information [109, 130]. Studies on the stability of DNA [117] under different storage conditions come the result, that the most important parameter is temperature: at room temperature DNA is stable for up to 60 weeks, at 37 °C for about 6 weeks, while the best storage conditions are at -20 °C independent of the storage medium (water, TE buffer or dried). For the experimental timescale of about 60 hours, DNA can therefore be assumed to be stable and hydrolysis can be neglected.

5.2 Hairpin formation in DNA

The flexibility of the DNA backbone (and also of the RNA backbone and proteins) is one of the reasons DNA forms double helices. The persistence length l_p gives a length scale, at which the correlation of orientation in ssDNA or dsDNA strands is lost. For ssDNA l_p heavily depends on the buffer salt concentration and reaches from 3 to 9 nm [25], which is in the order of only several bases. Due to the helical structure of dsDNA its persistence length is about five to ten times larger at about 50 nm [16].

As described before, double-stranded DNA is paired such, that the strands are oriented in opposing directions. ssDNA can therefore form internal double strands with unpaired subsections. The most common one is the hairpin, which consists of a double-stranded region and an unpaired loop. The minimum strand length to form a hairpin depends on the sequence and salt concentration, but it is usually assumed that there need to be at least four unpaired bases in the loop-section [52].

In unicellular organisms and even in mammalian cells so-called shRNA (small hairpin RNA) can interfere in gene expression [21, 96]. For such applications the hairpin strand evolved to

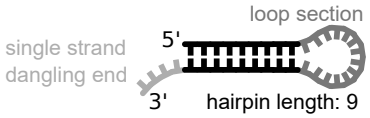

Figure 5.2 Sketch of a DNA hairpin:
DNA (and other polymers) can also bind to themselves. For DNA, the strands has two sections with reverse complement sequences called the stem and a loop-section in between them. As in the helical dsDNA, the strands are oriented in opposing directions in the hybridized section. The ends can be blunt ends, but also dangling ends on one or both sides are possible.

specifically interact with parts of the RNAi (*RNA i*nterference) machinery in the cell. DNA hairpins are common in nature, too: dsDNA is recognized by proteins involved in recombination, transcription and replication [12]. The 2010 review "Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes" from Bikard *et. al* [12] describes the different mechanisms of DNA extension by e.g. RNA induced priming of dsDNA hairpins.

For those reactions the interaction of the hairpin region with proteins is essential and the formation and dissociation of the double-stranded sections is mediated by said proteins. But in oligomer-only systems, like the presumed RNA world and especially in the Early Earth states when ribozymes had not yet emerged or were at least very rare, RNA or comparable oligomers were the only relevant molecules. With the lack of other mechanisms, the formation of hairpins was a process driven by diffusion, the bending energy of the phosphate backbone, the temperature and the strand concentration, only. In a bulk system with different strands and strand-lengths, short ssDNA strands compete for the hybridization locations on the longer template strands with the internal reverse complement sequence the hairpin-template includes. On the template strand the two reverse complement sections are spatially close which can be described as a high local concentration. Consequently, the binding-probability of short ssDNA to a template strand which can form a hairpin is low, if the concentration of the ssNDA is not comparable to the spatial distance of the two reverse complement sections of the template. For the template strand secondary structure, the bending energy of the phosphate backbone is a limiting factor. But as discussed above, the persistence length, or in this context the length at which the bending energy of the loop can be neglected, is short with 3 to 9 nm corresponding to 10 to 30 bases only. The bending energy can be assumed minimal for strands of length 48 nt when the reverse complement section are located at the 5'- and 3'-end respectively. For shorter template strands, the length of the reverse complement section is limiting the hairpin formation: As for dsDNA the number of bound bases determines the melting temperature of the complex (see Section 5.3). Strands need a certain length to be able to include the two reverse complement sections with appropriate lengths and the loop section. Too short strands have a very low melting temperature and are therefore predominantly single-stranded.

As described in the results section, the prevention of reverse complementary sequence motifs in longer template strands inhibits the formation of hairpins. Resulting strands are able to act as single-stranded templates or odd-parity extension cascade complexes in further templated ligation reactions. The experiments analyzed in Section 3.9.3 start with strands of length 12 nt, which are too short for the formation of hairpins from individual strands (at least for AT-only DNA and the system ligation-temperature). For a potential hairpin-prone template strand, the local concentration of 12mer "monomers" strands needs to be comparable to the concentration given by the spacial distance of the reverse complement sections on the template. The concentration of strand in the experiment is 10 µM and assuming a homogeneous distribution of all

strands the strand count per volume can be estimated.

$$
\begin{aligned}
10\,\mu\mathrm{M} * \frac{1}{4096} &\approx 0.0024\,\mu\mathrm{M} = 2.4 * 10^{-9}\,\frac{\mathrm{mol}}{\mathrm{L}} \\
&\approx 2.4 * 10^{-9}\,\frac{6.022 * 10^{23}}{1.0 * 10^{24}\,\mathrm{nm}^3} \\
&\approx 1.47 * 10^{-9}\,\frac{\#\text{of correct 12mer strands}}{\mathrm{nm}^3}.
\end{aligned}
\tag{5.1}
$$

Therefore, in a spherical volume with a radius of about 34 nm around a 48mer template strand there is about one other 12mer strand. With the sequence space of 4096 and 37 full length binding positions for 12mer strands (48-12+1 = 37) the probability of a fitting strand is about 1 %.

Literature also provides studies on the hybridization and renaturation of DNA. A rate of about 1/(µM s) is estimated from data of different experimental conditions [26, 71, 108, 129]. The sequence space dilutes the DNA concentration for correct binding partners, as shown in Figure 5.3 and discussed above. For 12mer dsDNA complexes the entire sequence space is relevant and the

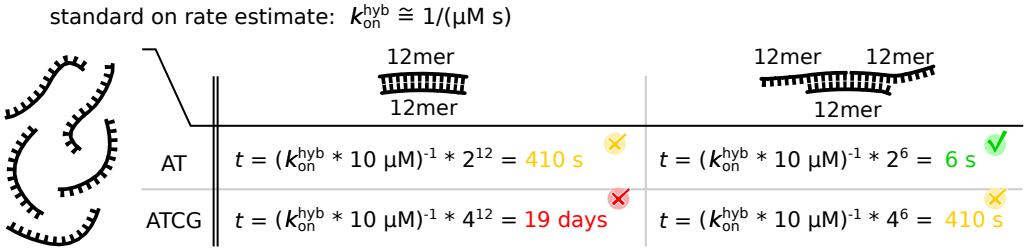

Figure 5.3 Hybridization on-rate estimation:

For systems with oligomers the k_{on}^{hyb} hybridization on-rate is estimated to be in the range of about 1/(µM s). The sequence space lowers this rate due to the reduction of possible binding partners. For a 12mer on 12mer complex with a 10 µM AT-only pool, the estimated on rate is about 410 s. In the case of a full random sequence pool k_{on}^{hyb} would increase to 19 days. For partial hybridization with dangling ends, the on-rate is lower. These are also the complex-configurations that are assumed to enable the first ligation reactions, as shown in Figure 2.1.

estimated on-rate is therefore multiplied by 2^{12} for AT-only sequences, resulting in $k_{on}^{hyb} \approx 410$ s. With all four bases and a similar complex k_{on}^{hyb} would grow to about 19 days. The first 24mer strands that emerge from the 12mer pool have probably been complexes made from three 12mer strands, two strands acting as the substrate and one as the template. The k_{on}^{hyb} in this case is lower, as only a small portion of the sequence space is now relevant as a hybridization partner. For 6 nt double-stranded sections the on rate decreases to 6 s, and in the ATGC-random case to 410 s. These estimates are the upper bound and the rates in the experiment will likely be lower.

Figure 3.41 describes the product concentration as a function of t_{lig} per cycle and total cycle count. For the AT-only sample, the product strands are observable for t_{lig} of 30 s, but not at 10 s. The calculated rate of 6 s seems slightly to high. For the x64 "double bases" sample discussed in Section 3.7.1 the product strands are fully developed for t_{lig} = 10 s already. The estimated on rate for a total concentration of 10 µM would in this case be $(k_{on}^{hyb}\ \mathrm{x}\ 10\,\mu\mathrm{M})^{-1}\ \mathrm{x}\ 2^3$ = 0.8 s. The PAGE analysis suggests, that the estimated on-rates are comparable to the experimental rates.

5.3 Melting curves in experimental settings and predictions by NUPACK.org

In contrast to detailed theoretical models of single strands or systems consisting of only a small number of strands, laboratory experiments usually need a parameter-description based on "bulk-materials". For example, the melting curve describes the temperature-dependent binding of strands to themselves or other strands. In the experiment the melting curve can be measured with the help of a FRET pair[i] or an intercalating dye. SYBR green is such an intercalating, fluorescent dye that binds to the minor groove of double-stranded DNA. The DNA sample is mixed with the dye and buffer in the concentration conditions of the experimental setting to assure best comparability. The sample is then (repeatedly) heated and cooled in a controlled fashion. At each temperature step a fluorescent measurement probes the amount of double-stranded complexes by the fluorescence of the SYBR green dye. Mergny and Lacroix published a detailed guide on the analysis of such melting curves [84]. Typically, the absorption or fluorescence data as a function of temperature has three distinct regions:

1. **low temperatures**: well below the melting temperature the fluorescence or absorption increase in a linear way due to a temperature dependence of the dye. Changes in the hybridization configurations of DNA strands are insignificant.
2. **transition region**: between the high and low temperatures the fluorescence and absorption change significantly, as double-stranded complexes separate into single strands. The point where 50 % of bases are unbound is called T_{melt}.
3. **high temperatures**: as for low temperatures, the hybridization configuration does not change (DNA and RNA are mostly single-stranded anyway) and increases or decreases in the absorption or fluorescence are due to the temperature-dependent-properties of the dye.

The main interest in melting curves are usually the melting temperature T_{melt}, the width of the transition and the amount of different hybridization domains (a strand of distinct poly-A/T and poly-G/C sections will dissociate partially at lower temperatures, and completely at elevated temperatures). The linear regions at low and elevated temperatures are linearly fit in standard experiments. All data points are then projected onto a new coordinate system based on the linear regions with the median line intersecting the data at T_{melt}. In some cases the linear regimes are not distinct or defined enough for a fit. Here, the derivative of the fluorescent signal over temperature gives an estimate for T_{melt} at the maximum point. The more the linear sections differ from a horizontal line (so the more temperature dependence the dye has), the more T_{melt} obtained by this method differs from the standard method.

Data points are recorded for at least one heating and cooling cycle, usually from about 4 °C to about 95 °C. In case the two data sets do not overlap, the system was not in equilibrium at each time-step. As already discussed in Section 5.2 and Section 3.8.3 the formation of double-stranded complexes is limited by the hybridization on-rate $k_{\text{on}}^{\text{hyb}}$. The system-$k_{\text{off}}$-rate does only depend on the binding energy, the total strand concentration and the temperature. Thus, in an out-of-equilibrium system data points for the cooling-curve are shifted towards lower temperatures.

[i] **FRET** (**F**örster **R**esonance **E**nergy **T**ransfer) is a microscopy technique in which two different fluorescent dyes are involved. The sample is only illuminated with light capable of exciting one of the dyes. The second dye can only be excited by the process of FRET, by which the excited electron of dye 1 is transferred to dye 2. Because dye 2 has a different electronic structure, the emitted light has a different wavelength than the light emitted from dye 1. The FRET process itself depends on the distance between dye 1 and dye 2 molecules. For binding essays the two dyes are attached to two different DNA strands that are then bound by Watson-Crick basepairing.

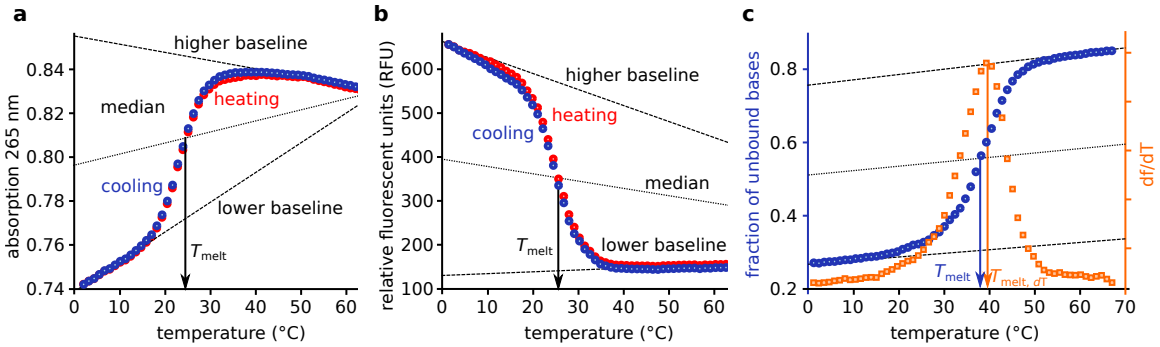

Figure 5.4 Melting curve analysis and baseline correction:
These figures are adapted from ref. [84]. Blue circles mark the data for cooling the sample down, red circles mark the data for heating the sample up. If there is hysteresis visible between the heating and the cooling pass, the temperature changes are too fast and the system is not in an equilibrium state. This is mostly due to the effective binding energy penalty arising from the sequence space of the pool of DNA: for the hybridization two strand do not only have to bind (heating: unbind) but also find their reverse complements.
a Absorption at the 265 nm as a function of temperature. Typically, the absorbance behaves in a linear fashion in regions, where the DNA binding is not significantly altered and only a function of temperature. Those regions, specifically at low and high temperatures, are fit in a linear way. The median line in between then intersects the melting curve at the T_{melt}-point.
b Typical graph as a PCR temperature cycler produces: relative fluorescent units (RFU) over temperature. Again, the low and high temperature regions are fit linearly, with the median line intersecting at T_{melt}.
c Final melting curve with fraction of unbound bases versus temperature (blue circles). The y-axis can start at fractions > 0 if not all bases can be in a Watson-Crick basepaired configuration, e.g. if poly-C tails are present in AT-only pools (might be used to identify certain strands). In some cases the melting curve is not as "well behaved" as shown here, which might make it impossible to extract T_{melt}. In those cases the derivative of the absorption, RFU or fraction of unbound bases over temperature can estimate T_{melt} at the maximum slope, as shown in orange squares. Due to a tilted background (see above: linear change of signal with temperature) T_{melt} might deviate from $T_{\text{melt, } dT}$.

5.3.1 NUPACK.org complex prediction and temperature-dependent binding

The online tool *NUPACK.org* [131] is based on the analysis algorithms of Dirks *et al.* [30–32] which are partly based on work of SantaLucia for structure and energy prediction of DNA [107] and Turner for RNA [123] (compare Section 3.9). As the temperature of the studied system is the key parameter (next to the binding energy) for the formation of dsOligomers this tool can calculate a hybridization-over-temperature curve as an idealized melting-curve prediction. In contrast to the experimental setting, the theoretical approach has certain limitations:

- **limited amount of strands:** Assessing the hybridization behavior of multiple strands computationally is difficult. All possible double-stranded conformations of at least three individual strands need to be calculated and evaluated. For large amounts of different strands, the complexity of the "complex-formation-space" becomes too large to handle.
- **interpolation for data concerning salt concentrations:** The hybridization data for the summaries and estimations in the publications of SantaLucia [107] and Turner [123] are based on several experimental studies with different salt conditions. Additional conditions are interpolated or to a certain degree extrapolated, and only accurate to a certain degree.
- **no parameter to change viscosity:** Enzymatic reactions often need specific reaction conditions, like salt (see above), surfactants, crowding agents, or simply fluid viscosity. *E.g.* the Taq DNA ligase reaction buffer has a high glycerol content and changes the sample's viscosity significantly. Additionally, complex formation is aided by the crowding agents like *Triton X-100*. These important parameters are not accessible in NUPACK but might change the melting curve noticeable.

Still, NUPACK can give an estimated binding curve for systems with only a few strands, and quite accurate melting profiles when testing the binding of two specific strands only. The prediction of structure can be used to test the hybridization behavior of designed strands, for which only the sequence is known. in comparing the structure for different complexes, the details of salt concentration and temperature are minor, due to the quasi-normalization to the same conditions.

5.3.2 Melting curves in experiments

Samples analyzed in melting profiles should be as similar to the experiment as possible. For the experiment here, the sample is pipetted as described in [74], but the ligase is substituted by a "dummy-ligase" solution which includes the salts and viscosity agents only. The DNA concentration is set to 10 µM and EVA green (intercalating dye to mark dsDNA) used as an indicator. As shown in Figure 5.5, the melting curve depends on the heating and cooling rate. While the dissociation does only depend on the effective $k_{\mathrm{on}}^{\mathrm{hyb}}$ and not on the sequence space, the heating path is not very sensitive to the exact rate. For the cooling path the formation of dsDNA depends on the sequence space, because a ssDNA needs to find a reverse complement strand for hybridization. With a larger variety of sequences at a set concentration, the abundance of possible hybridization partners becomes smaller. The result is an assembly-penalty depending on the sequence space and the cooling rate. If the system is in equilibrium at every measurement step, the heating and cooling path are similar. With a too large cooling rate, the cooling path data curve is shifted towards lower temperatures because only a local minimum was found in the hybridization energy landscape for the given set of oligomers. Figure 5.5a shows said hysteresis for a 48mer AT-only random sequence DNA sample. In panel b the cooling rate is significantly lower and the melting profile is similar for the heating and the cooling path with T_{melt} of about 50 °C.

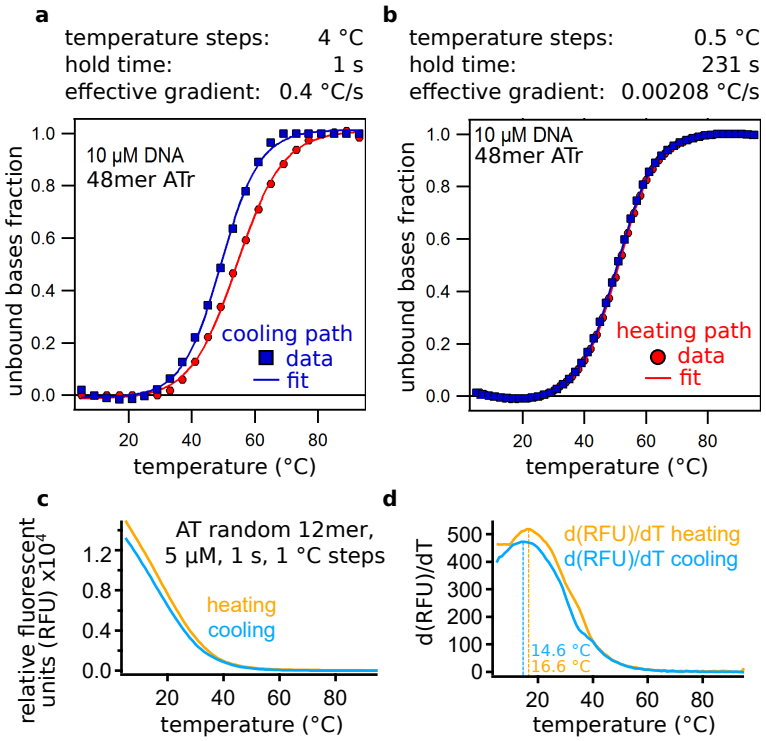

Figure 5.5 Experimental melting curves:
The cooling and heating rate of the 48mer random sequence DNA system alter the slope and the hysteresis between both paths:
a An effective heating and cooling rate of about 0.4 °C/s leads to hysteresis. The cooling path is shifted to lower temperatures due to the energetic penalty induced by the sequence space. While the dissociation does only depend on the concentration (balance of k_{on} and k_{off}) and the temperature of the system, strands also need to find a reverse complement sequence binding partner during the cool down.
b For significantly lower temperature ramps with an effective rate of 0.00208 °C/s the hysteresis is not observed, as the system is in equilibrium at each temperature step.
c For the 12mer AT-only random sequence sample at a concentration of 5 µM there are no clear linear sections for low temperatures.
d The RFU change over temperature reveals that T_{lig} is around 15 °C for both heating an cooling paths for the AT-only random sequence sample. The initial linear slope cannot be analyzed experimentally with the given conditions as water freezes at about 0 °C (sample including salt might change the freezing point of the sample).

For the 12mer AT-only random sequence sample, the experimental melting curve is missing the initial linear part. Analyzing the derivative of the relative fluorescent units over temperature indicates a T_{melt} of about 15 to 16 °C. The transition from the low temperatures to the medium temperatures, as discussed above, is likely in the negative Celsius degree range, where this analysis not possible due to the freezing of the medium. Importantly, even at elevated temperatures of up to 40 °C there are still double-stranded complexes, that might be ligated. And as shown in Figure 3.38a the experiment still works for $T_{\text{lig}} = 30$ °C. For experiments with a smaller sequence space like the x64 "double bases" pool even for higher T_{lig} are possible, see Figure 3.24.

In Section 3.8.3 the importance of "time per ligation step" to "temperature cycle count" ratio is discussed. Both ends of the spectrum, for either short T_{lig} of 10 s but multiple thousands of temperature cycles, or only a few cycles but very long T_{lig} steps, result in low product concentrations. This is due to two reasons: The system needs the temperature cycles to dissociate at least some of the double-stranded structures for further templated ligation reactions. At the same time, long enough T_{lig} are need for strands to find the correct hybridization partners in order to build ligatable complexes. The formation of dsDNA for the melting profiles follows the same estimation as shown in Section 3.8.3.

Binary sequence alphabets in DNA can be realized with either bases `A` and `T` or with bases `C` and `G`. Because the binding energy for the latter case is higher than for AT-only, the melting temperature of a similar random sequence system is higher, as shown in Figure 5.6.

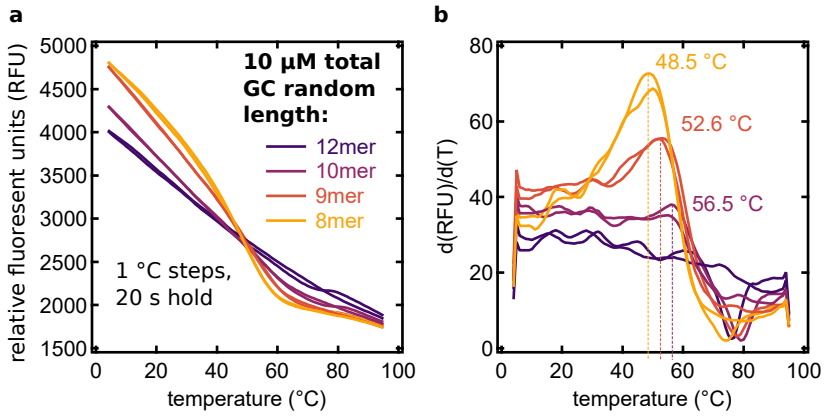

Figure 5.6 GC random sequence pool melting profile:
In all cases the effective heating rate was 0.05 °C/s. The heating and cooling path are almost identical for all lengths suggesting equilibrium conditions at each step.
a Relative fluorescent units (RFU) over temperature for EVA green dye with GC-only random sequence pools of lengths 12mer, 10mer, 9mer, and 8mer.
b Due to the steep slope at lower temperatures, the determination of the melting temperature by the linear fit method as described in Figure 5.4 cannot be applied. In this case here, the derivative method gives an upper boundary for T_{melt}.

As described above, for very steep slopes the determination of T_{melt} by fitting the linear regions in low and elevated temperature regimes can fail but the derivative method can give an estimated value as the upper limit. The 8mer GC-random sample has the lowest T_{melt} with the 9mer and 10mer having 2 °C higher T_{melt} each. The RFU over temperature for the 12mer GC-only sample is very linear and doesn't show the characteristic shape of a melting curve, which makes the determination of T_{melt} impossible. The 12mer GC-random pool NGS data (see Figure 3.22b) suggests an abundance of poly-`G` motifs, known as sticky bases. These structures might have

high melting temperatures which is further increased by the abundance of G-type strands causing a limited sequence space at the same total concentration.

5.4 NanoDrop ssDNA concentration measurement

DNA ordered from *biomers.net* is already set to the desired stock concentration of 200 µM, but this concentration adjustment seems unreliable when remeasuring. Therefore, all DNA stocks are again quantified with a different but uniform method. The measurement method of choice is the NanoDrop 3300 and NanoDrop One by *Thermo Scientific* which records the attenuation of light between wavelengths of 220 nm and 350 nm for liquid samples with DNA in small 2 µl droplets.

biomers.net includes the extinction coefficient for all strands in their data sheet. The extinction coefficient is a measure for the attenuation of light due to the presence of DNA, that absorbs light of certain wavelengths. Typically, for DNA the α-260 nm line is analyzed. For example, the sequence `TTTTTTAAAAAA` has an extinction coefficient of 147600 L/(mol*cm). In the NanoDrop measured 0.931, 0.906, and 0.916 1/cm for an expected concentration of 5 µM which gives the droplet concentration in the NanoDrop as:

$$\begin{aligned} \text{mean}(0.931, 0.906, 0.916) &= 0.9177\ 1/\text{cm} \\ \frac{0.9177\ \frac{1}{\text{cm}}}{147600\ \frac{\text{L}}{\text{mol*cm}}} 10^6 &= 6.2\,\mu\text{M}. \end{aligned} \tag{5.2}$$

In this case here, the concentration is about 24 % too high. During pipetting the sample stock with an ordered concentration of 100 µM is therefore treated as a stock with 124.35 µM. The extinction coefficient depends on the base composition of the oligomers: Poly-A strands and poly-T strands have distinctly different extinction coefficients (184800 L/(mol*cm)) for 12x `A` and 110400 L/(mol*cm) for 12x `T`). A composition of random sequence 12mer strands with all possible base fraction can therefore only be measured as a superposition of extinctions coefficients and only gives a bulk- or system concentration approximation. For the AT-only random sample *biomers.net* gives the extinction coefficient 147600 1/cm which is simply the coefficient for a 50:50 A:T-ratio strand.

Although determining the stock concentration in this way is also subjected to additional errors like pipetting or sub optimal mixing, all measurements were done in a very similar way and are therefore better comparable and more accurate than mixing the stock samples without remeasuring.

In Section 5.6 a quantifying method for the concentrations in PAGE gels is discussed in detail. While the measurement of concentration for single strand species is well suited for the NanoDrop machine, there are clear limitations to this approach. Especially the non-trivial oligomer length distribution in complex reaction buffers cannot be analyzed by this photometric attenuation method.

5.5 Enzymatic templated ligation

In cells and microorganisms ligases repair nicks in double-stranded DNA [121]. Without these enzymes nicked DNA double strands would degrade much faster due to processes like radiation, mutagens or recombination [77, 106]. The properties and the catalyzed bond formation of the

oligomer backbone has thus a widespread application in biotechnology [8, 22, 112] for applications like cloning and strands sequencing, as used in this study.

The chemistry of the templated ligation is explained on the basis of the Taq DNA ligase (also-called *Tth* ligase) and differences to other ligases are pointed out as fit. The Taq DNA ligase itself is extracted from the bacterium *Thermus thermophilus* which is indigenous to hot water springs and can withstand high temperatures. Ref [70] describes the ligation chemistry in detail: between the hydroxyl group of the 3'-end of the first strand and the phosphate of the 5'-start of the second strand the ligase catalyzes the formation of a phosphodiester bond. Cao [23] and Lehnman [75] found, that the ligation happens in three steps. First, the ligase reacts with NAD^+ and modifies an active site lysine residue with adenosine monophosphate (AMP), which is then transferred to the 5'-phosphate of the second substrate strand. Finally, the 3'-hydroxyl group reacts with the activated 5'-start, resulting in a native phosphodiester bond and the release AMP. In other ligases like the N°9 DNA ligase the first step can also be fueled by ATP.

In context of the experimental setting of this study, the exact chemistry of bond-formation is not discussed. Rather, the ligase interacts with complexes emerging from the random sequence pool with implications for all strands. The most important properties are then the ligase specificity, a possible sequence dependent ligation rate, the temperature-dependent stability, and the temperature-dependent activity of the ligase.

- **Ligase specificity:** Ideally for the experiment discussed here, the ligation reaction should be very specific and only ligate the perfect reverse complement sequence substrate strands hybridized to the template strand. But DNA ligases typically have some probability of ligating incorrectly paired strands. Extensive studies for the specificity of the Taq DNA ligase were for example done by Kim and Mrksich [70]. For the last 3'-end base of the upstream substrate strand they found a high specificity with almost no incorrectly ligated strands, but the overall yield was below 100 %. A study from Lohman *et al.* [78] analyzed the specificity for the last upstream and the first downstream base for templated ligation with the Taq DNA ligase at different temperatures and pH conditions. While the overall specificity was again high, especially for the eight possible `A`, `T` only cases (four 2 nt motifs, two all possible template and primer combinations each), there were incorrectly paired but which still resulted in ligated product strands. The unfortunate binning in the analysis of said study makes the detailed analysis difficult, but the result is somewhat consistent: only the template dimer `AA` templating the substrate dimer `TT` has a <80 % yield, while all other correctly paired configurations yield >80 %. This possible bias is included in the simulation in Figure 3.17 as a worst case scenario and only had a minor impact on the emerging sequences. *NEB* itself expects the incorrectly paired and thus ligated strands with a probability of 1.3 %.
- **Sequence dependent ligation rate:** Extending the above point, while the specificity of a ligase can be very high, that does not necessarily mean that all sequence motifs are ligated with the same efficiency. Overall, evolved enzymes like the Taq DNA liagse or the N9° DNA ligase with their biological origin in excision repairing of DNA, have a high efficiency and ligation yield for all correctly paired sequence motifs. In the study of Lohman *et al.* [78] the 2 nt sequence motif at position ±1 from the ligation site is analyzed for all possible template and substrate motifs. As the AT-only sample does not contain bases `G` and `C`, only eight data points are relevant. Unfortunately, the yield for each motif in this study is binned in only five coarse sections. The >80 % bin and the 50 %-80 % bin have their border in the region of the expected yield, as shown above. However, the yield for all motifs is >80 %, except for motif `AA` and template `TT`, which has a yield of 50 %-80 %. For a second experiment in said study the buffer pH is increased from 7.5 to 8.0 and all motifs

have similar yields of above 80 %. This suggests, that while the yield of motif `AA` might be slightly lower than the other motifs for the Taq DNA ligase, the difference is not substantial. The coarse binning of the ligation efficiency is used as an input for a worst-case ligation scenario in the corresponding simulation in Figure 3.17.

- **Temperature-dependent activity:** *New England Biolabs* reports the activity of the Taq DNA ligase in units/ml. One unit is defined with help of a standardized essay: "One unit is defined as the amount of enzyme required to give 50 % ligation of the 12-base pair cohesive ends of 1 µg of BstEII-digested λ DNA in a total reaction volume of 50 µl in 15 minutes at 45 °C". Figure 5.7 shows the manufacturers values. Importantly, as already discussed, this

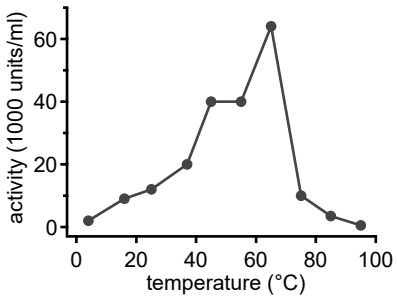

Figure 5.7 Taq DNA ligase activity over temperature:
For the given test scenario of ligating the 12 nt cohesive ends of λ DNA (see text) the ligases activity is high between 40 and 70 °C. For lower temperatures the activity is also lower, but the enzyme is still active. Data adapted from *New England Biolabs*.

is a system-activity and not measured per ligatable complex. The system ligation rate is *e.g.* dependent on the abundance of double-stranded complexes, the complex k_{off} dehybridization rate, and the enzyme concentration. Therefore, a low activity in the test-scenario by *NEB* is only an indicator for a higher activity at elevated temperatures. In the 12mer random sequence DNA pool of this study, the emergence of product strands seems to be governed by the hybridization on-rate, as shown in Section 3.8.3.

- **Temperature-dependent stability:** As described by the manufacturer, the ligase is stable at elevated temperatures for extended amounts of time. In experiments with $T_{\text{melt}} = 95\,°\text{C}$ (see Figure 3.38 the product stand distributions emerge in a comparable form as for $T_{\text{melt}} = 75\,°\text{C}$, suggesting that the ligase is active for the at least the majority of temperature cycles. Even after several thousand temperature cycles and more than 100 hours of experimental time the ligase enzyme is still active (see Figure 3.43 and Figure 3.53) for $T_{\text{melt}} = 75\,°\text{C}$. In the experimental context this means that the elongation mechanism by templated ligation is working in every temperature cycle and after extended time periods, though possibly with a reduced rate (compare Figure 4.3). Though, this is difficult to quantify, as the system-ligation rate depends on the complex concentration, as described above.

In addition to the above discussed ligase parameters, the ligation in the experimental setting in this study can also be interpreted on a "system-level". As assumed in the simulation in Section 3.9, the effective extension rate r_{ext} of a strand depends on the hybridization on-rate k_{on} the hybridization off-rate k_{off} and the "microscopic" ligation rate k_{lig} (that might be biased due to ligase preferences). Therefore, the extension of long complexes can become very likely due to their very low dissociation rate, while the k_{lig} ligation rate of the ligation site motif might be unfavorable (see Section 3.9.3). This is also the main result from the Gillespie-based simulation: Depending on the its length, a strand can fall into several dynamical regimes, dominated

by either dissociation, extension, or degradation. The about homogeneous ligation rate of the Taq DNA ligase can be the fastest or slowest rate compared to the other relevant rates in the experimental system, as those rates are not constant but depend on the complex-length.

5.6 PAGE concentration quantification LabView tool

As shown in Section 5.4 the quantification of a fluid sample DNA-concentration is easily done by analyzing the α-260 nm line. But this method has clear limitations:

- **length distribution:** A similar problem as described in Section 5.4 arises from the non-trivial length distribution. The extinction coefficient utilized in the NanoDrop analysis results from the amount and type of bases in a strand. Longer strands have higher extinction coefficients, but the NanoDrop only gives a single attenuation over wavelength characteristic which is not sufficiently analyzed by a single effective extinction coefficient.
- **buffer conditions:** The Taq DNA ligase buffer can change the binding interaction for single strands and therefore change the attenuation of the UV light in the NanoDrop. The surfactant *Triton X-100* also absorbs UV-light and changes the attenuation characteristic.

The standard analysis method for samples with different strand lengths is the PAGE analysis. With an electric field the charged DNA is pulled through the polyacrylamide gel acting as a sieve. Short strands can get through the fine pores faster than longer strands which results in a length resolution of the DNA strands in the sample. Strands of a similar length and charge (mind possible modification to the DNA, like a 5' POH) are pulled to the same position and called a "band". On each gel several different samples can be analyzed at the same time, one in every "lane" of the gel. To visualize the DNA it can either already include fluorophores as modifications to the DNA bases or backbone or the DNA can be stained after the samples are pulled into the polyacrylamide. This so-called post-staining is typically done with an intercalating dye that is bound either to the major or minor groove of double helix DNA (like SYBR green) or a dye that is bound in between the bases for ssDNA like SYBR gold, as used in this study. In denaturing PAGE conditions the urea concentration and temperature are high and the loading dye includes formamide to dissociate all dsDNA and inhibit rehybridization of complexes. PAGE gels can also be used in the non-denaturing conditions to analyze dsDNA. In the final step the gel is imaged on a device that consists of a light-proof box with an internal camera and LEDs to excite the fluorescent dyes (ChemiDoc Mp, *BioRad*). The details of sample preparation for the PAGE analysis are explained in Section 5.11.

Generally, calculating the concentration of individual bands on PAGE gels is not possible and the only measurements derived from the gel are a coarse quantification of the fluorescent response of the bands. Those limitations arise due to the possible non-linear fluorescent response of the dye depending on the length of the dyed strand, changing base-compositions per strand length, unknown starting concentrations and dsDNA even in denaturing gels.

In this case here, the properties of the AT-only DNA sample allow for quantification of the concentration on the PAGE gel, as long as some requirements are met. First of all, the total number of 12mer oligomers does not change from lane to lane, only their distribution. Without ligation, all oligomers are 12mer "monomers" while for temperature cycled samples some 12mers are ligated and are therefore found in longer strands (an 48mer consists of four 12mers that are now missing form the 12mer pool). Then, SYBR gold has an about linear fluorescence response with the increase of strand length. For a 24mer band and a 48mer band of same concentration, the 48mer band has an about doubled fluorescence intensity. This means every 12mer subsection

has about the same probability to be stained by the SYBR gold dye. The length distribution is also well defined with known lengths of multiples of 12 nt and even bands of long strands can be distinguished from the neighboring lengths.

When preparing the experiment, for example to screen the influence of the ligation temperature T_{lig} on the length distribution (see Figure 3.36), the sample is prepared as one well mixed master sample which is then distributed to multiple tubes for each experimental condition to ensure best comparability. Therefore, the total intensity of each lane must be similar for all lanes. Differences in the intensity are most likely due to the pipetting error in the low-volume but high-viscosity samples. This can easily be corrected by scaling the total intensity to the reference lane that only includes the 12mer monomers and was not cycled but kept in the fridge at 4 °C.

A LabView tool assists in the concentration quantification (CQ, CQ-tool), the individual steps are as follows:

1. **Load image and define lanes:** The unaltered gel image is loaded into the LabView tool. All lanes on the gel are selected with the help of four cursor points, see Figure 5.8b. The points mark the top and bottom of the left-most lane as well as the the top and bottom of the right-most lane. Lane locations in between are linearly interpolated and marked with white lines. The gel background is also linearly interpolated from the marker data and plotted with red lines.

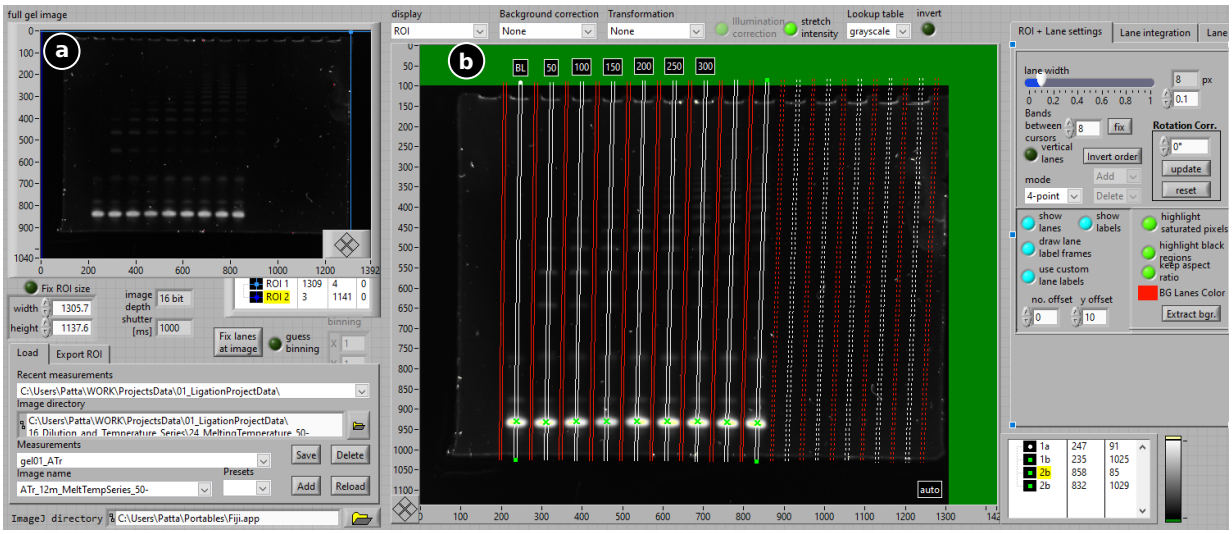

Figure 5.8 Image import and lane selection:
a The gel image can directly be imported from the ChemiDoc gel imaging station.
b Four cursor points are used to select the top and bottom of the outermost gel lanes. The lanes and background regions in between are linearly interpolated from the cursor location data.

2. **Baseline correction:** From the imported gel image and the marked lanes all intensity values per position are plotted, as shown in Figure 5.9a, and background intensities are also plotted in Figure 5.9b. The x-axis for each lane is an arc with length 1. For comparison, the 12mer peak position and the original gel pocket are used to align all lanes and background regions. The tool offers additional adjustments like baseline corrections or lane normalization. In the analysis performed here, the those options are not used, only the background lanes are smoothed slightly (range of eight data points).

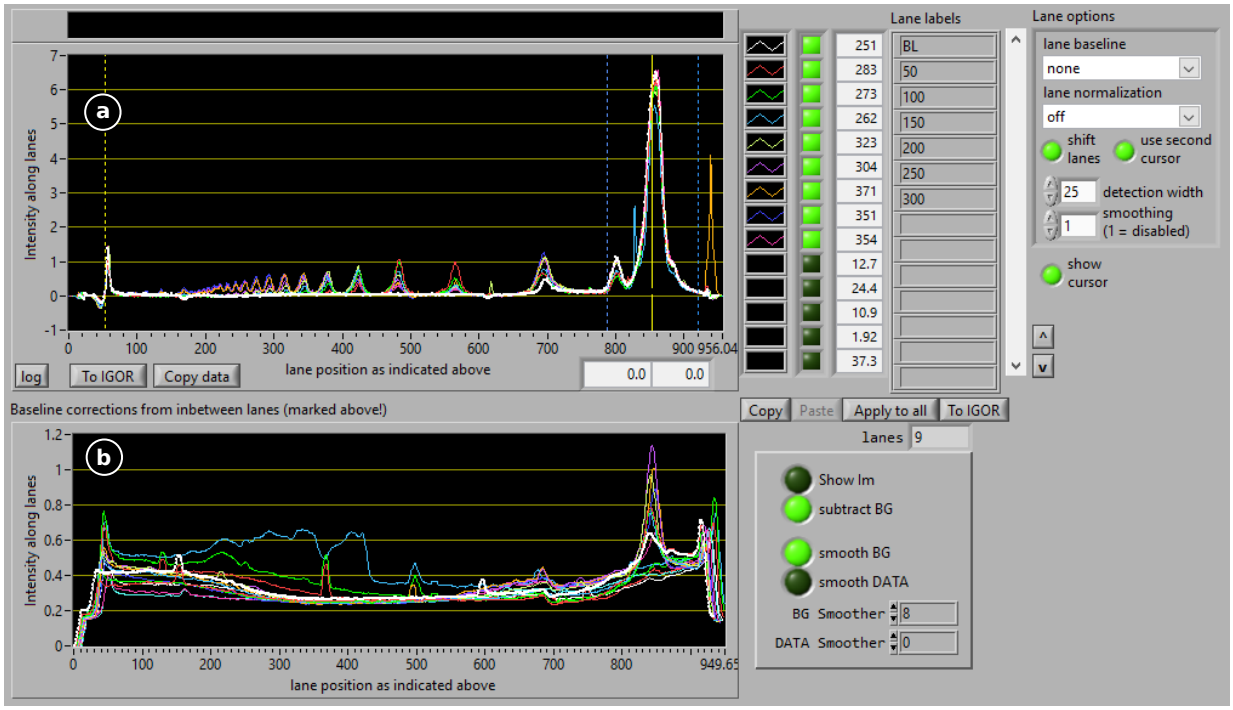

Figure 5.9 Lane orientation and background correction:
a Two markers define the position of all lanes at similar locations. The leftmost marker gives best alignment results when focused on the gel pocket on top of the gel The rightmost marker selects the 12mer monomer pool position as the smallest size DNA in the experiment.
b Inter-lane regions marked red in Figure 5.8b define the background left and right of each analyzed lane. The options on the right hand side enable smoothing of the background (selected: smoothing windows of eight data points).

3. **Calculate uniform x-axis and adjust y-scaling:** For comparison of lanes the x-axis must be homogeneous. In Figure 5.10a the tool imposes a new x-axis with a factor of x more data points than the original data set. For all lanes the intensity data is linearly interpolated between the original points. The right hand side graph crops the x-axis to the relevant section and can scale the y-axis to arbitrary values.
4. **Mark peak areas and calculate concentration:** The intensity graph shown in Figure 5.10b is loaded in Figure 5.11a. With vertical markers the peak regions are manually selected. A second mode allows for automatic peak selection, but this mode struggles with the low peak heights for very long strands. Selected regions are numbered and the areas for each peak are calculated and divided by the length of the corresponding strand length. The total peak area for each lane is summed up and that sum is normalized to the entered value. In Figure 5.11b the peak-areas are scaled to the known concentration of the baseline lane.

The data can further be adjusted to give a better approximation of the measurement quality. In an ideal baseline lane only the peak for 12mer "monomer" data should be visible. But as already seen in the PAGE gels shown before and in the CQ-analysis in Figure 5.9, the baseline has more intensity signals. The most obvious are two bands at approximately 16 nt and 24 nt, as shown in Figure 5.13b. At the shorter length, the artifacts are only visible if DNA and the Taq reaction buffer are combined, suggesting a buffer component being stained. The artifact at a length of approximately 24 nt is observed for all conditions, which suggests it could be part of the DNA

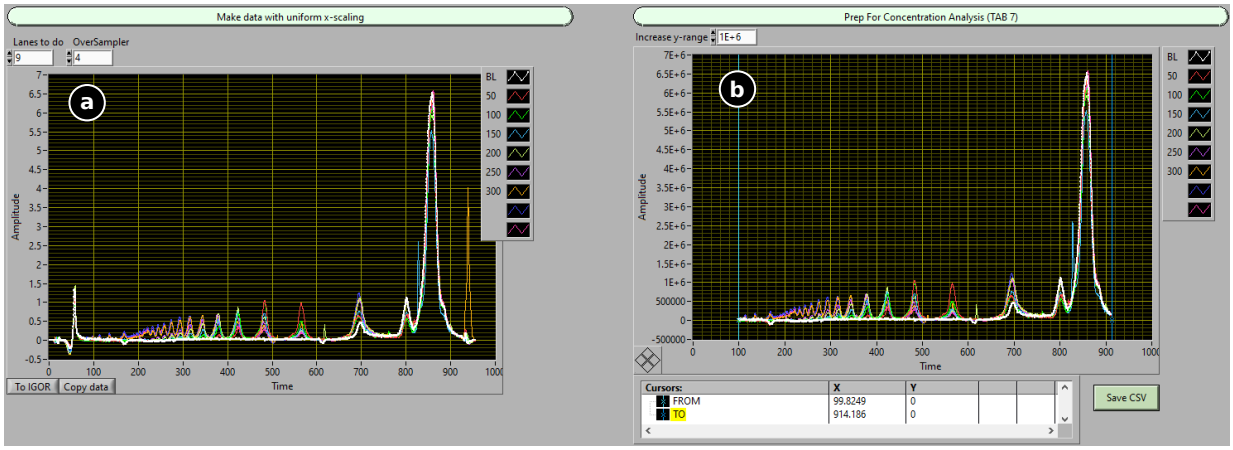

Figure 5.10 Uniform x-axis and cropping:

a The lanes selected in Figure 5.8a and shifted in Figure 5.9a have individual x-axes. Here, a new universal x-axis is imposed for all lanes. Original data points are linearly interpolated on the universal x-axis.

b Two markers define the cropping region. Additionally, the y-axis scaling can be changed to improve readability.

sample. It is likely, that those are DNA strands which accidentally developed a second backbone at the 3' nucleoside during synthesis. Here, one extension is regularly extended at the phosphate backbone, but a second one is extended from the first base (3'-end). The resulting molecule has one 3'-end and two 5'-ends and runs approximately like a regular 24mer in the PAGE analysis. Additionally, the baseline is not exactly zero, even after correcting the background, as shown in Figure 5.9b. Lanes with DNA are almost always slightly lighter than the surrounding gel.

This information can be used to improve the CQ-data: Figure 5.13 describes the corresponding analysis steps: In panel a the unedited data from Figure 5.11b is shown, the baseline in red. For the 24mer position including the incorrectly synthesized (and likely inactive) 24mer artifacts and all following baseline values that simply mark the base intensity for a lane are subtracted from all other lanes. The baseline intensity for the 24merr+ position is extrapolated on the lin-log plot from the data of the 36mer and longer of the baseline. Because the baseline only plots the intensity of the background for said range, the given concentration can be described as the detection limit of the CQ method here.

The reproducibility of this analysis is shown in Figure 3.36 with error bars marking the average from six measurements. The limits of this method are clearly the resolution of the gel image and the background noise. Especially for long products strands, it becomes successively more difficult to distinguish neighboring bands and therefore peaks. At longer lengths the systematic error becomes larger as well, because the concentration is equivalent to the peak area marked by the intensity of the band as a low intensity is more susceptible to noise. As shown in Figure 5.13c the detection limit allows for the quantification of concentrations down to 0.001 µM for strands at a length of 144 nt with seemingly negligible amounts of noise-induced jitters resulting in smooth curves.

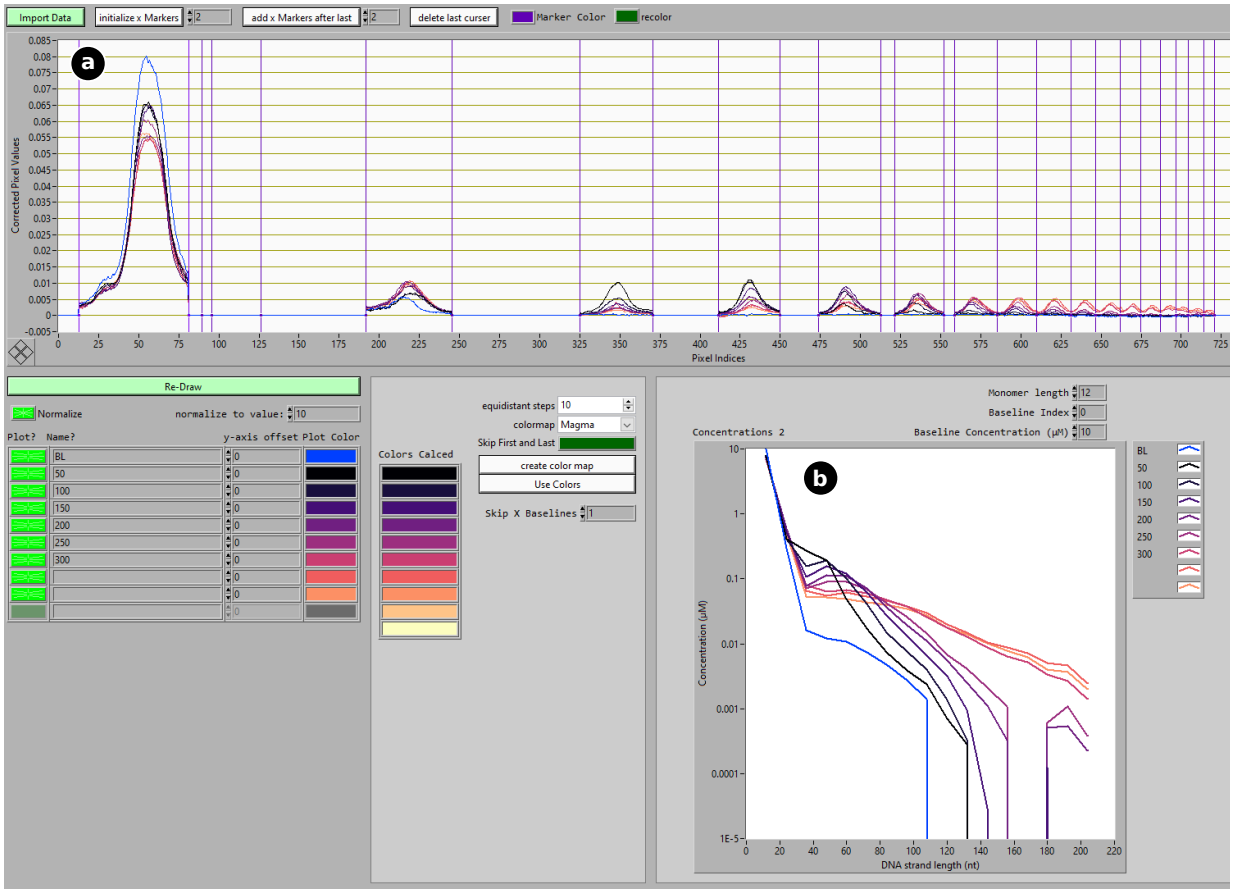

Figure 5.11 Peak selection and concentration quantification:

a All lanes are imported from Figure 5.10b. Vertical cursors are used to define the peak positions by hand, as automatic peak detection struggles with recognizing the small peaks for long product strands. Regions in between peaks are excluded from the analysis. Every peak area is divided by the length of the corresponding lane (first peak divided by 12, second by 24, third by 36, etc. depending on the length of the monomer building block lengths).
b All peak areas are summed up for each lane and the y-axis is scaled to normalize the total area to the baseline lane.

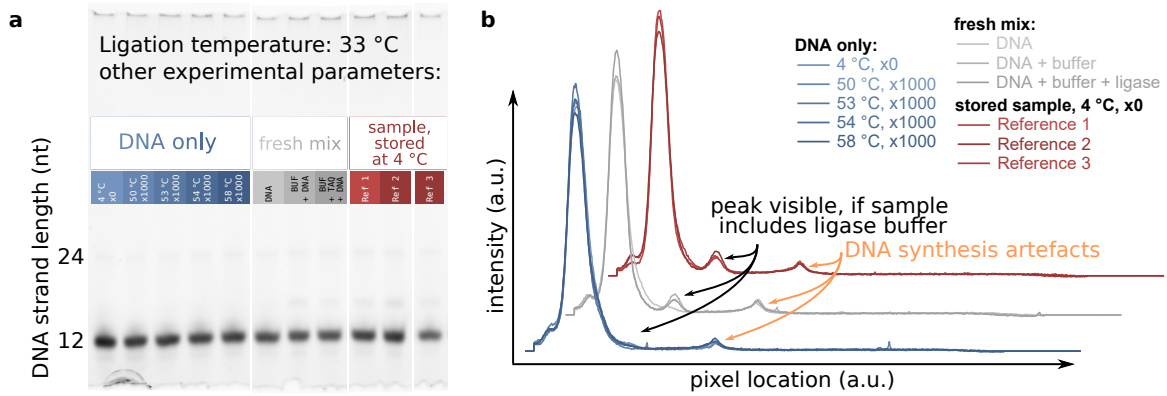

Figure 5.12 PAGE baseline artifacts:

a For altered experimental conditions the PAGE shows different artifacts.
b For DNA only, without the Taq reaction buffer and the Taq DNA ligase, the artifact at a length of about 16 nt cannot be observed. A fresh mix of DNA and buffers is very similar to the a mix stored in the fridge for about 65 hours. All samples show the artifact at a length of 24 nt, which is likely due to a faulty DNA synthesis.

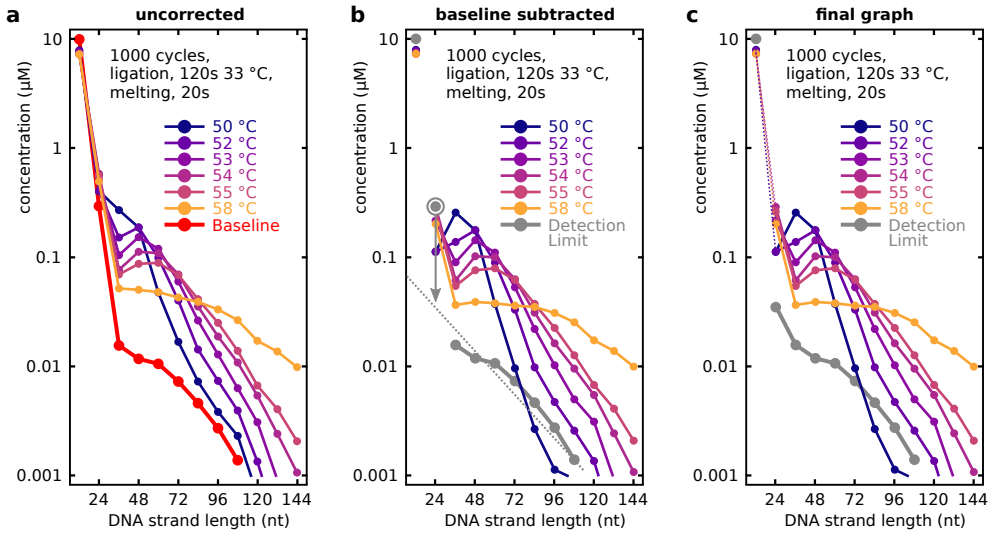

Figure 5.13 CQ-tool additional baseline correction:

a Original data with baseline marked red.
b The baseline for "dimers" and longer is subtracted from all data lanes. The 24mer baseline is extrapolated from the lin-log plot of the slightly lighter background for longer strands. This technically marks the detection limit of the CQ-tool.
c The final graph shows the concentrations as accurately as possible with the CQ-tool.

5.7 DNA sequencing

When it comes to length and structure prediction for DNA, RNA or proteins electrophoresis is usually the way to go. This method uses an electric field to drag DNA through a gel substrate acting as a sieve: short strands travel further than long strands. For size comparison one lane is often reserved for a ladder, a sample with known DNA strand lengths and concentrations. A common medium for DNA strands form single nucleotides up until a length of 500 nt is polyacrylamide. In combination with TRIS-boric acid low EDTA buffer, 8.0 M of urea and a temperature of about 55-60 °C this analysis is called a denaturing gel, as double-stranded DNA is melted and retained as single strands. Without these conditions DNA hybridization and complexes are kept and different configurations might be distinguished [43, 91] (also compare Section 5.11).

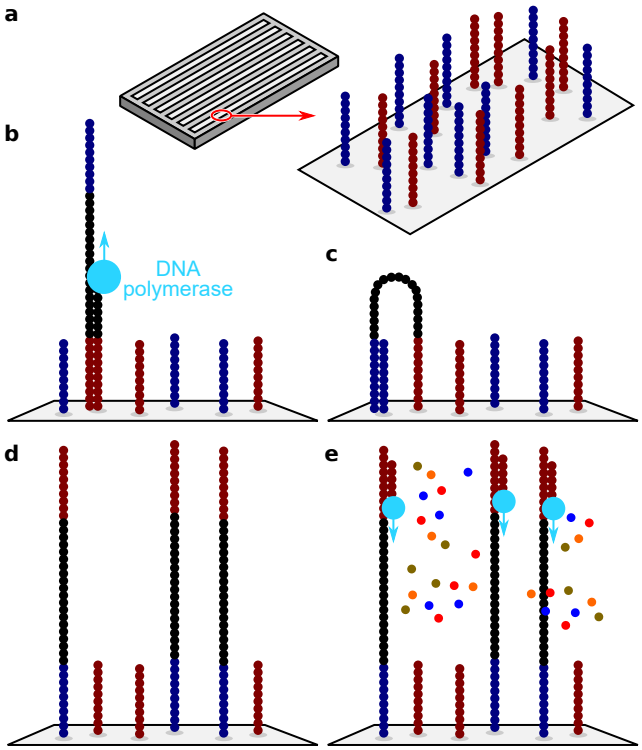

Figure 5.14 *Illumina* deep sequencing technique:
a The flow cell has several lanes for different samples. In each lane there is a so-called "lawn" of short DNA fragments that bind the DNA adapters.
b Hybridized DNA strands are replicated by a DNA polymerase. The original strand is dissociated and washed away.
c The remaining replicated DNA folds over to another "lawn" strand and is again rebuild by a polymerase. This step is called bridge amplification and resulting in forward and reverse strands for all lawn DNA fragments.
d For sequencing, the reverse complement strands are cleaved and washed away.
e The DNA strands are sequenced by synthesis. Each mono-nucleotide is labeled and its attachment to the DNA strand is detected in the NGS device by fluorescence.

When it comes to the sequence-level of DNA analysis, sequencing is the only possibility to analyze and understand long strands. The term Next generation Sequencing (NGS) describes a sequencing method that is able to provide highly parallelized and fast analysis of DNA. There are several methods that have advantages for different kinds of sample: NanoPore sequencing

is best suited for very fast and cheap sequencing of long DNA strands of several thousand bases, but lacks in accuracy and read quality. In this work we used a PCR-driven sequencing method with optical detection in a flow-cell from *Illumina*. The details of the sample preparation and analysis are discussed in detail in Section 5.12. As a summary: adapters are attached on the 5' and 3' end of ssDNA.

Figure 5.14 shows, that these adapters bind the DNA to the flow cell. *Illumina* utilizes sequencing by synthesis: The DNA is first hybridized to the so-called "lawn" of DNA fragments with adapter sequences fit to bind the prepared DNA (see Section 5.12). Clusters with similar DNA sequences are formed by bridge amplifications. Here, a hybridized strand is rebuild by a polymerase. The new end folds over to a neighboring "lawn" adapter strands and is then again rebuild by the polymerase. This results in clusters of similar sequences, which makes detection for the *HiSeq* device easier, as several strands are subjected to the same reaction at the same time and location: The sequencing itself is performed by synthesis of the DNA with fluorescently labeled mono-nucleotides in a base-by-base fashion. All sequence data is evaluated for it's read quality and exported in a single file with the front adapter sequences removed already.

5.8 Sequence distance metrics

Simple sequence analysis might included operations like the comparison of entropy or the base-content of a strand. Those methods usually do not need to compare different strands on a sequence level. But at some point there is a need to compare two strand in a meaningful and accurate way. Because sequencing data of a single strand does indeed look like a long word, text based analysis is the standard method. This is easily explained with an example: lets compare sequences `ATAAA` and `AGAAA`. The obviously differ in position 2. A simple algorithm like the Hamming distance will classify this difference between the two strands with a value of 1, describing the amount of operations needed to change one strand into the other. There are several *edit distance* types, differing in the kind of operations that are included. The before mentioned Hamming distance for example, does only allow for substitution operations. Selecting the applicable edit distance for an analysis is vital: NanoPore sequencing results do often have insertion and deletion as their primary failure mode. In the following case, the Hamming distance will deliver wildly inaccurate results: `AATTACAC` and `AATACAC` have a Hamming distance of 5. An edit distance such as the "Levenshtein distance" that does include the operations "insertion" and "deletion" is needed. For the example above, the Levenshtein distance is 1: The recognition of the deletion on the 3rd position `T` is a way more accurate description of the distance between the two strands, while taking the reason and mechanism for the difference into account. This distance metric is used in the sequencing error estimation discussed in Section 5.16.

For two strands those metrics are a valuable comparison. Indeed, such analysis is used all the time e.g. in text prediction - but it lacks in certain properties in order to compare large sets of sequences. Evaluating the distance between two coordinate points in 2D space $P_0(x_0, y_0)$ and $P_1(x_1, y_1)$ is an easy task:

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}. \tag{5.3}$$

"Hidden" in the coordinates, there is a description of the *direction*, to go from one point to the other. This a fundamental difference to edit distances: there is no information about the position where two strands differ, nor a possibility to space them in a useful coordinate system (it is possible to have a $4^{\text{strand lenth}}$-sized coordinate system). Additionally, the algorithms determining

those distances all have compute times in the order of O^2. This limits the usability drastically, as DNA sequence information is typically long and the amount of different strands huge. Therefore, distance metrics are usually only involved in comparing an analyzed strand to a known strand like a spiked sequence or primer sequence.

5.9 A-type/ T-type bias model and kinetic simulation of 24mer assembly (ref. [74])

For the A:T-ratio analysis in Section 3.3 the simulation of Tkachenko and Maslov [119] is adapted, as shown below.

5.9.1 Base composition evolution mediated by internal hairpins

Besides the emergence of the sequence patterns that depend on the position in the multimer strand, the rapid shift from a binomial base-composition distribution to a bimodal one is the most obvious and reproducible effect in the random sequence pool. From longer multimers, especially 72mers and 84mers, the distribution gets narrower and the sequences in A-type and T-type groups become very self-similar. The strands in both groups are unfit for internal hairpin formation, because the only reverse complement stretches are AT-patterns on the ligation sites. But the poly-`A` and poly-`T` stretches in between can't hybridize. This self-segregation into the A-type and T-type groups appears, because the ability to template further ligation reactions is reduced for a strand that can form internal hairpins. The formation of such hairpins can easily be suppressed by a base composition bias.

A sequence with length N and binary base composition p (the fraction of bases `A` of the entire strand) can have an internal hairpin with length $l_0 \geq 0$ consisting of the left part l_{0l} and the right part l_{0r} each also of length l_0. The probability of two bases in the sequence to be complementary is $2p(1-p)$. The probability for a complementary section in the strand with length l_0 is then $(2p(1-p))^{l_0}$. There are $(N-2l_0)^2/2$ ways of choosing two non-overlapping segments in a strand of length N, assuming very long strands. For the expected amount of hairpins of length l_0 being 1, we get an equation relating the strand length to the maximum hairpin length:

$$N = 2l_0 + \sqrt{2}(2p(1-p))^{-l_0/2}. \tag{5.4}$$

If the ligation reaction would generate random sequences, the resulting ensemble would have the maximum entropy. The corresponding distribution is a Gibbs-Boltzmann distribution, with an abundance of $\mathrm{e}^{\lambda p}$ for each individual sequence. If the selection in the pool would only depend on the base composition p, a maximum entropy ensemble will still give a correct PDF, $f(p)$. The number of strands with a given compositions can again be approximated with a Gaussian curve $\sim \mathrm{e}^{-2N(p-1/2)^2}$ with $p = x + 1/2$ and the mean compositions at $1/2 \pm x_0$:

$$\mathrm{e}^{-2N(x)^2} \to \mathrm{e}^{-2N(x \pm x_0)^2} = \mathrm{e}^{-2N(x^2+x_0^2)}\mathrm{e}^{\mp 4Nxx_0}. \tag{5.5}$$

Including the parameter β describing the observed uneven abundances of A-types and T-types and summing up both terms for x_0 and $-x_0$ yields

$$P(x) \sim (\beta^{\alpha}\mathrm{e}^{-4Nxx_0} + \mathrm{e}^{4Nxx_0})\mathrm{e}^{-2N(x^2+x_0^2)}. \tag{5.6}$$

The A-type bias is clearly visible from 36mers on, so $\alpha = N/12 - 2$. The integral over all A:T-compositions x yields:

$$P(x) = \sqrt{\frac{2N}{\pi}} \left(\frac{\beta^{(\frac{N}{12}-2)} \mathrm{e}^{-4Nx_0x} + \mathrm{e}^{4Nx_0x}}{\beta^{(\frac{N}{12}-2)} + 1} \right) \mathrm{e}^{-2N(x^2+x_0^2)} \tag{5.7}$$

with factors $\sqrt{\frac{2N}{\pi}}(\beta^{\alpha}+1)^{-1}$ for normalization.

5.9.2 Kinetic simulation of 24mer formation from random sequence 12mer pool

In their 2018 publication [118] Tkachenko and Maslov expand their hypothesis from 2015 [119] by a model for the selection of fittest sequences in a random sequence ensemble. The mechanism that is described therein is comparable to a templated ligation reaction. They found a two-stage entropy reduction effect where the first dip in entropy stems from a slightly inhomogeneous ligation rate, depending on the sequence motifs. Selected sequences in the second major reduction in entropy have a vastly smaller sequence space of only $2N$ starting from N^2. This is due to the second part of the effect, where the surviving sequences compete for a limited amount of substrate sequences. For the publication [74] and study shown here, the model was extended to include the possibility of internal hairpin formation. The description for the autocatalytic formation of 24mers, with the concentrations of the products d_{ij} and their "left" and "right" 12mer subsequences l_i and r_j

$$d_{ij} = \lambda(\alpha_{j^*i^*} d_{j^*i^*})(\alpha_i l_i)(\alpha_j r_j) \tag{5.8}$$

now has an additional term d_{j*i*} for the concentration of the complementary 24mer. This strand acts as the sole template for ij in this model. In contrast to the original publication, the ligation rate λ was set to be independent of the sequence. The activity of the respective sequences is decribe by α and depends on the longest internal hairpin l_s for sequences s:

$$\alpha_s = \frac{1}{1 + \mathrm{e}^{-(G_0 + \Delta G l_s)/kT}}. \tag{5.9}$$

$\Delta G \approx 1.5kT$ is the hybridization free energy per base for AT-random strands, G_0 the threshold free energy accounting for the ends of the hybridized regions (assumed to be about $1.5kT$ per side and two times for the formation of the internal loop, in total $G_0 \approx 6kT$). In the simulation the activity and longest internal hairpin for all possible 2^{12} 12mers and 2^{24} 24mers is determined. Equation(5.8) is then solved numerically.

For this simulation an initial seed of 24mers is required to template the reaction. This seed is randomly selected and has a binomial base-composition distribution.

5.10 Sample preparation

All DNA was ordered from *biomers.net* either dried or suspended in MilliQ water. The datasheet from the manufacturer provided information about the concentration and dilution. To ensure similar concentrations across all different samples, the stock solutions of 200 µM DNA in MilliQ was diluted by a factor of 20 and measured on a *NanoDrop Spectrophotometer*, as described in Section 5.4. The 260 nm absorbance line value was recorded and gives the concentration of the sample by dividing it by the extinction coefficient noted in the data sheet.

For the experiments an aluminum metal block with holes to hold reaction tubes is placed on ice acting as a heat-buffer too keep the chemicals at 0 °C during preparation. To ensure the

samples in a single time-series or temperature-series experiment are similar, a mastermix is prepared and then split into smaller volumes in separate reaction tubes. The tubes are then transferred to a PCR cycler. After temperature cycling the PCR machine cools the samples to 4 °C until they are removed and stored in the fridge for further analysis, like PAGE (Section 5.11 and Section 5.6) or NGS (Section 5.7).

5.11 Polyacrylamide gel electrophoresis (PAGE)

For analyzing the dynamics and product yield of the random sequence ligation we use polyacrylamide gel electrophoresis (PAGE) with SYBR gold post-staining. The gels are 15 % acrylamide and are run with 50 % urea in 1x TBE buffer at about 50 °C and, therefore, pose denaturing conditions. The gel is mixed from the *Roth* Rotiphorese DNA sequencing system. One 0.75 mm thick gel with a 15 tooth comb needs about 5 ml gel mixture which contains 3 ml gel concentrate, 1.5 ml gel diluent, 0.5 ml buffer concentrate, 25 µl APS and 2.5 µl TEMED. After 30 min of pre-run at 400 V, the gel pockets are loaded with a total of 4 µl of sample made from 0.89 µl of 10 µM sample and 3.11 µl of 2x loading dye (for about 10 ml add 9.5 ml formamide, 0.5 ml glycerol, 1 µl EDTA (0.5 M), and 100 µl Orange G dye from *New England Biolabs*). The sample is drawn into the gel in a first step of the run with 50 V for 5 min, then the gel electrophoresis is run for about 30 min at 300 V.

After the PAGE run, gels are submersed into 50 ml of 1x TBE buffer with 5 µl of 10.000x *SYBR Gold Nucleic Acid Gel Stain* from *Thermo Scientific* for 5 min. The stained gel is washed in 1x TBE buffer two times and imaged in a *Bio-Rad* ChemiDoc MP System. Analysis of the gel images are done in self-written LabView code (see Section 5.6) or GIMP and inkscape.

5.12 *Illumina* sequencing library preparation

For the library preparation the *Swift Biosciences* Accel-NGS 1S kit is used. The protocol of the manufacturer is followed but only one quarter of each specified volume is used in order to get more samples from the very expensive kit.

There are four basic steps in the library preparation chemistry:

1. **Back-primer attachment:** in a combined step the 3'-end gets phosphorylated and a random sequence of bases `C` and `T` attached by a terminal transferase. The concentration of the DNA relative to the nucleotides-concentration is set to achieve an about 8 nt long CT-tail. In an immediately following constant temperature ligation step the back primer starting with the sequence `AGAT` is attached by templated ligation. The CT-tail acts as the template for the double-stranded primer complex with a dangling end of bases (probably) unspecifically binding to CT-bases (this might simply be a G-overhang).
2. **Strand duplication:** a single cycle PCR reaction builds the reverse complement of the DNA strand with a single `A` overhang on the 5'-end of the initial DNA.
3. **Front adapter attachment:** the front adapter is attached by an additional ligation reaction.
4. **Barcode attachment:** the barcode combinations for identifying the sample after sequencing are attached by a PCR reaction. This reaction is run several times to ensure barcode attachment and sufficient library-DNA concentration.

In between the preparation steps there are several cleaning steps with AMPure XP beads from *Beckman Coulter*. The magnetic beads are stored in a high salt concentration solution which

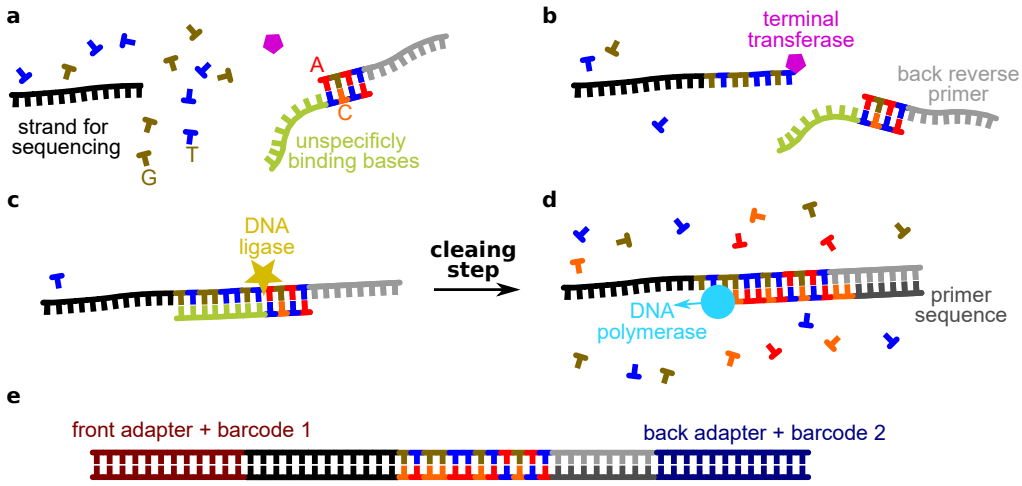

Figure 5.15 ssDNA library preparation sketch:
a ssDNA strands are mixed with the first buffer and reaction educts.
b A terminal transferase attaches a stretch of about eight bases with random sequence `C` and `T` to the 3'-end of DNA strands.
c The double-stranded back-adapter with the starting sequence `AGAT` binds (probably) unspecifically to the random sequence CT-tretch and is ligated. The helper strand is washed away in a subsequent cleaning step.
d Starting at the new adapter sequence a primer is hybridized and extended by a DNA polymerase.
e In subsequent steps a front-adapter is ligated and the front- as well as the back-barcode are added by additional PCR cycles. The resulting strand is added to the flow cell of the sequencing machine, see Section 5.7.

binds DNA to the bead surface. A magnetic tube holder pulls the formerly dispersed beads to the tube wall. The supernatant is carefully removed and the beads are cleaned with a 80 % ethanol 20 % MilliQ solution. A low salt solution is added and the tube vortexed and centrifuged. In the low salt conditions the DNA detaches from the bead surface and goes back into solution. In the final step, the beads are again magnetically pulled to the tube wall and the DNA-sample is pipetted from the tube. Adjusting the ratio of sample-volume to bead-solution volume changes the length-dependent binding affinity of the beads. Low bead concentrations result in a selection for long strands. In the library preparations done for the experiments here, a ratio of 1:2 of sample to beads was used to recover as much short DNA as possible.

5.13 Demultiplexing of sequence data

The *Illuimna* HiSeq sequencing device has a flow cell with several lanes, see Section 5.7. In standard procedures multiple different samples with distinct barcode combinations are run in several lanes to achieve the highest possible variety of bases and sequences per lane. This increases the probability of unique and located sequence clusters in the flow cells that improve read quality. If two strands with resembling sequences form clusters close to each other, the sequencing-by-synthesis approach can have a larger error by mistaking overlapping cluster regions.

The sequencing device produces three FASTQ/ FASTA text-files: a file each for the sequenced front-barcode, back-barcode and the actual sequence for each analyzed strand (-cluster, by bridge amplification, see Section 5.7). The nth entry in each list corresponds to the same sequenced

DNA strand. The demultiplexing software is supplied with a list of the front- and back-barcode pairs of the sample strands. It then reads through the two barcode-files and creates an output file including only the sequences from which the barcodes match the supplied pairs in the before mentioned list (see Section 5.12).

5.14 Filtering and sorting of FASTQ databases

Illumina sequencing devices like the HiSeq system used for the sequence analysis in this thesis, save the respective information in so-called FASTQ/ FASTA files in a text format. Information for a analyzed strand is encoded in four lines:

1. starting with an at-symbol (@) marking the beginning of a new read, followed by a sequence identifier and optionally with additional information *e.g.* read length
2. the measured sequence in capital letters (`A`:adenine, `T`:thymine, `C`:cytosine, `G`:guanine)
3. starting with a plus-symbol (+) and optionally the same sequence identifier as in 1. and additional information
4. the read quality per base encoded in a single ASCII character per base.

Quality encoding starts from ASCII symbol 0x21 (!) and has 94 steps to 0x7e (~):

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNO
PQRSTUVWXYZ[\]^_‘abcdefghijklmnopqrstuvwxyz{|}~
```

The quality of the read for each sequence depends on numerous things:

- The molecule concentration in the flow cell
- The sequence itself; DNA strands with stretches of the same base pose a detection problem for the machine. Read quality can be lower.
- The length of the sequenced strand - the read quality decreases towards a length of 100 bases.

A standard procedure in filtering FASTQ files analyzes if the quality score of each base falls short of a given threshold. In FASTQ files the quality score is not linear but follows a Phred quality score. The quality is defined as

$$Q = -10 \log_{10} P. \tag{5.10}$$

To account for single outliers that might be in a strand with overall high quality reads, a second threshold is used to allow a maximum of "bad bases". In samples with only two bases, as it is the case in the main part of this study, all reads that are read as one of the other two possible bases (in AT-only there should be no reads of `C` or `G`) are necessarily bad bases. Therefore, special filtering tactics can be applied in these cases.

The filtering mechanism, as described in Section 5.15, searches for specific primer sequence motifs that follow a strand with a length of a monomer-multiple. In only analyzing the quality of presumably correctly read bases, it is assured, that even multiple bad reads will not influence

Table 5.1 The quality score in FASTQ files is given with the Phred quality score, as given in (5.10)

Phred Quality Score	Probability of incorrectly read base	accuracy
10	0.1	90 %
20	0.01	99 %
30	0.001	99.9 %

good reads in the remaining strand. In the analysis part, the amount of analyzed bases as function of the position in the strand might vary, as wrong bases don't contribute, but adjoining good reads do.

5.15 Regular expression filtering of sequences

Regular **Ex**pression (RegEx) filtering is a method of finding, cutting and replacing single characters or whole text segments in string type variables. For the analysis of sequence data, this is a very suitable approach: Characters denote the four bases in DNA `A`, `T`, `C` and `G`. A string then consist of a sequence of these characters. Filter operations in RegEx give a lot of freedom and it is possible to combine different operations.

A typical sequence in a results file from the *Illumina* sequencer consist of the sequence, followed by a stretch of `C` and `T` in random sequence, again followed by the back primer (compare Section 5.12). The analysis script utilizes this distinct pattern and first searches for a CT-random-stretch with a minimum length of four, followed by the start of the back primer sequence `AGAT`. In the subsequent operation, the length of the preceding sequence must be a multiple of the monomer length. The RegEx algorithm will look as follows for a read length of 100 bases:

```
(^[ATCG]{12}|[ATCG]{24}|[ATCG]{36}|[ATCG]{48}|[ATCG]{60}|[ATCG]{72}
|[ATCG]{84})(?=([CT]{4,}AGAT))
```

This yields a list of sequences with the correct length. In a second filtering step the actual base content is analyzed. In AT-only samples (or respectively CG-only) a limited amount of bases `C` or `G` (`A`, or `T`) are allowed. In the following example it is two wrong reads (highlighted):

```
^(?!(?:.*?(G|C))2,)^([ATCG]{12,})
```

Resulting sequences are stored in a LabView variant variable, that combines an unique object name (the sequence) and its value (count). The variant variable can be exported as a *.bin* file for resource efficient storage and fast import into the analysis program. The file size for bin-files scales linearly with the amount of different reads.

5.16 Sequencing, demultiplexing and qualityscore filtering error estimation

Generally, the easiest way to asses the error rate in sequencing samples is by comparing a strand to known sequence. This is possible by covalently linking the ligated or polymerized strand to the template strand. By utilizing a hairpin mediated growth mode, where the primer strand and template strand are connected by a single-stranded loop section, elongated strands are already a single complex. Duzdevich *et al.* [34] characterize the ligation of single bases to a self-priming hairpin complex mediated by an intermediate dimer. Because the known template and primer are sequenced as the same strand as the ligated section, they can perform extensive characterization of the error rate and also identify the ligation yield of each single nucleoside. In contrast, the sequences in the random sequence pool can only be treated as a "bulk property": As a baseline, the original pool that did not experience any ligation events can be sequenced. This is shown in Figure 3.5 and as already stated in the corresponding discussion section (Section 4.6), the difference of the 12mer pools before and after the ligation experiment is equivalent to the strands that are preferably ligated and give rise to the product strands. Those exact strands are therefore missing in the NGS data set after 1000 temperature cycles. In several *Illumina* deep se-

quencing runs similar results are achieved, as *e.g.* shown in Figure 3.5. Here, a similar sample, but pipetted, run and sequenced separately shows very comparable features. This suggests that the *Illumina* NGS method is at least robust on the scale of "bulk samples" that are prepared at different times and not sequenced together.

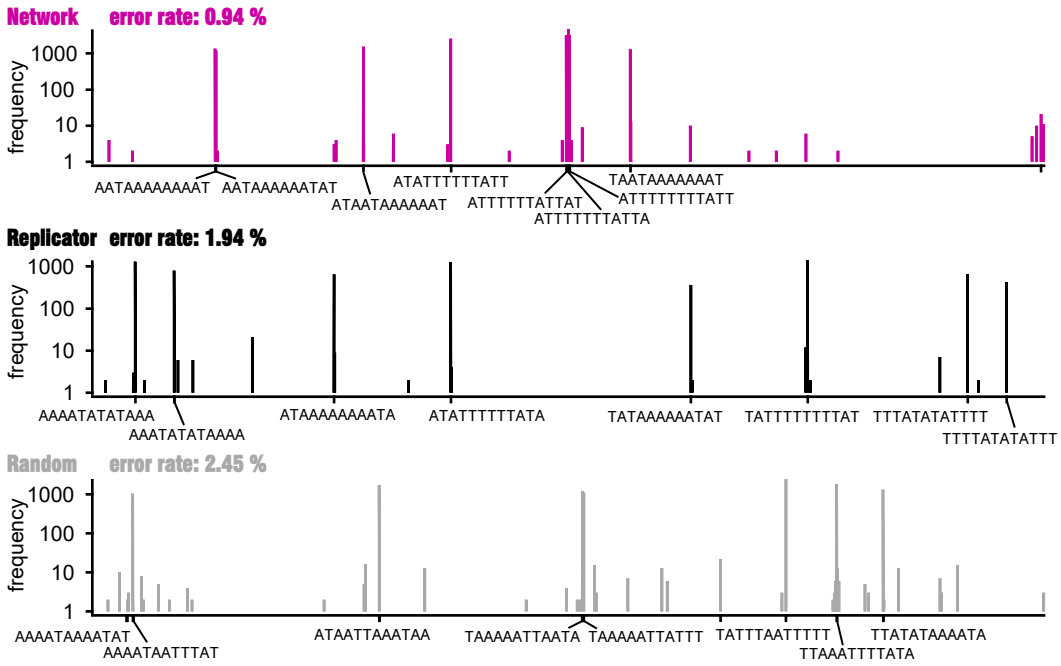

Figure 5.16 AT-only 12 nt sequence motif abundance for the x8-samples:
The correctly sequenced motifs in all three samples `Network`, `Replicator`, and `Random` are clearly distinguishable from the incorrectly read sequences, simply by their abundance. The error rate is between 0.94 % for the `Network`-sample to 2.45 % for the `Random`-sample (incorrectly sequenced motifs/ correctly sequenced motifs). This error rate includes a possible error or bias from sequence synthesis. The y-axis is plotted as in log-scale.

On the other hand the error quantification becomes difficult on the scale of single strands. In contrast to the entire pool, where single incorrectly read strands do not significantly alter the overall sequencing result, there is no baseline for single strands. Because each and every single strand is random (though strands might have a bias at the 3'-end due to the synthesis reaction, see Section 3.3), it's impossible to quantify a per-strand error. A similar limit holds true on the scale of single-bases: While bases `C` and `G` are obvious false reads in an AT-only sample, for a poly-`A` strand the HiSeq device might accidentally read too few or too many single `A` resulting in a wrong read length. The frequency of CG-reads in AT-only samples is also position dependent: a 12mer strand with an unread-base and thus a read length of 11 nt might include `C` at position 12, that was attached as part of the CT-random sequence tail during *Illumina* library preparation. This is a wrong read as a 12mer, but correctly read CT-tail. For quantification of a poly-base read error a specialized separate sample would be needed, like a 12 nt `A`-only for detailed analysis.

Here, the sequencing device provides the before mentioned Phred quality score (see Section 5.14) that estimates the quality of each base in each read by the fluorescent signal during sequencing by synthesis. During filtering, all bases are sorted to be above a defined quality score. Still, some strands have obvious errors despite their high Phred score. This can be observed in detail for the x8-samples `Network`, `Replicator`, and `Random` shown in Section 3.7.2. However, the analysis of all possible AT-only 12 nt motifs for the sequencing baseline without temperature

cycling in Figure 5.16 shows between 0.94 % and 2.45 % incorrectly sequenced strands. While the correctly sequenced strands are easily identified simply by their abundance, several other incorrect strands are sequenced more than once.

A method to quantify the difference between two sequences is explained above in Section 5.8. Calculating the Levenshtein difference of all sequenced strands to all known strands in the x8-samples results in a list of weighted nodes (sequence motifs) and weighted edges (edit distance between the strands). This list can be visualized with the software tool *Gephi* [10] in a network graph, as employed for the AT-random product sequences. In contrast to Figure 3.27, Figure 3.28, or Figure 3.33 which are sorted by hand to highlight the relation of subgroups, the network graphs here utilize a energy-minimization method called *ForceAtlas2* [63]. The algorithm moves the nodes to minimize a system of equations describing the weighted edges in terms of repulsion and attraction of the nodes. Figure 5.17 shows the networks with interactions for Levenshtein distances. Likely due to the large amount of edges and the comparably low amount of nodes (edges/nodes ≈ 8.05) the algorithm does not converge in a reasonable time frame. The plotted networks are shown after ≈ 20 s of runtime. Still, distinct features in each network are visible. For readability purposes, and as discussed later in Figure 5.18, the edges for Levenshtein distances ≥ 4 are removed from the graph. The `Network`-sample clearly shows two groups of motif-clusters, again A-type and T-type. Incorrectly sequenced strands are spatially close to the correct sequence motifs and tend to differ by only one or two Levenshtein-operations (mutations). The `Replicator`-sample also features the A-type and T-type group, but both are subdivided into two groups with either the two sequences with a poly-base motif or the alternating pattern at the center section. The `Random`-sample shows that incorrectly sequenced strands are mostly related to the correct sequences, but the network does not show a clear segregation as for the other samples. `Network`-sample and `Replicator`-sample both obtain function by their autocatalytic or designed selection process. The `Random`-sample has no selected sequence motifs and therefore no clear spacial segregation.

The network graphs suggest, that most incorrectly read sequences are somewhat closely related to the correct strands, as expected. With $leq2$ of the operations included in the Levenshtein metric (substitution, insertion, and deletion) those incorrect reads could still be taken into account for analysis. Figure 5.18 plots the frequency of sequence motifs reached by applying ≥ 1 edit operations starting from the known correct motifs of the three x8-samples. `Network` and `Replicator` both reach over 99.5 % of all analyzed strands with a Levenshtein distance of ≤ 2. From the incorrectly sequenced strands between 76.6 % (`Random`) and 82.4 % (`Network`) differ only in 1 operation from the correct strand. Therefore, most strands are correctly read (more than 97 % of all cases) and incorrectly read strands predominantly differ in only 1 operation. During sequence motif analysis, strands which include single or double `C` or `G` are still analyzed in the region with AT-only motifs, as the probability that those regions are correctly sequenced are very high. It is important to note, that a possible bias from synthesis is not found by this method.

As depicted by the size of the nodes in Figure 5.17, the correct sequences are not equally abundant, as it would be expected for similar concentrations. The eight unique strands were ordered separately, the concentration was measured with the help of the NanoDrop (see Section 5.4) and the sample mixed with equimolar concentration. As seen here, the abundance after sequencing is not homogeneous. Possible errors include the pipetting error before and during NanoDrop analysis, the pipetting error in sample preparation, an errors from synthesis and from sequencing. While it seems like T-type strands are more abundant in the `Network`-sample, A-type strands are more abundant in the `Replicator`-sample. The base on the 3'-end has no prominent influence on the abundance of a sequence, unlike it might be suspected by the high frequency of 3'-end `AT` motifs. The reproducibility of the NGS method suggests that a possible bias in the

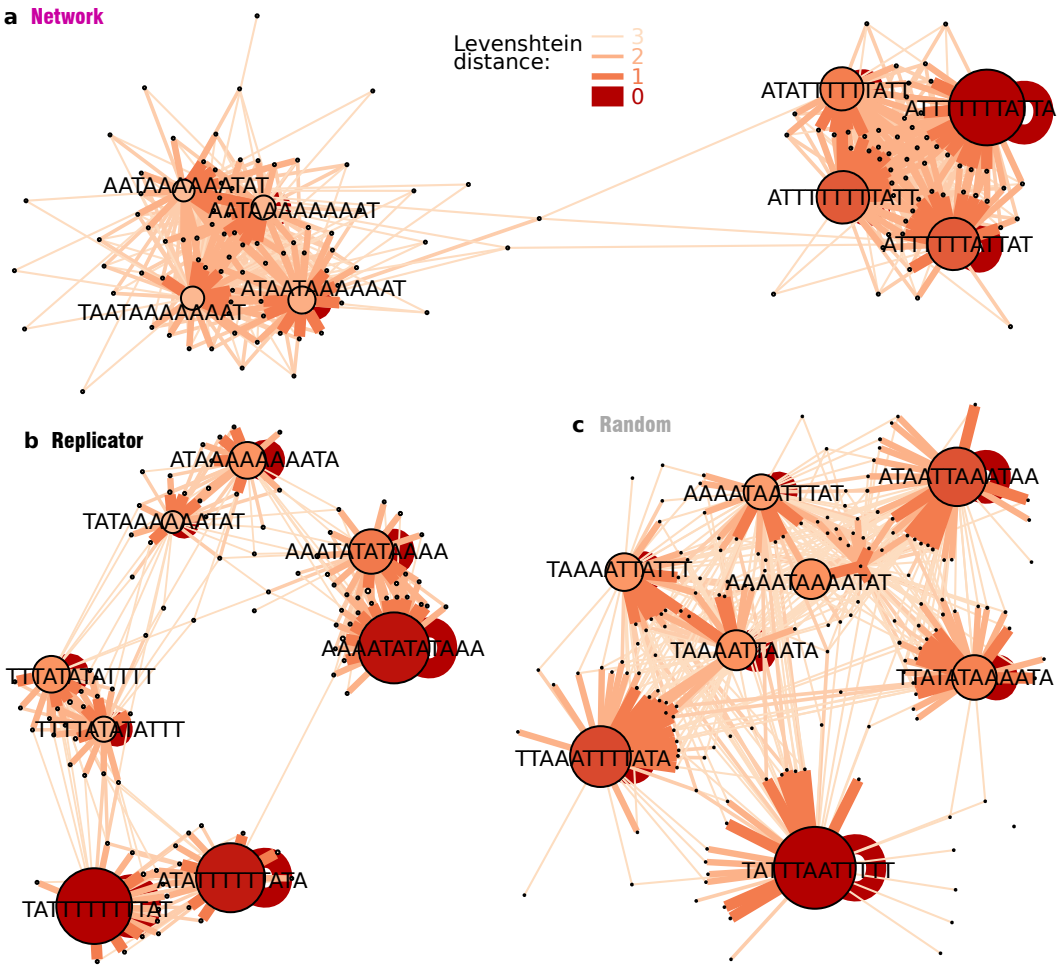

Figure 5.17 ForceAtlas2 sequence motif networks plotting the Levenshtein distance from known motifs:

a `Network`-sample shows two distinct groups, as already known for the emerging sequences themselves, but also for the incorrectly sequenced strands.

b `Replicator`-sample also forms two groups with A-type and T-type, but each with two subgroups: Poly-`T` or poly-`A` at the center of the strand and `ATAT` alternating at the center of the strands.

c `Random`-sample is more loosely connected, because the strands are neither designed (`Replicator`) nor selected from a pool (`Network`) and therefore lack the "connectivity" and inter-templation ability that enables the cluster formation of the other samples.

library preparation and subsequent sequencing is small. The homogeneous binomial A:T-ratio distribution of 12mers (see Section 3.3) further suggests that the assembly of these random sequences is only slightly biased. In contrast, for the concentration estimation of the stock DNA, small volumes need to be pipetted for the dilution prior to the NanoDrop analysis, as well as for the pool stock mix. Here, it is possible that already inconsistent concentrations could unintentionally be enlarged, although the intention of the NanoDrop step is the opposite. The NanoDrop measurement itself is fairly reliable due to the nature of the absorbance method and triplicate measurements. As long as strands are predominantly single-stranded (all single strands in all three x8-samples are very likely single-stranded) and the extinction factor is correct (theoretically calculated for the correct sequence length and composition) the only factors altering the

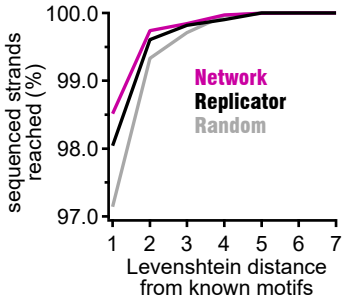

Figure 5.18 Sequence coverage as a function of the Levenshtein distance from known motifs: Plotting the percentage of sequences reached (y-axis) by an increasing cumulative Levenshtein distance(x-axis) from the known eight sequences per sample shows, that the majority of incorrectly read strands are only 1 or 2 Levenshtein-operations different from the known motifs. For `Network` and `Replicator` over 99.5 % of strands only have a maximum edit distance of two. Here, all incorrect sequences are analyzed, including those which include one or two `C` and/ or `G`.

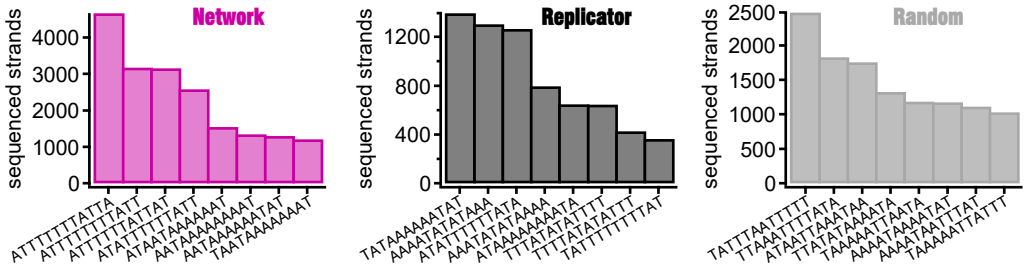

Figure 5.19 Biased pools after NGS of x8-samples: Abundances of the eight known sequences in the x8-samples `Network`, `Replicator`, and `Random`. In the `Network`-sample and the `Random`-sample the most prominent sequence is up to four times more abundant than the rarest sequence.

absorbance are the concentration, or that the included sequences are, at least partially, incorrect. The last point might apply to a certain degree: in the PAGE analysis all samples show an artifact at length 24 nt that probably stems from DNA synthesis, as explained in Figure 5.12. This artifact might skew the absorbance data and lead to false concentration estimation of the stock sample, especially if it differs in the individual components of *e.g.* the x8-samples.

6 Bibliography

[1] E. D. Agerschou, C. B. Mast, and D. Braun. Emergence of life from trapped nucleotides? nonequilibrium behavior of oligonucleotides in thermal gradients. *Synlett*, 28(01):56–63, 2017.

[2] E. C. Anderson, W. F. Libby, S. Weinhouse, A. F. Reid, A. D. Kirshenbaum, and A. V. Grosse. Radiocarbon from cosmic radiation. *Science*, 1947.

[3] R. Appel, B. Niemann, and W. Schuhn. Synthesis of the first triphosphabutadiene. *Angew. Chem. Inf. Ed. Engl.*, 119:932–935, 1986.

[4] A. Arango-Restrepo and J. M. Rubi. The Soret coefficient from the Faxén theorem for a particle moving in a fluid under a temperature gradient. *European Physical Journal E*, 42(5):2–7, 2019.

[5] J. Attwater and P. Holliger. Origins of life: The cooperative gene. *Nature*, 491(7422):48–49, 2012.

[6] J. Attwater, A. Wochner, V. B. Pinheiro, A. Coulson, and P. Holliger. Ice as a protocellular medium for rna replication. *Nature Communications*, 1(6):1–8, 2010.

[7] P. Baaske, F. M. Weinert, S. Duhr, K. H. Lemke, M. J. Russell, and D. Braun. Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proceedings of the National Academy of Sciences of the United States of America*, 2007.

[8] F. Barany. The ligase chain reaction in a pcr world. *PCR Methods Appl*, 1(1):5–16, 1991.

[9] J. A. Baross and S. E. Hoffman. Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins of Life and Evolution of the Biosphere*, 1985.

[10] M. Bastian, S. Heymann, M. Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *Icwsm*, 8(2009):361–362, 2009.

[11] S. a. Benner. Defining life. *Astrobiology*, 10(10):1021–30, 2010.

[12] D. Bikard, C. Loot, Z. Baharoglu, and D. Mazel. Folded dna in action: Hairpin formation and biological functions in prokaryotes. *Microbiology and Molecular Biology Reviews*, 2010.

[13] K. Binder, D. Heermann, L. Roelofs, A. J. Mallinckrodt, and S. McKay. Monte carlo simulation in statistical physics. *Computers in Physics*, 7(2):156–157, 1993.

[14] K. R. Birikh, P. A. Heaton, and F. Eckstein. The structure, function and application of the hammerhead ribozyme. *European Journal of Biochemistry*, 245(1):1–16, 1997.

[15] M. Bjelˇ. Correlation between thermophoretic behavior and hydrophilicity for various alcohols. *Eur. Phys. J. E*, pages 1–7, 2019.

[16] V. Bloomfield and D. M. Crothers. *Nucleic acids: structures, properties and functions*. University Science Books: Sausalito, CA, 2000.

[17] P. D. Boyer. The atp synthase - a splendid molecular machine. *Annu. Rev. Biochem.*, 66:717–749, 1997.

[18] C. Briones, M. Stich, and S. C. Manrubia. The dawn of the rna world: toward functional complexity through ligation of random rna oligomers. *Rna*, 15(5):743–749, 2009.

[19] C. Brochu. Tyrannosaurus rex: The tyrant king. life of the past. edited by peter larson and kenneth carpenter. bloomington (indiana): Indiana university press. *The Quarterly Review of Biology*, 2009.

[20] B. Brown, R. S. Lull, and H. F. Osborn. *Tyrannosaurus and other Cretaceous carnivorous dinosaurs. Bulletin of the AMNH ; v. 21, article 14*. New York: Published by order of the Trustees, American Museum of Natural History, 1905.

[21] T. R. Brummelkamp, R. Bernards, and R. Agami. A system for stable expression of short interfering rnas in mammalian cells. *Science*, 2002.

[22] W. Cao. Dna ligases and ligase-based technologies. *Clinical and Applied Immunology Reviews*, 2(1):33–43, 2001.

[23] W. Cao. Dna ligases: Structure, function and mechanism. *Current Organic Chemistry*, 6(9):827–839, 2002.

[24] P. P. Chan and P. M. Glazer. Triplex dna: fundamentals, advances, and potential applications for gene therapy. *Journal of Molecular Medicine*, 75(4):267–282, 1997.

[25] Q. Chi, G. Wang, and J. Jiang. The persistence length and length per base of single-stranded dna obtained from fluorescence correlation spectroscopy measurements using mean field theory. *Physica A: Statistical Mechanics and its Applications*, 2013.

[26] I. I. Cisse, H. Kim, and T. Ha. A rule of seven in watson-crick base-pairing of mismatched sequences. *Nature Structural and Molecular Biology*, 2012.

[27] F. Crick. The origin of the genetic code. *Journal of Molecular Biology*, 38:367–379, mar 1968.

[28] F. H. Crick and J. D. Watson. The complementary structure of deoxyribonucleic acid. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 223(1152):80–96, apr 1954.

[29] J. Derr, M. L. Manapat, S. Rajamani, K. Leu, R. Xulvi-Brunet, I. Joseph, M. A. Nowak, and I. A. Chen. Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences. *Nucleic Acids Research*, 40(10):4711–4722, 2012.

[30] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM review*, 49(1):65–88, 2007.

[31] R. M. Dirks and N. A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of computational chemistry*, 24(13):1664–1677, 2003.

[32] R. M. Dirks and N. A. Pierce. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *Journal of computational chemistry*, 25(10):1295–1304, 2004.

[33] K. Dose. Peptides and amino acids in the primordial hydrosphere. *BioSystems*, 1975.

[34] D. Duzdevich, C. E. Carr, and J. W. Szostak. Deep sequencing of non-enzymatic rna primer extension. *Nucleic acids research*, 2020.

[35] E. Edeleva, A. Salditt, J. Stamp, P. Schwintek, J. Boekhoven, and D. Braun. Continuous nonenzymatic cross-replication of dna strands with in situ activated dna oligonucleotides. *Chemical Science*, 10(22):5807–5814, 2019.

[36] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, 1971.

[37] M. Eigen, W. Gardiner, P. Schuster, and R. Winkler-Oswatitsch. The origin of genetic information. *Scientific american*, 244(4):88–119, 1981.

[38] M. Eigen and P. Schuster. The hypercyde emergence of the hypercycle. *Die Naturwissenschaften*, 64:541–565, 1977.

[39] E. H. Ekland, J. W. Szostak, and D. P. Bartel. Structurally complex and highly active rna ligases derived from random rna sequences. *Science*, 269(5222):364–370, 1995.

[40] G. Ertem, R. M. Hazen, and J. P. Dworkin. Sequence analysis of trimer isomers formed by montmorillonite catalysis in the reaction of binary monomer mixtures. *Astrobiology*, 7(5):715–722, 2007.

[41] A. Eschenmoser. The search for the chemistry of life's origin. *Tetrahedron*, 63(52):12821–12844, 2007.

[42] A. C. Fahrenbach, C. Giurgiu, C. P. Tam, L. Li, Y. Hongo, M. Aono, and J. W. Szostak. Common and potentially prebiotic origin for precursors of nucleotide synthesis and activation. *Journal of the American Chemical Society*, 139(26):8780–8783, 2017.

[43] S. G. Fischer and L. S. Lerman. Dna fragments differing by single base-pair substitutions. *Proceedings of the Natural Academy of Sience, Biochemistry*, 80(March):1579–1583, 1983.

[44] M. D. Frank-Kamenetskii and S. M. Mirkin. Triplex dna structures. *Annual review of biochemistry*, 64(1):65–95, 1995.

[45] W. Fuller, W. Wilkins, H. Wilson, and L. Hamilton. The molecular configuration of deoxyribonucleic acid: Iv. x-ray diffraction study of the a form. *Journal of molecular biology*, 12(1):60–IN9, 1965.

[46] J. E. Gee and D. M. Miller. Structure and applications of intermolecular dna triplexes. *The American journal of the medical sciences*, 304(6):366–372, 1992.

[47] M. A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *Journal of Physical Chemistry A*, 2000.

[48] W. Gilbert. Origin of life: The rna world. *nature*, 319(6055):618–618, 1986.

[49] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[50] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[51] D. T. Gillespie. Gordon conference, stochastic physics in biology. In *From the Stochastic Simulation Algorithm to the Small-Voxel Tracking Algorithm*, Four Points Sheraton, Ventura, CA, January 2017.

[52] C. A. Haasnoot, S. H. De Bruin, R. G. Berendsen, H. G. Janssen, T. J. Binnendijk, C. W. Hubers, G. A. Van Der Marel2, and J. H. Van Boom. Structure, kinetics and thermodynamics of dna hairpin fragments in solution. *Journal of Biomolecular Structure and Dynamics*, 1983.

[53] G. J. Handschuh, R. Lohrmann, and L. E. Orgel. The effect of mg2+ and ca2+ on urea-catalyzed phosphorylation reactions. *Journal of Molecular Evolution*, 1973.

[54] P. G. Higgs and N. Lehman. The rna world: molecular cooperation at the origins of life. *Nature Reviews Genetics*, 16(1):7–17, 2015.

[55] A. G. Hinnebusch. Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiology and Molecular Biology Reviews*, 75(3):434–467, 2011.

[56] D. P. Horning and G. F. Joyce. Amplification of rna by an rna polymerase ribozyme. *Proceedings of the National Academy of Sciences*, 113(35):9786–9791, 2016.

[57] P. P. Hsu and D. M. Sabatini. Cancer cell metabolism: Warburg and beyond. *Cell*, 134(5):703–707, 2008.

[58] W. Huang and J. P. Ferris. Synthesis of 35–40 mers of rna oligomers from unblocked monomers. a simple approach to the rna world. *Chemical Communications*, (12):1458–1459, 2003.

[59] W. Huang and J. P. Ferris. One-step, regioselective synthesis of up to 50-mers of rna oligomers by montmorillonite catalysis. *Journal of the American Chemical Society*, 128(27):8914–8919, 2006.

[60] N. V. Hud. Searching for lost nucleotides of the pre-RNA World with a self-refining model of early Earth. *Nature Communications*, 9(1):1–4, 2018.

[61] N. V. Hud, B. J. Cafferty, R. Krishnamurthy, and L. D. Williams. The origin of rna and "my grandfather's axe". *Chemistry & biology*, 20(4):466–474, 2013.

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

[63] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6):e98679, 2014.

[64] D. Jost and R. Everaers. A unified poland-scheraga model of oligo-and polynucleotide dna melting: salt effects and predictive power. *Biophysical journal*, 96(3):1056–1067, 2009.

[65] G. F. Joyce. Toward an alternative biology. *Science*, 336(6079):307–308, 2012.

[66] G. F. Joyce and L. E. Orgel. Progress toward understanding the origin of the rna world. *Cold Spring Harbor Monograph Series*, 43:23, 2006.

[67] D. Kacian, D. Mills, F. Kramer, and S. Spiegelman. A replicating rna molecule suitable for a detailed analysis of extracellular evolution and replication. *Proceedings of the National Academy of Sciences*, 69(10):3038–3042, 1972.

[68] L. M. Keil, F. M. Möller, M. Kieß, P. W. Kudella, and C. B. Mast. Proton gradients and ph oscillations emerge from heat flow at the microscale. *Nature Communications*, 8(1):1–9, 2017.

[69] D. S. Kelley, J. A. Karson, G. L. Früh-Green, D. R. Yoerger, T. M. Shank, D. A. Butterfield, J. M. Hayes, M. O. Schrenk, E. J. Olson, G. Proskurowski, M. Jakuba, A. Bradley, B. Larson, K. Ludwig, D. Glickson, K. Buckman, A. S. Bradley, W. J. Brazelton, K. Roe, M. J. Elend, A. Delacour, S. M. Bernasconi, M. D. Lilley, J. A. Baross, R. E. Summons, and S. P. Sylva. A serpentinite-hosted ecosystem: The lost city hydrothermal field. *Science*, 2005.

[70] J. Kim and M. Mrksich. Profiling the selectivity of dna ligases in an array format with mass spectrometry. *Nucleic Acids Research*, 38(1):1–10, 2010.

[71] M. Kinjo and R. Rigler. Ultrasensitive hybridization analysis using fluorescence correlation spectroscopy. *Nucleic Acids Research*, 1995.

[72] J. A. Kozlov and L. E. Orgel. Nonenzymatic template-directed synthesis of rna from monomers. *Molekulyarnaya Biologiya*, 34(6):921–930, 2000.

[73] R. Krishnamurthy. Experimentally investigating the origin of dna/rna on early earth. *Nature Communications*, 9(1):5175, 2018.

[74] P. W. Kudella, A. V. Tkachenko, A. Salditt, S. Maslov, and D. Braun. Structured sequences emerge from random pool when replicated by templated ligation. *accepted in PNAS*, 2021.

[75] I. R. Lehnman. Dna ligase: structure, mechanism, and function. *Science*, 186(4166):790–797, 1974.

[76] L. Li, N. Prywes, C. P. Tam, D. K. Oflaherty, V. S. Lelyveld, E. C. Izgu, A. Pal, and J. W. Szostak. Enhanced nonenzymatic rna copying with 2-aminoimidazole activated nucleotides. *Journal of the American Chemical Society*, 139(5):1810–1813, 2017.

[77] T. Lindahl, P. Karran, and R. D. Wood. Dna excision repair pathways. *Current opinion in genetics & development*, 7(2):158–169, 1997.

[78] G. J. S. Lohman, R. J. Bauer, N. M. Nichols, L. Mazzola, J. Bybee, D. Rivizzigno, E. Cantin, and T. C. E. Jr. A high-throughput assay for the comprehensive profiling of dna ligase fidelity. *Nucleic Acids Research*, 44(2), 2016.

[79] R. Lohrmann. Formation of nucleoside 5′-polyphosphates from nucleotides and trimetaphosphate. *Journal of Molecular Evolution*, 1975.

[80] B. Maddox. *Rosalind Franklin: The dark lady of DNA*. HarperCollins New York, 2002.

[81] N. R. Markham and M. Zuker. Unafold. In *Bioinformatics*, pages 3–31. Springer, 2008.

[82] C. B. Mast and D. Braun. Thermal trap for dna replication. *Physical Review Letters*, 104(18):1–4, 2010.

[83] C. B. Mast, S. Schink, U. Gerland, and D. Braun. Escalation of polymerization in a thermal gradient. *Proceedings of the National Academy of Sciences*, 110(20):8030–8035, 2013.

[84] J.-L. Mergny and L. Lacroix. Analysis of thermal melting curves. *Oligonucleotides*, 13(6):515–537, dec 2003.

[85] S. L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 1953.

[86] D. R. Mills, R. Peterson, and S. Spiegelman. An extracellular darwinian experiment with a self-duplicating nucleic acid molecule. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):217, 1967.

[87] S. Miyakawa and J. P. Ferris. Sequence-and regioselectivity in the montmorillonite-catalyzed synthesis of rna. *Journal of the American Chemical Society*, 125(27):8202–8208, 2003.

[88] F. M. Möller, F. Kriegel, M. Kiess, V. Sojo, and D. Braun. Steep ph gradients and directed colloid transport in a microfluidic alkaline hydrothermal pore. *Angewandte Chemie International Edition*, 56(9):2340–2344, 2017.

[89] M. Morasch, J. Liu, C. F. Dirscherl, A. Ianeselli, A. Kühnlein, K. Le Vay, P. Schwintek, S. Islam, M. K. Corpinot, B. Scheu, D. B. Dingwell, P. Schwille, H. Mutschler, M. W. Powner, C. B. Mast, and D. Braun. Heated gas bubbles enrich, crystallize, dry, phosphorylate and encapsulate prebiotic molecules. *Nature Chemistry*, 11(9):779–788, 2019.

[90] H. Mutschler, A. Wochner, and P. Holliger. Freeze-thaw cycles as drivers of complex ribozyme assembly. *Nature Chemistry*, 7(6):502–508, 2015.

[91] R. M. Myers, S. G. Fischer, L. S. Lerman, and T. Maniatis. Nearly all single base substitutions in dna fragments joined to a gc-clamp can be detected by denaturing gradient gel electrophoresis. *Nucleic acids research*, 13(9):3131–3145, 1985.

[92] L. Nguyen, M. Döblinger, T. Liedl, and A. Heuer-Jungemann. Dna-origami-templated silica growth by sol–gel chemistry. *Angewandte Chemie - International Edition*, 2019.

[93] J. Oró. Mechanism of synthesis of adenine from hydrogen cyanide under possible primitive earth conditions. *Nature*, 1961.

[94] L. E. Orgel. Evolution of the genetic apparatus: A review. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:9–16, 1987.

[95] R. Österberg, L. E. Orgel, and R. Lohrmann. Further studies of urea-catalyzed phosphorylation reactions, 1973.

[96] P. J. Paddison, A. A. Caudy, E. Bernstein, G. J. Hannon, and D. S. Conklin. Short hairpin rnas (shrnas) induce sequence-specific silencing in mammalian cells. *Genes and Development*, 2002.

[97] R. Pascal, A. Pross, and J. D. Sutherland. Towards an evolutionary theory of the origin of life based on kinetics and thermodynamics. *Open biology*, 3(11):130156, 2013.

[98] L. Pauling and R. B. Corey. Structure of the nucleic acids. *Nature*, 1953.

[99] A. Poghossian, A. Cherstvy, S. Ingebrandt, A. Offenhäusser, and M. J. Schöning. *Possibilities and limitations of label-free detection of DNA hybridization with field-effect-based devices*. Elsevier, 2005.

[100] V. Presnyak, N. Alhusaini, Y.-H. Chen, S. Martin, N. Morris, N. Kline, S. Olson, D. Weinberg, K. E. Baker, B. R. Graveley, et al. Codon optimality is a major determinant of mrna stability. *Cell*, 160(6):1111–1124, 2015.

[101] A. D. Pressman, Z. Liu, E. Janzen, C. Blanco, U. F. Müller, G. F. Joyce, R. Pascal, and I. A. Chen. Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for self-aminoacylating rna. *Journal of the American Chemical Society*, 141(15):6213–6223, 2019.

[102] A. Priye, Y. Yu, Y. A. Hassan, and V. M. Ugaz. Synchronized chaotic targeting and acceleration of surface chemistry in prebiotic hydrothermal microenvironments. *Proceedings of the National Academy of Sciences*, 114(6):1275–1280, feb 2017.

[103] B. K. Rima and N. V. McFerran. Dinucleotide and stop codon frequencies in single-stranded rna viruses. *Journal of General Virology*, 78(11):2859–2870, 1997.

[104] J. Rosenberger, T. Göppel, P. W. Kudella, D. Braun, U. Gerland, and B. Altaner. Understanding templated ligation: extension cascades, persisting complexes and emergent length scales. *under review*, 2020.

[105] A. Salditt, L. M. Keil, D. P. Horning, C. B. Mast, G. F. Joyce, and D. Braun. Thermal habitat for rna amplification and accumulation. *Physical Review Letters*, 125(4):048104, 2020.

[106] A. Sancar. Dna excision repair. *Annual review of biochemistry*, 65(1):43–81, 1996.

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

[108] I. Schoen, H. Krammer, and D. Braun. Hybridization kinetics is different inside cells. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21649–21654, 2009.

[109] G. K. Schroeder, C. Lad, P. Wyman, N. H. Williams, and R. Wolfenden. The time required for water attack at the phosphorus atom of simple phosphodiesters and of dna. *Proceedings of the National Academy of Sciences of the United States of America*, 2006.

[110] W. G. Scott, J. B. Murray, J. R. P. Arnold, B. L. Stoddard, and A. Klug. Capturing the structure of a catalytic rna intermediate: The hammerhead ribozyme. *Science*, 274(5295):2065–2069, 1996.

[111] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[112] S. Shuman and C. D. Lima. The polynucleotide ligase and rna capping enzyme superfamily of covalent nucleotidyltransferases. *Current opinion in structural biology*, 14(6):757–764, 2004.

[113] D. Sievers and G. Von Kiedrowski. Self-replication of hexadeoxynucleotide analogues: Autocatalysis versus cross-catalysis. *Chemistry - A European Journal*, 4(4):629–641, 1998.

[114] J. Summers and S. Litwin. Examining the theory of error catastrophe. *Journal of virology*, 80(1):20, 2006.

[115] J. Swetina and P. Schuster. Self-replication with errors: A model for polvnucleotide replication. *Biophysical chemistry*, 16(4):329–345, 1982.

[116] J. W. Szostak. The eightfold path to non-enzymatic rna replication. *Journal of Systems Chemistry*, 3(1):2, 2012.

[117] I. D. Technologies. Oligonucleotide stability study. `https://sfvideo.blob.core.windows.net/sitefinity/docs/default-source/technical-report/stability-of-oligos.pdf?sfvrsn=c6483407_10`, 2014.

[118] A. V. Tkachenko and S. Maslov. Spontaneous emergence of autocatalytic information-coding polymers. *The Journal of Chemical Physics*, 143(4):045102, jul 2015.

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

[120] M. J. Tobin. April 25, 1953: Three papers, three lessons, 2003.

[121] A. E. Tomkinson, S. Vijayakumar, J. M. Pascal, and T. Ellenberger. Dna ligases: structure, reaction mechanism, and function. *Chemical reviews*, 106(2):687–699, 2006.

[122] S. Toyabe and D. Braun. Cooperative ligation breaks sequence symmetry and stabilizes early molecular replication. *Physical Review X*, 9(1):011056, 2019.

[123] D. H. Turner and D. H. Mathews. Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, 38(suppl_1):D280–D282, 2010.

[124] N. Vaidya, M. L. Manapat, I. A. Chen, R. Xulvi-Brunet, E. J. Hayden, and N. Lehman. Spontaneous network formation among cooperative rna replicators. *Nature*, 491(7422):72–77, 2012.

[125] E. A. Venczel and D. Sen. Synapsable dna, 1996.

[126] G. Walter. The rna world. *Nature*, 319:618, 1986.

[127] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 1953.

[128] F. Westheimer. Biochemistry: Polyribonucleic acids as enzymes. *Nature*, 319(6054):534–536, 1986.

[129] J. G. Wetmur and N. Davidson. Kinetics of renaturation of dna. *Journal of Molecular Biology*, 1968.

[130] N. Williams. Dna hydrolysis: mechanism and reactivity. In *Artificial Nucleases*, pages 3–17. Springer, 2004.

[131] J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce. Nupack: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry*, 32(1):170–173, jan 2011.

[132] T. Zhang, C. Hartl, K. Frank, A. Heuer-Jungemann, S. Fischer, P. C. Nickels, B. Nickel, and T. Liedl. 3d dna origami crystals. *Advanced Materials*, 2018.

[133] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research*, 31(13):3406–3415, 2003.

7 Acknowledgements

Thanks Dieter, you have been a great mentor and guided me to the scientist and person I am today. I learned how sometimes one has to just go do something without worrying too much beforehand. The cooperation on the publications, the data as well was the cooperation with Alexei and Sergei have taught me so much about good scientific practices and working together. I really enjoyed my project and all tasked involved: prototyping, measuring, analyzing, scripting, troubleshooting, gathering data and finally writing everything down as publications and this thesis.

Thanks Braun-Lab! I mean, without a nice lab the dissertation is as dull as sitting in a cellar. Coffee in the morning, lunch at noon, then a kicker-break were as important to me as the scientific discussion we had. Special thanks goes to the colleagues in the inner circle, who were always in early and experience that unstressed Morning-Patta. With Annalena, Alex, Thomas, Max and Chrissy I did not only find colleagues, but friends I was happy to see on Monday mornings after weekends and vacations. And in long, dark Corona-winter-mornings the occasional morning-coffee-zoom-session could lift the mood significantly. I am sure I will be seeing all of them often in the years to come.

I also want to highlight the colleagues with whom I worked closely on the projects. Thanks to Annalena Salditt, Alan Ianeselli, Joachim Rosenberger, Tobi Göppel and Bernhard Altaner, who all had a profound impact on the results and the insights that were gained on this project and thesis. An additional special-thanks goes to Alexandra Kühnlein for proof-reading my thesis and Julian Stein for extended conversations during the learning phase for the oral exam!

Thanks Christof! I wish every lab had a postdoc like you. You are such a vital piece in everyday operation, know everything, handle everything and you're a friend on top. I'm sure I'll be talking to you, whether it's about a new laptop, bike, camera or 3D printer.

Thanks Ela! Starting as a long-term girlfriend and finishing my PhD as my wife, you've always been helpful in either pushing me to go back to work, or lending me an interested ear to spout my ideas into. And even though you claim to not have gotten a lot of what I was brabbeling about, I'm sure some important ideas would not have happened without you.