Fakultät für Physik
Technische Universität München

Self-assembly of informational polymers via templated ligation

Joachim H. Rosenberger

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Friedrich C. Simmel
Prüfende der Dissertation:
1. Prof. Dr. Ulrich Gerland
2. Prof. Dr. Dieter Braun

Die Dissertation wurde am 17.11.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 21.01.2022 angenommen.

Abstract

Understanding how systems replicating information can emerge from the interaction of informational polymers (DNA, RNA) is crucial to comprehend the origins of life on Earth. System-level properties arising from the interaction of informational polymers must be identified and understood to reveal the underlying mechanisms that lead to the emergence of these replicating systems. To this end, we developed a stochastic simulation capable of simulating the reactions of thousands of strands.

We start with a conceptional embedding of the herein presented research in the field by considering a hypothetical protocell. We then discuss the properties and representations of informational polymers and their energy models. Thereafter, we describe in detail the group of models that the simulation can, in principle, represent. It should also become apparent how the domain of representable models could be extended. Following that, the theoretical background of the stochastic simulation is presented in detail.

We then pursue a statistical physics approach, abstracting and coarsening the interaction details to find universal rules governing systems of interacting informational polymers. In an initial study, we focus on the self-assembly of informational polymers via templated ligation, which is the reaction of two strands becoming covalently linked while being hybridized on a third strand (template). Despite its importance for prebiotic information processing, a systematic study of self-assembly and strand growth via templated ligation has been missing. In this work, we use a simple model to study the problem from first principles. We assume the binding energy of two strands to be simply proportional to their overlap length, while the sequence dependence is ignored. We find that the competition of the length-dependent time scale of dehybridization with the time scales of other dynamical processes leads to a non-monotonous strand length distribution, expressing a minimum and a maximum: While short strands dehybridize fast, longer strands stay hybridized long enough for the extension with another strand. Above a first characteristic strand length, L^*, extension cascades rapidly turn duplexes with overhang into fully-hybridized double strands without overhang. When the length of these double strands reaches a second length scale, $L \geq L^\dagger$, their strong binding renders them inert on a transient or degradation time scale given by, e.g., an outflux rate. The dynamics of strands of length $L^* < L < L^\dagger$ are dominated by auto- and heterocatalytic cycles, which lead to the increase of the length distribution between minima and maxima. An analytic theory allows the prediction of these characteristic length scales from elementary model parameters.

We further demonstrate the emergence of the non-monotonous length distributions and the validity of our predictions in an experimental system using DNA strands with random sequences from a binary alphabet (A,T). These experiments were conducted by Patrick Kudella in Dieter Brauns's Lab at the Ludwig Maximilian University of Munich.

Subsequently, I present the first results obtained with a sequence-dependent energy model and give an outlook on future work.

Zusammenfassung

Um den Ursprung des Lebens zu entschlüsseln, ist es notwendig zu verstehen, wie aus der Interaktion von Informationspolymeren (RNA/DNA) informationserhaltende Systeme emergieren können. Um wiederum die zugrunde liegenden Mechanismen zu verstehen, welche zur Emergenz eines solchen informationserhaltenden Systems führen, müssen zunächst makroskopische Eigenschaften identifiziert werden, welche sich aus der Interaktion von Informationspolymeren ergeben.

Zu diesem Zweck haben wir eine stochastische Simulation entwickelt, welche es ermöglicht, die Reaktionen zwischen Tausenden von Komplexen, welche sich durch Hybridisierung von Strängen zusammensetzen, zu simulieren.

Wir beginnen mit der konzeptionellen Einbettung unserer Arbeit in das Forschungsfeld anhand der Betrachtung einer hypothetischen Protozelle. Im Anschluss werden die Eigenschaften, Repräsentationen und Energiemodelle von Informationspolymeren dargestellt. Daraufhin beschreiben wir, welche Modelle prinzipiell durch die Simulation abgebildet werden können. Es sollte durch diese Betrachtung ebenfalls evident werden, wie die Menge an abbildbaren Modellen durch Modifikation der Simulation erweitert werden könnte. Anschließend diskutieren wir den konzeptionellen Aufbau der Simulation im Detail.

Wir verfolgen einen „Statistischen Physik“ Ansatz und abstrahieren und reduzieren die Interaktionsdetails, um universelle Regeln, welche diesen Systemen zugrunde liegen, zu identifizieren. In einer ersten Untersuchung fokussieren wir uns auf die Selbstassemblierung von Informationspolymeren via templierter Ligation. Templierte ligation ist eine Reaktion bei der zwei Stränge, welche nebeneinander auf einem dritten Strang, dem Templat, gebunden sind, kovalent verbunden werden. Trotz der Wichtigkeit im Hinblick auf die präbiotische Informationsprozessierung fehlte bisher eine systematische Untersuchung von Selbstassemblierung und Strangwachstum via templierter Ligation in der Literatur. Um diese Fragestellung ausgehend von einfachen Grundprinzipien zu untersuchen, benutzen wir ein einfaches Modell, welches eine analytische Beschreibung der Resultate zulässt. Wir nehmen an, dass die Bindeenergie zweier hybridisierter Stränge proportional zu deren Überlapplänge ist, wobei wir eine Sequenzabhängigkeit vernachlässigen.

Es zeigt sich, dass die Kompetition der längenabhängigen Zeitskala der Dehybridisierung mit der Zeitskala der übrigen dynamischen Prozesse zu einer nicht monotonen Längenverteilung führt: Während kurze Stränge schnell dehybridisieren, bleiben längere Stränge lange genug hybridisiert, um mit einem anderen Strang verlängert zu werden. Ab einem gewissen Schwellenwert der Stranglänge, L^*, wandeln sogenannte Extensionskaskaden einen Duplex in einen vollhybridisierten Duplex ohne Überhang um. Wenn die Länge dieser Doppelstränge einen zweiten Schwellenwert, $L \geq L^\dagger$, erreicht, werden diese Duplexe durch ihre hohe Bindeenergie inaktiv im Bezug auf eine weitere Transiente oder degradierende Zeitskala, welche z. B. durch eine Ausflussrate gegeben sein kann. Die Dynamik der Stränge der Länge $L^* < L < L^\dagger$ wird bestimmt durch auto- und heterokatalytische Zyklen, welche die ansteigende Flanke in der Stranglängenverteilung zwischen Minima und Maxima verursachen. Die Werte für die Schwellenwerte (charakteristische Längen) können analytisch aus elementaren Modellparametern abgeschätzt werden.

Wir demonstrieren die Emergenz der durch Computersimulationen entdeckten, nicht

monotonen Längenverteilung und die Validität unserer Vorhersagen in einem Laborexperiment. Das Experiment wurde mit DNA Strängen mit randomisierten Sequenzen basierend auf einem binären Alphabet (A,T) von Patrick Kudella im Labor von Dieter Braun an der Ludwig-Maximilians-Universität durchgeführt.

Nachfolgend präsentiere ich erste Resultate eines Systems mit einem sequenzabhängigen Energiemodell und gebe einen Ausblick im Hinblick auf weiterführende Studien.

Contents

[1] These section builds upon the corresponding section in [106]

1. Introduction

The universe started 13.8 billion years ago in a highly symmetric and hence homogeneous state. To our belief, this point is the origin of a set of fundamental rules governing this universe. The interplay of these elementary rules leads to the emergence of larger, more complex structures, such as atoms, which themselves lead to the formation of larger structures and the introduction of new rules (chemistry) governing their interactions. One could conclude that since the initial rules were set, an enormous self-assembly process started, consequently leading to the formations of galaxies, solar systems, planets, etc., whose trajectories in space and time are described by the rules of general relativity. Apparently, on at least one of these planets, the self-assembly led to the emergence of a new process called evolution. We refer to this event as the origins of life. Evolution is based on the mutation of a genetic code of individuals, which leads to an advantage or disadvantage with respect to their ability to reproduce.

It is still an unsolved question how the first system undergoing evolution could emergence on earth.

Evolution requires individuals with an information storage (genetic code) that couples to the fitness of individuals. The genetic code is assumed to be stored in heteropolymers (RNA,DNA,XNA,TNA, etc.), which we call informational polymers or simply a strand. The residues of the strands are so-called nucleotides.

The selection of advantageous modifications of the genetic code (mutations), with respect to the ability of creating more (almost identical) offspring, requires individuals.

Individuals must be kept constantly out of chemical equilibrium with their environment, as they would otherwise dissolve in the surrounding solvent, which requires the constant consumption of energy [38] in the form of heat and work while releasing entropy. Hence at the origins of life, the following premises must come together:

- compartmentalization (individuals)
- copying of genetic information with a sufficient fidelity
- coupling of genetic code to fitness
- ability of individuals to consume energy

1.1. Protocell

We can speculate about the function of a protocell that combines all these requirements *cf.* [107] (see Figure 1.1). We consider a vesicle formed by a fatty acid bilayer. The vesicle is embedded in a rich medium that provides the building blocks needed for the assembly and growth of the protocell, such as short RNA oligonucleotides (monomers, dimers, and trimers) and fatty acids. The medium must also provide the chemistry to constantly activate (or supply activated) oligonucleotides.

The membrane encloses longer RNA strands (genome), which cooperatively form a copy cycle that enables information replication. The genome is trapped inside the vesicle as "oligonucleotides of four residues or longer cannot cross fatty acid membranes" [61]. The

short oligonucleotides supplied by the environment diffuse through the cell membrane into the protocell, where they are consumed by the copy cycle.

The genome must now somehow couple to cell growth and division. There is a multitude of possibilities for how this coupling could be realized in detail.

An indirect coupling could be the control of membrane properties through the genome, enabling faster integration of fatty acids and hence cell growth. Thereby the protocell is thought to grow up to a critical size needed for division.

A direct coupling would be the increase in RNA concentration inside the vesicle, which leads to osmotic swelling, which again leads to more incorporation of fatty acids into the membrane [51]. In this case, a swollen vesicle grows at the expense of neighboring vesicles with lower osmotic pressure. A genome expressing a faster copy cycle, incorporating mass at a faster rate, would thus lead to a faster-growing cell. However, spherical osmotic swollen vesicles turned out to be hard to divide [87].

In turn, abundant addition of fatty acids into the existing vesicle membrane leads to the formation of long filamentous vesicles that are divided easily by agitation similar to the pearling instability [87]. A significant advantage of this mode of division is that the volume to the surface area, which is broken upon division of a thread-like filament, is relatively small. We can therefore expect the loss of encapsulated material to be small upon division compared to dividing a spherical vesicle *e.g.*, via extrusion, see [51].

The required abundant incorporation of fatty acids into the membrane can be achieved via the addition of phospholipids. Hence a protocell containing more phospholipids in its membrane will grow at the cost of its neighbors containing fewer phospholipids [87].

Hence an inheritable advantage could be created by the coupling of the copy cycle to the production of an enzyme that is able to catalyze the production of phospholipids.

1.2. Self-replicating network of informational polymers

The concept of a self-replicating network consisting of informational polymers might seem simple at first glance, but it might be the most challenging open question considering the emergence of life. A comprehensive list of problems that need to be overcome can be found in [61].

A premise for informational polymers to form a self-replicating network is their ability to hybridize and form double-stranded complexes where the stability is strongly sequence-dependent. Thereby the binding energy dependce on the sequences of both strands in and at the vicinity of the binding site [83, 52, 32, 23], but typically grows with the number of paired nucleotides. Due to Watson–Crick base pairing, informational polymers preferentially hybridize onto strands with complementary sequences. This selective recognition of complementary strands constitutes the basis of biological information storage both in modern cells as well as in prebiotic scenarios [34, 71, 16, 31, 3, 61, 94, 55].

In order to assemble copies of the genome, strands have to grow by forming noncovalent bonds with other strands[1]. Throughout this work, we refer to this covalent bonding as *ligation*. We refer to *random ligation* as the ligation of two unhybridized strands in solution. In contrast *templated ligation* [58, 64, 47, 12] occurs when two strands are adjacently hybridized on a third strand.

The combined interaction of strands inside the protocell via hybridization, dehybridization, and ligation must somehow form a self-replicating network.

[1] Monomers are seen as strands of length one

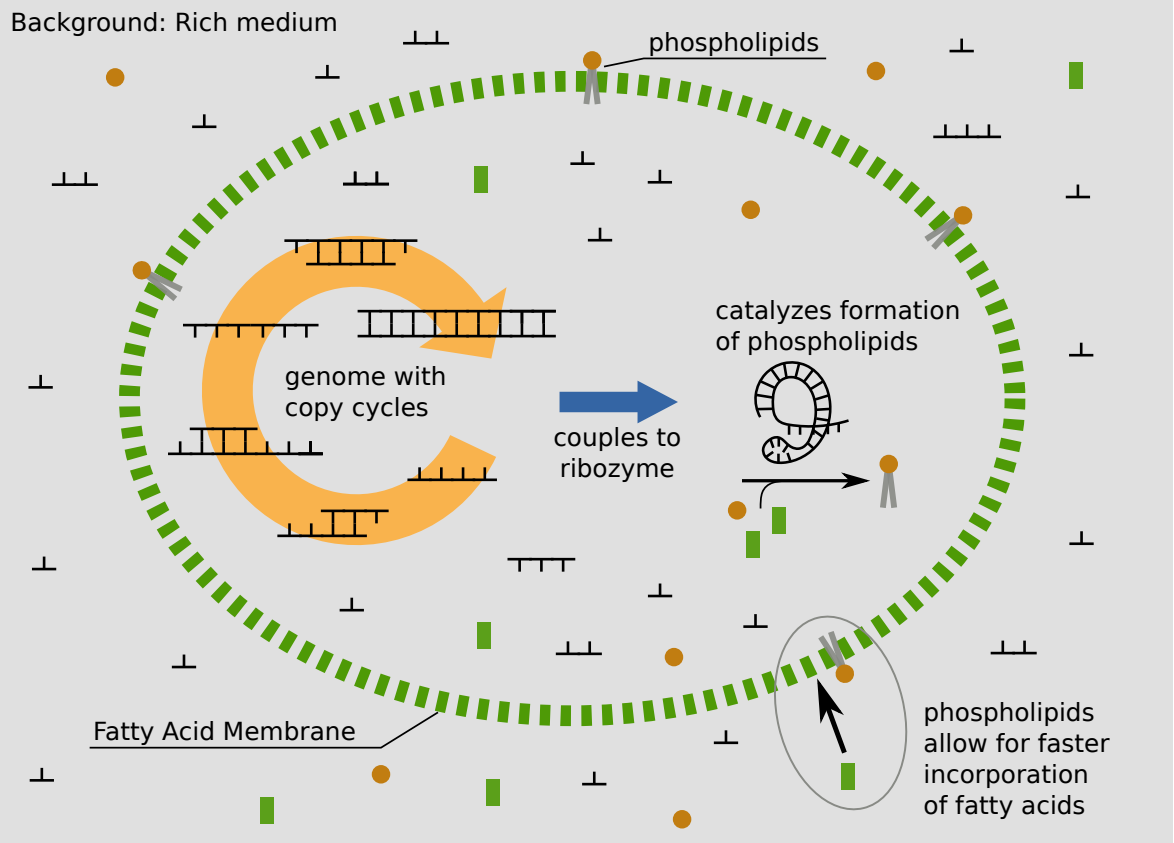


Figure 1.1.: Conceptual protocell where copy cycles favor the creation of a ribozyme which catalyzes the formation of phospholipids which again increases the incorporation rate of fatty acids into the membrane. This increases the growth rate of the cell membrane leading to a faster cell division. Hence a coupling between the genome and a higher reproduction rate can be envisioned, leading to the onset of evolution.

So far most experimental attempts were restricted to the study of a single process of the whole network: The (partial) copying of a strand in a so called primer-extension scenario, see [99, 90, 53, 72, 45, 49, 79, 86], with short oligonucleotides (building blocks). Thereby a strand (primer) is assumed to be hybridized stably [2] onto a longer strands (template) such that there is a 5′ overhang. Short oligonucleotides contained in the solution are thereby assumed to be in (competitive) binding equilibrium with the primer-template duplex, which allows combining hybridization and subsequent ligation into an effective extension process. Also, special cases such as the influence of a so-called helper strand creating a pocket for the hybridization of nucleotide building blocks were considered. These experiments, even though they consider only one process of the whole network, are of great complexity: As mentioned above, hybridization is sequence-dependent, but also ligation depends on the sequence of the building block and the sequence of the primer-template duplex at the ligation site. The leaving group used for the activation of the nucleotides also influences the ligation rate (also sequence-dependent). But even the exact process of how the nucleotides were activated, such as the pH during activation, influences the results. The latter is due to the formation of di-nucleotides, wherein the process of ligation a complete nucleotide serves as a leaving group [81].

However, focusing on understanding every detail of a single process of the reaction network will not allow us to discover possible emergent phenomena that can arise in a self-replicating network of informational polymers. Work on the self-assembly of complexes starting from an initial set of short building blocks has (with some notable theoretical [36] and experimental [105, 100] exceptions) been limited. To my knowledge, there was no conclusive study including the aspect of the initial self-assembly of small building blocks into stable structures. The most reasonable approach is to start with simple interaction rules between strands, such as *e.g.*, : (i) assuming the binding energy of hybridization being proportional to the overlap length instead of considering a sequence-dependent energy model (ii) assuming a constant context-independent templated ligation rate (iii) restricting the number of possible hybridization configurations. After understanding the emergent phenomena arising in these simplified systems, one can progress by, e.g., considering a binary alphabet energy model or connecting systems of different physical control parameters via in and outflux, etc.

1.2.1. State of the art of theoretical studies [‡]

Previous theoretical work studying templated ligation was mostly based on effective models. Usually, the description of the state space has been reduced to strand lengths, without taking into account the hybridization complexes explicitly [78, 74, 89, 47, 82, 70, 92, 98, 95, 88, 8]. In such a coarse-grained picture, (de-)hybridization and templated ligation are not elementary reactions but are combined into an effective extension reaction. In order to specify the corresponding rate r_{ext}, the intricacies of the assembly process are neglected and *a priori* assumptions regarding the relevant configurations are made [70, 78, 82, 89, 47, 95]. Many models (implicitly or explicitly) neglect the dependence of the binding energy on the number of paired nucleotides, *i.e.*, the length of the hybridization site [70, 78, 82, 89, 47, 95, 88, 98]. Others consider a length-dependent dehybridization rate only up to some cut-off length such that the time scale of ligation is always much larger than the time scale of the dehybridization kinetics [74]. A study addressing the full complexity of the assembly process was limited by small system-sizes [36].

[2] Stable in the sense that the probability of the primer dehybridizing during the experiment is negligible.
[‡] These section builds upon the corresponding section in [106]

In this thesis, the emergent phenomena of oligonucleotide self-assembly and growth via templated ligation is studied from first principles. While an *ab initio* scenario circumvents the need for *a priori* assumptions, it is computationally demanding. In particular, a vast combinatorial complexity arises as the number of different complexes grows rapidly with the overall nucleotide mass in the system. This required the development of a new simulation tool able to handle the complexity of such systems.

1.3. Structure of this work

In the next Chapter 2 we will give a more detailed description of informational polymers and will allow us to understand the so-called nearest neighbor energy models. We will then discuss the possible class of systems that can be modeled by the simulation framework that we created in Chapter 3. Next we present the theoretical background of the simulation framework Chapter 4 followed by a technical outline of the actual implementation in Chapter 5. We will then present what we call the null-model for studying emergent phenomena in systems on informational polymers undergoing templated ligation, see Chapter 6. This model assumes a binding energy that is solely proportional to the length of the hybridization site but not sequence-dependent. Then we present first preliminary results obtained by a more advanced binary alphabet energy model in Chapter 7, which showcases the expandability of our simulation framework. In the last Chapter 8 we are going to discuss possible extensions of the simulation and follow-up investigations.

2. Informational polymers

Nucleotides of a particular alphabet $\mathcal{A}$ can be linked covalently, thereby forming a sequence. The letters of the alphabet are called bases. The nucleotides are directional, and the linkage preserves the directionality and thereby transfers it onto the sequence. The directionality determines the direction from which the encoded information of the sequence must be read. We call a directional sequence a strand. Two strands aligned in opposite directions brought into the vicinity of each other can form non-covalent bonds, thereby forming a double-stranded structure. The stability of the double-stranded structure depends on the sequences of the two aligned strands. For each letter $L \in \mathcal{A}$ there is a complementary letter $L^* \in \mathcal{A}$ in the sense that if a double-stranded structure consists of a succession of complementary base pairs (bp), notated as $L.L^*$, the double-strand can adopt a particularly stable form called a helix. The stability of the helix is due to the non-covalent interaction of the nucleotides. We will generally refer to the helix formation of two strand segments as hybridization. If the above criteria are fulfilled, the strand is called an informational polymer.

2.1. Nucleic acids

We kept the definition of information polymers general as, in principle, there can be other informational polymers than the biologically relevant Nucleic acids DNA, and RNA.

Nucleic acids found in organisms have in general an alphabet of 4 letters $\mathcal{A} = \{A, G, C, U\ (T)\}$, where the letters stand for (A)denine, (G)uanine, (C)ytosine, (U)uracile, (T)hymine (see Figure 2.1 (a)). In biological system, RNA contains almost exclusively the base U instead of T, which applies vice versa for DNA. A is complementary to U (T), and G is complementary to C. The complementary letters can form so-called Watson-Crick base pairs $G.C, A.U/T$, which results in the canonical A and B helix of RNA and DNA, respectively.¨ In the case of RNA also U and G are partially complementary, forming a so-called wobble base pairs $G.U$ (see [25]). The $G.U$ wobble base pairs of RNA is relatively to the Watson Crick base pair $A.U$ more stable than $G.T$ is to $A.T$ in DNA [52, 32], which explains why wobble base pairs are associated with RNA and not DNA.

The four letters encode via motifs of length 3 (codons) for all amino acids required to build the proteins used in all organisms. However, also informational polymers of more than the four nucleotides used by nature can be created. For example the Hachimoji nucleic acids consist of an 8 letter alphabet $\mathcal{A} = \{A, G, C, U, P, Z, B, S\}$, where additionally P is complementary to Z and B is complementary to S. It is standing to reason that the selection of the four biological relevant nucleotides is either a consequence of their abundance at the origins of life or/and a result of evolution where the four nucleotide alphabet turned out to be optimal.

The covalent bonds linking nucleotides are phosphodiester bonds between the 5′ carbon and the 3′ carbon of two adjacent nucleotides, see Figure 2.1 (b). The directionality of the strand as a whole is a result of the directionality of the nucleotides themselves. The stability of the helix is mainly due to two types of non-covalent bonds: (base-pairing) hydrogen bonds between the nucleotides within a base pair and (base-stacking) π-π stacking of the

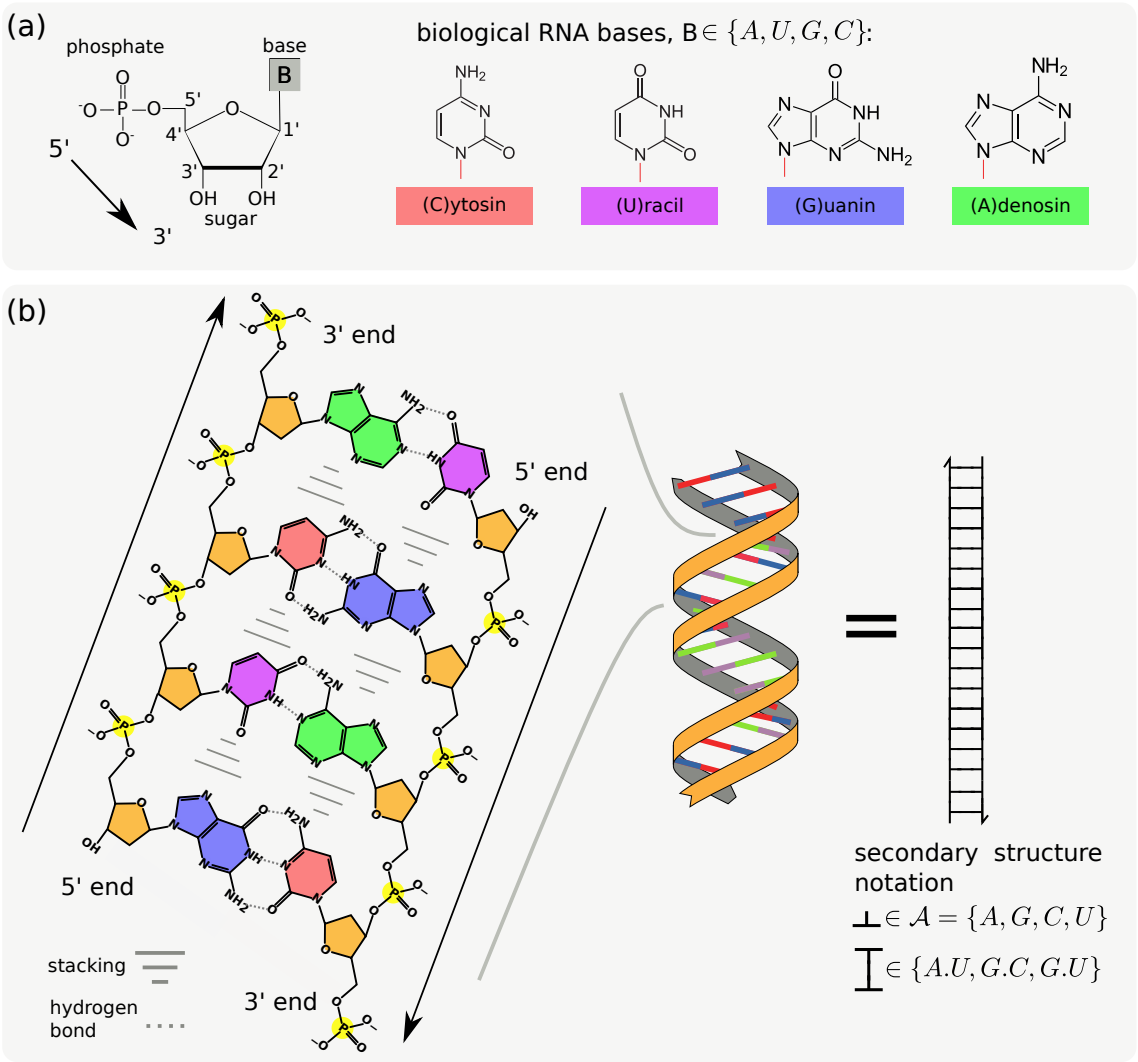


Figure 2.1.: (a) (left) chemical structures of a RNA nucleotide consisting of a phosphate, a D-ribose (sugar) and a base $B \in \mathcal{A}$. The carbons on the sugar are labeled 1′ to 5′, where 5′ to 3′ is used to notate the directionality of the nucleotide. (right) The bases C and U are pyrimidines (one ring of size 6, 4 carbons, and two nitrogen atoms), and G and A are purines (one pyrimidine plus a carbon and two nitrogen atoms forming a second ring of size 5). **(b)** The nucleotides are linked via phosphodiester bonds transferring their directionality onto the resulting strand. Matching nucleotides opposite each other (base pair) form hydrogen bonds. Adjacent nucleotides that are in base pairs interact with each other via stacking. Both effects stabilize the helix. The four bases illustrated in detail on the left are part of a larger helix (middle). We can abstract the helix in a simplified secondary structure notation (right). (b) left got adapted from Wikimedia Commons: 'Molecular structure of DNA' by Madeleine Price Ball under CC0 1.0 License.

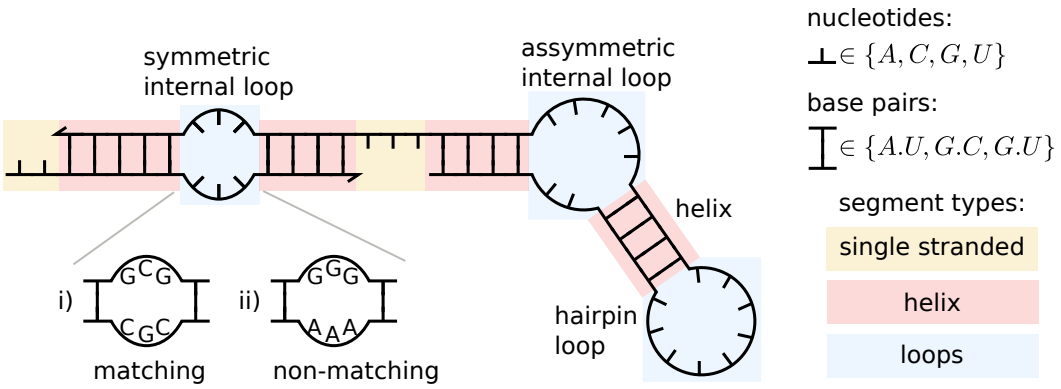


Figure 2.2.: example of a secondary structure consisting of basic secondary motifs: two single-stranded segments, three helices (red) and three loops (blue). The loops are a symmetric and asymmetric internal loop and a harpin loop. Note that the nucleotides in the loop are in an open configuration, which does not imply that they are non-matching, as can be seen in *cf.* i) and ii).

aromatic rings of the bases of adjacent nucleotides. Base-stacking makes the most significant contribution to helix stability and even dominantly determines its salt, and temperature dependence [35]. However, both contributions interact with each other and are not clearly separable [33]. The stability is highly sequence-dependent and is strongest for G, C rich sequences.

We will in the following use the abstraction for polymers as depicted on the right of Figure 2.1 (b), where each $\perp \in \mathcal{A}$ corresponds to a nucleotide.

2.2. Possible conformations of informational polymers

The stiffness of informational polymers is length-dependent: Strands shorter than the so-called persistent length behave like a rigid rod, whereas strands longer than the persistent length are modeled using a worm-like chain model. The persistent length depends on the sequence, the temperature, the salt concentrations etc. Naturally, the persistent length of a helix is longer than that of a single strand. The persistent length can be determined by measuring the diffusion coefficients via *e.g.*, FCS (fluorescence correlation spectroscopy) and comparing the measured values to results obtained by simulation [84].

We now discuss the possible conformations such a flexible information polymer or multiple information polymers can, in principle, adopt. Without loss of generality, we conduct this discussion at the example of the four nucleotide alphabet $\mathcal{A} = \{A, G, C, U\}$. We restrict ourselves to secondary structures; hence the strand is assumed to be confined in the two-dimensional plane and is transformed via bending. Thereby strand segments can align each other such that, if the sequences are complementary, contiguous base pairs are formed, resulting in the formation of helices. An example of such a secondary structure of two strands is shown in Figure 2.2. It can be split into basic secondary motifs: two single-stranded segments (yellow), four helices (red), and three-loop (blue).

Two facing nucleotides can be in an open or closed configuration, where only complementary nucleotides can adopt a closed configuration forming a base pair. Hence a loop does not necessarily consist of non-matching nucleotides, as depicted in i) and ii) of Figure 2.2. However, the probability that a loop forms at a position where the sequences do match is suppressed due to a penalty in free energy related to that structure, as discussed in the next section.

2.3. Energy models of secondary structures

In this chapter, we briefly review the concept of calculating free energies of secondary structures by so-called nearest-neighbor energy models. The RNA parameters for these calculations can be found in the Turner NNDB, see Ref. [52], and the parameters for DNA are published in Ref. [32].

These rules assign each basic secondary motif m that is not a helix a free energy penalty of formation $\Delta G^\circ_{\text{form,m}}$ that is particular to its type and shape, see Fig. 2.3 (a).

In the literature, there is also an initiation penalty associated with helix formation $\Delta G^\circ_{\text{ini}}$. I believe this naming is misleading. A better name would be bimolecular acceptance penalty. In contrast to the other basic secondary motifs, this penalty is not given for each helix (red in Fig. 2.3 (a)) but rather for the number of bimolecular reactions that were necessary to assemble the complex. A complex consisting of n strands, naturally, had to undergo $n-1$ bimolecular hybridization reactions.

Additionally each *block of four* that contains at least one base pair is assigned a sequence dependent free energy $\Delta G^\circ_B(N_1, N_2, N_3, N_4, (\text{bp}_1, \text{bp}_2))$, $N_i \in \mathcal{A}$, $\text{bp}_i \in \{0, 1\}$, as illustrated in Fig. 2.3 (b).

The libraries for DNA and RNA include tabulated measured values for small loops in dependence of their sequence, such as, e.g., 2x2 internal loops. Note that in this case, the free energy associated with a block of four containing one base pair and one mismatch is already included in the tabulated values.

There are also a couple of other rules and special cases such as a penalty for an A.U base-pair terminating a helix $\Delta G^\circ_{\text{AU}}$ or a symmetry penalty $\Delta G^\circ_{\text{sym}}$ for self-complementary strands. The symmetry penalty is equal to $\Delta G^\circ_{\text{sym}} = 0.43$ kcal/mol which corresponds to $\frac{\Delta G^\circ_{\text{sym}}}{k_\text{B}T} = \ln(2)$. This symmetry penalty will appear naturally in Section 4.7 when relating the elementary rates to the rate constants.

In general, the free energy of a complex C consisting of n strands is given by

$$\begin{aligned}\Delta G^\circ_{\text{tot}}(C) = &\sum_{m \in \text{motifs not helix}} \Delta G^\circ_{\text{form,m}} + \sum_{B \in \text{blocks}} \Delta G^\circ_B(N_1, N_2, N_3, N_4, (\text{bp}_1, \text{bp}_2)) \\ &+ k\Delta G^\circ_{\text{AU}} + \sigma \Delta G^\circ_{\text{sym}} + (n-1)\Delta G^\circ_{\text{ini}}\end{aligned} \tag{2.1}$$

where σ is 1 if the complex is rotationally symmetric under a 180° rotation in the plane and zero otherwise, and k is the number if AU(GU) ends of motifs. We call

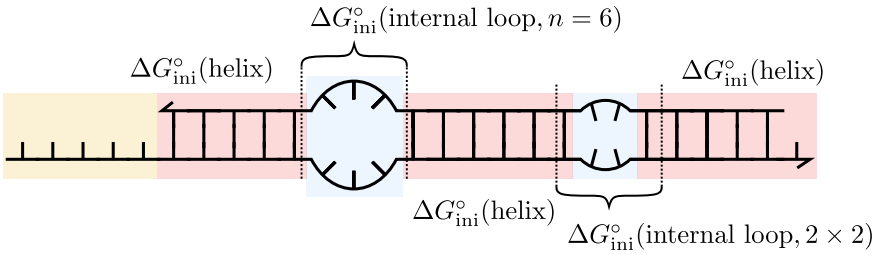
$$\Delta G^\circ_{\text{b,tot}} = \sum_{m \in \text{motifs not helix}} \Delta G^\circ_{\text{form,m}} + \sum_{B \in \text{blocks}} \Delta G^\circ_B(N_1, N_2, N_3, N_4, (\text{bp}_1, \text{bp}_2)) \tag{2.2}$$

the total binding energy. Hence

$$\Delta G^\circ_{\text{tot}}(C) = \Delta G^\circ_{\text{b,tot}} + k\Delta G^\circ_{\text{AU}} + \sigma \Delta G^\circ_{\text{sym}} + (n-1)\Delta G^\circ_{\text{ini}} \tag{2.3}$$

Simplifications can be undertaken, such as, e.g, neglecting the AU end penalty or assigning dangling ends (block of four with three nucleotides) an energy of zero.

(a) Free energy due to initiation penalty of basic secondary motifs:



(b) Free energy due to blocks of four:

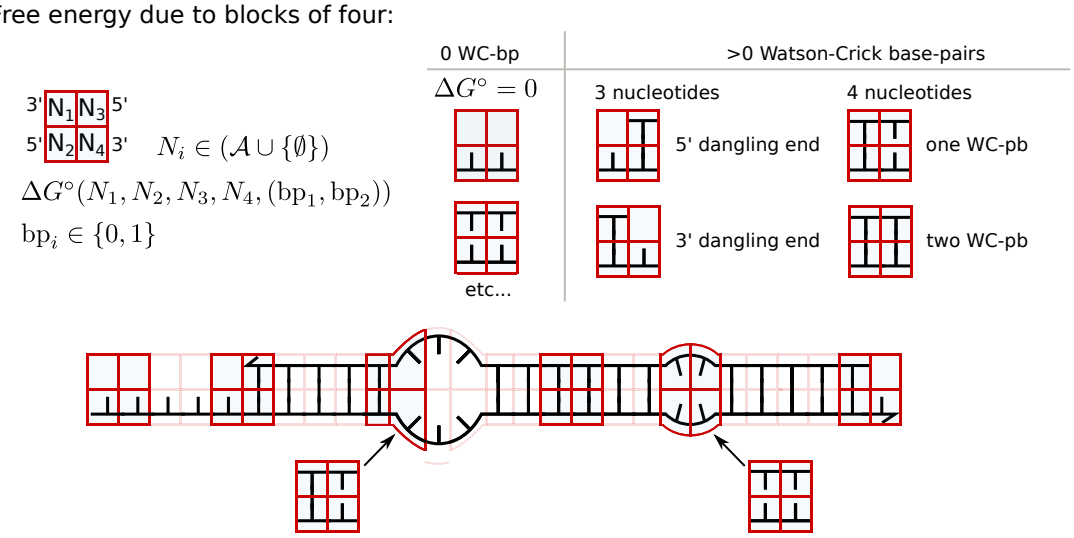


Figure 2.3.: principle outline of the nearest neighbor energy models: **(a)** We assign each basic secondary structure motif (e.g. internal loops, helix, hairpin loop etc.) a specific initiation penalty of formation $\Delta G^\circ_{\text{ini}}$. $\Delta G^\circ_{\text{ini}}$ depends on the type and particular geometry of the motif. **(b)** We basically assign the helices and their adjacent base pairs a free energy. This can be interpreted as assigning each *block of four* a free energy $\Delta G^\circ(N_1, N_2, N_3, N_4, \text{bp}_1, \text{bp}_2)$. This free energy depends on the nucleotides contained in the block of four and if N_1 N_2 and N_3 N_4 are in a closed configuration forming a Watson-Crick base pair (WC-bp). As mentioned in Section 2.2 even though, e.g., N_1 and N_2 of a block can in principle form a WC-bp, they can be assumed to be in an open configuration, not forming a WC-bp.

3. Class of models addressed by the simulation

In this chapter, we present the class of models that can be studied by our simulation framework. We first give an overview of the complex-structures that can be formed by hybridization of informational polymers that are included in our model. Next, we outline the reactions that these complexes can undergo. Subsequently, we illustrate the assembly processes that are introduced through the allowed complex structures and reactions. The system is assumed to be a well-mixed reaction vessel of volume V and is, therefore, zero-dimensional (see Figure 3.4 (a)). Hence, the rates of the elementary processes have no spatial dependence.

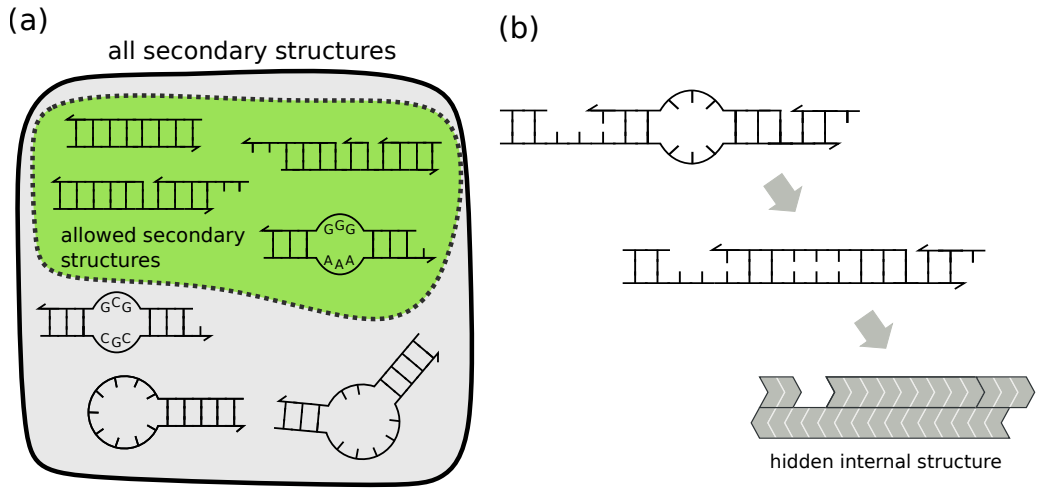


Figure 3.1.: (a) Types of secondary structures that can be handled by the algorithm: single strands, helices and symmetric internal loops. Non symmetric loops such as non symmetric internal loops or hairpins can not be represented by the data structure. We further only allow for symmetric internal loops of non-matching opposing nucleotides. Hence effectively, we consider all secondary structures that can be formed by alignment of the two strands where all possible base pairs are formed. **(b)** As we exclude internal loops of matching base pairs, non-matching nucleotides can be represented unambiguously by a gap. We can simplify the representation even further by hiding all internal structure.

3.1. Complexes and strands

The basic element of our dynamics is a directed oligomer called a strand-like described in Chapter 2.

Virtually all secondary structures are taken into account that can be formed by aligning rigid strands. Therefore the only secondary structures that can be taken into account are single strands, helices, and symmetric internal loop (see Figure 3.1 (a), green area).

We further assume that two matching nucleotides facing each other in a hybridization

site adopt a closed configuration and hence form a base pair. This restriction to one possible configuration upon hybridization excludes symmetric internal loops consisting of matching base pairs. Non-matching pairs of nucleotides can now unambiguously be represent by a helix with a gap (Figure 3.1 (b) middle). We facilitate the representation even further by hiding the internal configuration of a hybridization site using the representation of rigid blocks (Figure 3.1 (b) bottom).

3.2. Segments

We further virtually divide strands into so-called segments, see Figure 3.2. A segment starts or ends with a strand or when the strand changes from a double-stranded configuration into a single-stranded configuration or vice versa. Hence a strand is segmented into single and double-stranded segments.

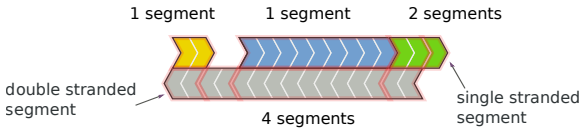


Figure 3.2.: The strands of the complex shown in Fig. 3.1 (b) are divided into so-called segments.

3.3. Elementary processes

In this section, we present the elementary processes (or reactions) informational polymers can undergo in our simulation framework. The internal elementary processes compatible with the data structure of our simulation are: hybridization, dehybridization, templated ligation, and cleavage (see Figure 3.3 (a)). In addition, influx and outflux couple the system to its environment (see Figure 3.3 (b)).

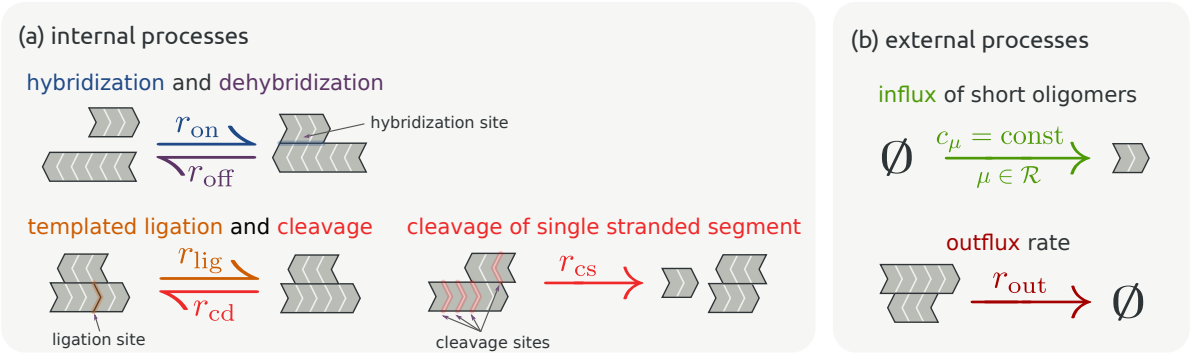


Figure 3.3.: (a) Overview of all internal reactions. Hybridization and dehybridization rates are chosen thermodynamically consistent. Cleavage reactions can have different rates depending on if a single strand (r_{cs}) or a strand in a double strand configuration (r_{cd}) is cleaved. Covalent bonds between a double strand and a single strand are supposed to be cleaved at the single strand rate r_{cs}. **(b)** Overview of external reactions that couple the system to two external reservoirs.

3.3.1. Influx and outflux

The influx of single strands of lengths $m \in \mathcal{R}$ is implemented by keeping their concentration c_m constant, which can be interpreted as a coupling of the system to a large external reservoir. Each complex is removed from the system with an outflux rate denoted as r_{out}.

The chemostat concentrations c_m will be used to model the influx of small building blocks such as monomers, dimers, and trimers. This can be thought to represent the situation discussed in Chapter 1.1, where a rich medium surrounded a protocell, and building blocks entered the cell via diffusion. Further, choosing *e.g.*, a constant outflux rate $r_{\text{out}} = \text{const}$ could account *e.g.*, for a serial dilution upon division of the protocell. At the same time, an outflux can facilitate studying the properties of a system as a stationary state can be reached even if cleavage is excluded from the model.

3.3.2. Hybridization and dehybridization

Hybridization and dehybridization are assumed to be elementary reactions occurring with rates r_{on} and r_{off}, defined for a single hybridization site. We assume that hybridization and dehybridization are one step reactions: Upon hybridization of a particular alignment of two strands, all possible base pairs are formed, and upon dehybridization all base pairs are dis-ruptured. Thermodynamic consistency [54, 80] connects the elementary rates r_{on} and r_{off} for hybridization and dehybridization to the standard binding free energy ΔG_b° of a hybridization site:

$$\frac{r_{\text{off}}}{r_{\text{on}}} = (V N_A c^\circ) e^{\beta \Delta G_b^\circ}, \tag{3.1}$$

where $\beta = (k_{\text{B}} T)^{-1}$, k_{B} is Boltzmann's constant and T denotes the (absolute) temperature, N_A is the Avogadro constant and $c^\circ = 1\,\text{mol/l}$ is the standard concentration.

3.3.3. Templated ligation and cleavage

When two strands of length L_1 and L_2 are hybridized adjacently on a third strand L_3, they form a ligation site, see Fig. 3.3 (a). In such a configuration the strands L_1 and L_2 can ligate at rate r_{lig}, forming a strand of length $L_1 + L_2$. This process is called templated-ligation, with the third strand understood as the template.

Cleavage, on the contrary, breaks strands apart. Cleavage of a bond of a single-stranded segment, with elementary rate r_{cs}, forms two separate complexes. In principle, cleavage of bonds in double-stranded segments could also be implemented in a straightforward manner. It would also be possible to assign the bonds located at the transition from a double-stranded segment to single-stranded segment or bonds located at a templated ligation site a particular cleavage rate.

3.3.4. Random ligation

Note that we did not include random ligation of two complexes at this point. However, the short building blocks, such as, e.g., dimers, must form through non templated ligation of monomers in the first place. Therefore, it is desirable to include random ligation into the model. Including this reaction into the code of the simulation framwork is manageable. However, the physical parametrization in order to still provide a thermodynamic consistent hybridization energy model is highly challenging.

3.3.5. Example dynamics

An example of successive reactions is illustrated in Figure 3.4 (b), thereby tracking a specific dimer (blue). The model dynamics were implemented in *C++* based on the Gillespie algorithm [7, 5, 24]. An important feature of our stochastic simulation is that the state space

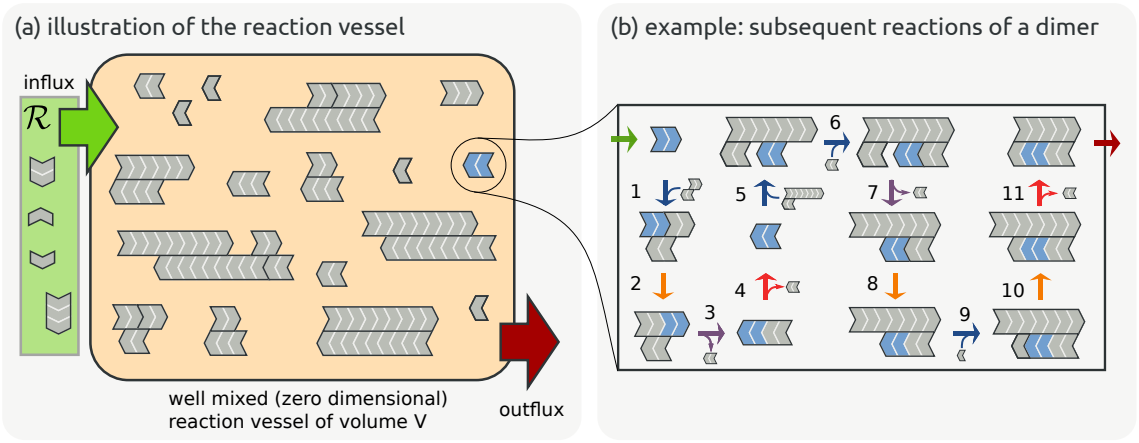


Figure 3.4.: (a) Illustration of the system (reaction vessel). **(b)** Illustration of successive reactions by tracking the fade of a specific dimer (blue): (1&2) After the dimer enters the system, it becomes ligated to a second dimer via templated ligation. (3) After the third dimer, which served as a template, dehybridizes the newly formed single-stranded tetramer undergoes cleavage (4). Thereafter the dimer hybridizes onto a duplex (5). As there is a gap between the dimer and its neighboring strand on the template, no ligation site is created. (6) Subsequently, another dimer hybridizes onto the triplex, whereby a ligation site is formed. The duplex, apart from the ligation site dehybridizes (7), and the triplex becomes a duplex via templates ligation (8). Next, the duplex is extended by a monomer (hybridization and subsequent ligation) (9&10). Cleavage of the remaining overhang leads to a fully-hybridized duplex (11) (duplex without overhang), which thereafter leaves the system.

of complex configurations is dynamically generated rather than completely generated in advance.

4. Theoretical foundation of the simulation framework*

4.1. Markov Chain on the state space of copy numbers

Let us consider a system of N_n^{tot} complexes, where n defines the exact state of the system. We define the countable set of all possible complex species by $\mathcal{S}$. Each individual complex i belongs to a species $S_{j_i} \in \mathcal{S}$. We further define the state n as

$$n = (N_1, N_2, N_3, ...), \tag{4.1}$$

where N_i is the copy number (occupation number) of a species $S_i \in \mathcal{S}$. The total number of complexes is thus $N_n^{\text{tot}} = \sum_i N_i$. We also define the set of species present in state n as $\mathcal{S}_n = \{S_i : S_i \in \mathcal{S} \text{ and } N_i > 0\}$. The total number of species in state n is thus $S_n^{\text{tot}} = |\mathcal{S}_n|$. All possible states n span the state space of copy numbers $\mathcal{N}$.

The complexes can undergo reactions $\mu \in \mathcal{M}_n$, where $\mathcal{M}_n$ is defined as the set of all possible reactions of a system in state n. Thus the number of possible reactions in state n is $M_n = |\mathcal{M}_n|$. Within our model μ is either a hybridization, dehybridization, ligation or outflux of a complex, for which we use the labels *on*, *off*, *lig* and *out* respectively. The index μ can be mapped onto a tuple

a) (on, i, j, c) for hybridization,

b) (off, i, c) for dehybridization,

c) (lig, i, c) for ligation,

d) and (out, i) for outflux,

where i, j label the species of the reactants, and c specifies the hybridization or ligation site (channel) between two complexes or within a complex. We assign each reaction μ a so-called elementary rate r_μ (or reaction parameter, see [5]). It is the rate at which individual complexes (monomolecular reactions) or tuples of complexes (bimolecular reactions) undergo a specific reaction. The elementary rates have units of time t_0. Using the mapping of μ onto its corresponding tuples, the elementary rates r_μ can be written as $r_{\text{on}}(i, j, c)$, $r_{\text{off}}(i, c)$, $r_{\text{lig}}(i, c)$ and $r_{\text{out}}(i)$.

The above reactions connect the state space $\mathcal{N}$ and define the transition rates of a Markov chain: If $n' \in \mathcal{N}$ can not be reached from $n \in \mathcal{N}$ via the above reactions, the transition rate is zero, $w(n \to n') = 0$. If $n \in \mathcal{N}$ can transition to state $n' \in \mathcal{N}$ via one of the above reactions, the transition rate is equal to $w(n \to n') = h_\mu r_\mu$, where h_μ is the number of possible reactant combinations for the reaction μ to occur. In our model each transition rate is related to a specific reaction μ, we can thus simplify the notation of the transition rates to $w(n \to n') = r_\mu^{\text{tot}}$, which we call the total rate of reaction μ:

$$r_\mu^{\text{tot}} = h_\mu r_\mu. \tag{4.2}$$

We call h_μ the combinatorial factor of the reaction μ. We will derive the combinatorial factors for the canonical choice of chemical species in Sec. 4.7, which we will use to connect the

*This chapter got adapted from the supplement of [106]

elementary rates r_μ to the chemical rate constants k_μ. In the next section, we first present the principles of the Gillespie algorithm, which we use to compute the time evolution of the Markov chain.

4.2. Gillespie algorithm

We use the Gillespie algorithm [7] to compute the time evolution on the Markov chain. Assuming that at time t the chain is in state $n \in \mathcal{N}$, each possible transition rate r_μ^{tot} is calculated. The probability $p(\mu)$ that the transition μ is chosen to be the next transition to occur is given by the ratio of r_μ to the total transition rate $r^{\text{tot}} = \sum_{\mu=1}^{M_n} r_\mu^{\text{tot}}$,

$$p(\mu) = \frac{r_\mu}{r^{\text{tot}}}. \tag{4.3}$$

The waiting time τ for the next reaction to happen is exponentially distributed

$$\tau \sim r^{\text{tot}} e^{-r^{\text{tot}}\tau}. \tag{4.4}$$

Exponentially distributed waiting times are generated from uniformly distributed random variables u ($p(u) = \mathcal{U}(0,1)$ where $\mathcal{U}(0,1)$ is the uniform distribution on the unit interval) and the aid of the inverse transform sampling theorem:

$$p(\tau) = -\frac{\ln(u)}{r^{\text{tot}}}. \tag{4.5}$$

When a reaction is chosen, the system time is updated according to $t \to t + \tau$. The Markov chain transitions from state n to state n'.

4.3. Implementation of the Gillespie algorithm

The algorithm's performance can be improved by drawing the next reaction μ using a binary tree [24]. Thereby, the scaling of the simulation time reduces from $\mathcal{O}(S_n^{\text{tot}})$ to $\mathcal{O}(\ln(S_n^{\text{tot}}))$.

The state $n \in \mathcal{N}$ is represented in a data structure on the computer. Performing the reaction implies modifying the data structure representing n according to the rules of the reaction such that it represents n'. Also the set of possible reactions needs to be updated to match the new state:

$$\left\{r_\mu\right\}_{\mu \in I_n} \to \left\{r_\mu\right\}_{\mu \in I_{n'}} \tag{4.6}$$

Instead of an update, one could recalculate all possible reactions of state n'. However, this would be disadvantageous performance-wise. Our implementation of the algorithm tries to keep the number of reactions and the data structures that need to be updated after each reaction as small as possible while still keeping the bookkeeping manageable. One way to achieve this is not to sample bimolecular reactions (hybridizations) directly but instead decompose them into collision events, followed by the selection of one possible hybridization (channel) upon collision. The reaction channels need to be calculated only upon collision. Furthermore, choosing the collision rate to be the same for all pairs of potentially colliding complexes allows the compression of all hybridizations into a single collision event. We will explain this procedure in detail in Sec. 4.6.

4.4. Decomposition of the hybridization rate

As introduced in the previous sections, each hybridization represents a transition from a state n to a new state n', as depicted in Fig. 4.1(a). The system contains three species. Several hybridizations are possible (thin blue lines). The magenta line highlights the chosen hybridization.

In our implementation, we separate the transport process that brings the molecules into proximity (collision) from the specific interactions upon encounter. To this end, we decompose r_{on} into the rate at which the reactants of species $S_i, S_j \in \mathcal{S}$ collide (r_{coll}), times the probability that the collision leads to a hybridization ($:= p_{\text{a(cceptance)}}$), times the probability for the specific reaction channel ($:= p_{\text{c(hannel)}}$):

$$r_{\text{on}}(i,j,c) = r_{\text{coll}}(i,j)p_a(i,j)p_c(i,j,c). \tag{4.7}$$

After a collision, we directly evaluate whether an interaction occurs (probability $p_a(i,j)$). If the reaction is reactive, we choose a hybridization channel. The hybridization is then performed immediately. This can be interpreted as the introduction of a transient state with a lifetime of effectively zero, see Fig. 4.1(b).

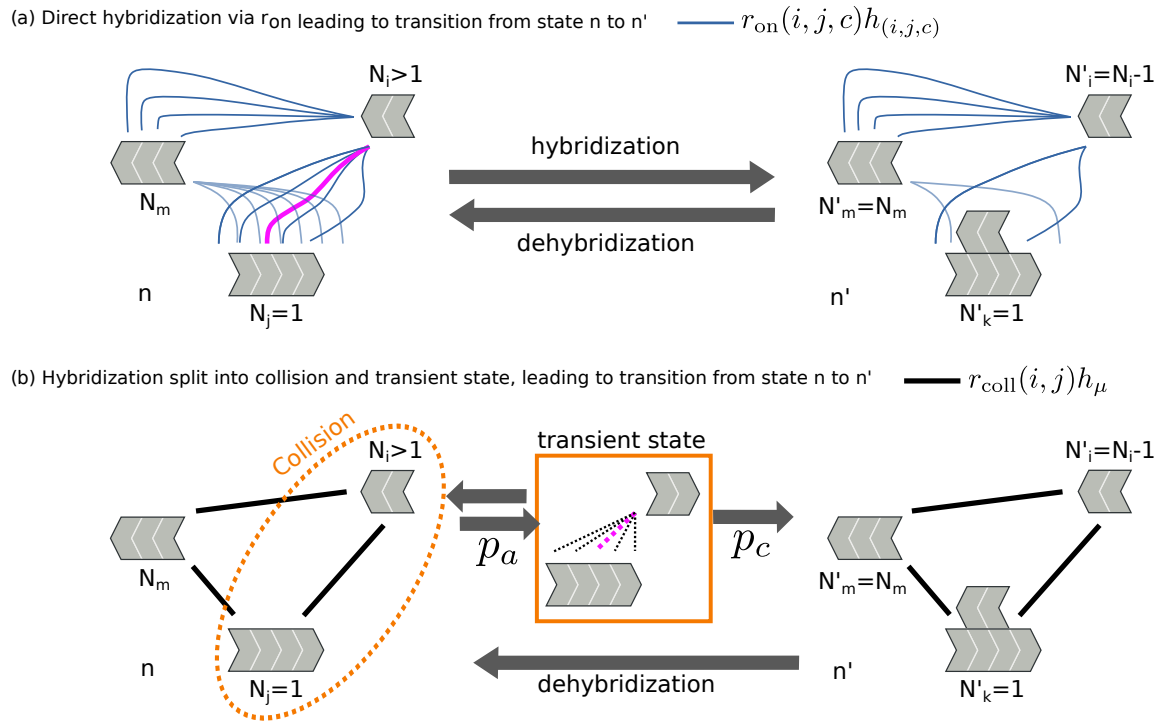


Figure 4.1.: Hybridization on the Markov chain on the space of copy numbers $\mathcal{N}$ where (a) hybridizations are individual reactions and (b) a hybridization is split into a collision and a subsequent transient state (b). The system contains three complex species S_m, S_i, S_j with copy numbers $N_m, N_i > 1, N_j = 1$. The hybridization performed is between a complex of species S_i and S_j (magenta line in (a)), leading to the new species S_k. (a) Each blue line corresponds to a possible hybridization reaction with elementary rate $r_{\text{on}}(i,j,c)$. (b) Hybridization is interpreted as a two step reaction. First, the two complexes highlighted by the dotted orange ellipse collide. The complexes form a transient state with probability p_a, and the reaction gets rejected with probability $1 - p_a$. Upon formation of the transient state, a channels c is selected with probability p_c.

4.5. Extracting the volume dependence from the collision rate

Assigning each reaction an elementary rate that only depends on the reacting chemical species neglects any spatial heterogeneity and assumes a spatially homogeneous distribution of the reactants [5]. Furthermore, we assume that the local equilibrium distribution of the velocities is restored sufficiently fast after a reactive reaction. Both criteria require that the nonreactive (elastic) encounters are much more frequent than the reactive (inelastic) encounters. A system fulfilling these criteria is also called well-mixed. This could lead to the conclusion that the acceptance probability p_a must be small in our model. However, the criteria can also be fulfilled by introducing nonreactive species, which are not explicitly modeled. For RNA/DNA systems, those nonreactive species simply correspond to the solvent molecules.

If the criteria are fulfilled, the elementary collision rate is equal to the ratio of the (ensemble) averaged collision volume per unit time $\langle dV_{\text{coll}}(i,j)/dt \rangle$ to the system volume V (see [5]):

$$r_{\text{coll}}(i,j) = \frac{1}{V} \left\langle \frac{dV_{\text{coll}}(i,j)}{dt} \right\rangle \tag{4.8}$$

$\langle dV_{\text{coll}}(i,j)/dt \rangle$ will depend on the details of the transport process, the temperature, the solvent etc. However, the $1/V$ scaling is a generic feature due to the spatial homogeneity assumption. This equation also shows that the rate constant for a bimolecular reaction, $k_\mu = V N_A r_\mu$ is actually (as it should be) independent of the volume.

4.5.1. Example: Decomposition of the bimolecular reaction rate of hard spheres

Instead of looking at the more complex hybridization reactions, let us consider a reaction between two gases S_1, S_2 of hard spheres as a simple example for the decomposition of a bimolecular reaction rate as done in [5]. The bimolecular reaction rate is colloquially called the on-rate r_{on}. The radii of the reactants are R_1, R_2, and their masses are given by m_1, m_2. Whenever the distance of two spheres becomes equal to the sum of their radii $d_{12} = R_1 + R_2$, a collision takes place. Let us assume that for a reaction to occur, the reactants further have to overcome an energetic barrier E. The on-rate for this reaction is consequently given by:

$$r_{\text{on}} = \frac{1}{V} \left\langle \frac{dV_{\text{coll}}(i,j)}{dt} \right\rangle e^{-\frac{E}{kT}} = \frac{1}{V} \pi d_{12}^2 \langle v_{12} \rangle e^{-\frac{E}{kT}} = \frac{1}{V} \pi d_{12}^2 \left(\frac{8kT}{\pi m_{12}} \right)^{\frac{1}{2}} e^{-\frac{E}{kT}}, \tag{4.9}$$

where $\langle v_{12} \rangle$ is the average velocity of an arbitrary S_1 molecule relative to an arbitrary S_2 molecule and $m_{12} = m_1 m_2/(m_1 + m_2)$ is the reduced mass. We can identify $r_{\text{coll}} = \frac{1}{V} \pi d_{12}^2 \left(\frac{8kT}{\pi m_{12}} \right)^{\frac{1}{2}}$, $p_a = e^{-\frac{E}{kT}}$, and $p_c = 1$ as there is only one possible reaction channel.

4.5.2. Effective collisions rate r_0

For convenience we define

$$r_0(i,j) := N_A c^\circ \langle dV_{\text{coll}}(i,j)/dt \rangle, \tag{4.10}$$

where N_A is the Avogadro constant and $c^\circ = \text{mol/l}$ is the standard concentration. We call r_0 the effective collision rate. Hence, the elementary collision rate can be written as (*cf.* Eq. (4.8))

$$r_{\text{coll}}(i,j) = \frac{1}{V N_A c^\circ} r_0(i,j). \tag{4.11}$$

The last equation will be useful when deriving the free energy of a hybridization site in the next section.

4.6. Constant collision rate allows for a single total collision rate

A constant collision rate $r_{\text{coll}} = \text{const.}$, independent of the properties of the colliding complexes, allows us to condense all collisions into a single total collision occurring with rate

$$r_{\text{coll}}^{\text{tot}} = \frac{N_n^{\text{tot}}(N_n^{\text{tot}} - 1)}{2} r_{\text{coll}}, \tag{4.12}$$

where N_n^{tot} is the total number of complexes in state $N \in \mathcal{N}$. When the Gillespie algorithm has selected the total-collision as the next reaction to be performed, two individual complexes that undergo the collision must be drawn. In order to obtain the correct combinatorial factors for bimolecular reactions, we simply draw each pair of individual complexes with the same probability. The total collision rate within a species S_i is thus given by

$$r_{\text{coll}}^{\text{tot}}(i,i) = \frac{N_i(N_i - 1)}{2} r_{\text{coll}}, \tag{4.13}$$

whereas the total collision rate between two different species S_i, S_j is

$$r_{\text{coll}}^{\text{tot}}(i,j) = N_i N_j r_{\text{coll}}. \tag{4.14}$$

The two combinatorial factors h_μ are illustrated in *cf.* Figure 4.2. We can write Eq. (4.12) as a sum of collisions within a species and between species.

$$r_{\text{coll}}^{\text{tot}} = \frac{N_n^{\text{tot}}(N_n^{\text{tot}} - 1)}{2} r_{\text{coll}} = \sum_{i=1}^{S_n^{\text{tot}}} \frac{N_i(N_i - 1)}{2} r_{\text{coll}} + \sum_{\substack{i,j \\ i<j}} N_i N_j r_{\text{coll}}. \tag{4.15}$$

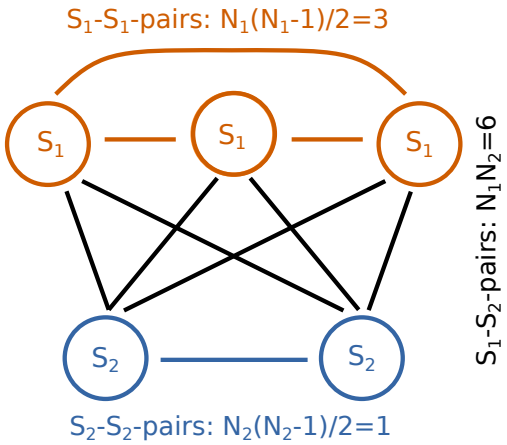


Figure 4.2.: Combinatorial factors for simple chemical species: The number of possible reactant pairs within a species S_i is given by $h_{(on,i,i)} = N_i(N_i - 1)/2$. For reactions between two species $S_i \neq S_j$ the combinatorial factor is equal to $h_{(on,i,j)} = N_i N_j$. Orange lines: Reacting pairs within species S_1. Blue lines: Reacting pairs within species S_2. Black lines: Reacting pairs between species S_1 and S_2.

4.6.1. Constant effective collision rate

As the collision rate is constant, also the effective collision rate r_0 is constant (*cf.* Equation (4.11)). We set $r_0 = 1/t_0$, where t_0 is the arbitrary unit of time. Thus, the elementary collision rate is given by

$$r_{\text{coll}} = \frac{1}{V N_A c^\circ} r_0 = \frac{1}{V N_A c^\circ} \frac{1}{t_0}. \tag{4.16}$$

4.7. Combinatorial factor h_μ and the relation between rate constants and elementary rates

In chemical equilibrium, the ratio of the rate constants for hybridization and dehybridization is given by (determines) the change in Gibbs free energy associated with the reaction:

$$\ln\left(\frac{k_{\text{off}}(i,j,c)}{c^\circ k_{\text{on}}(i,j,c)}\right) = \Delta G^\circ_{\text{hyb}}(i,j,c). \tag{4.17}$$

Further, in chemical equilibrium, the concentration of species that can be formed by hybridization, are determined by the initial concentrations and $\Delta G^\circ_{\text{hyb}}(i,j,c)$. We aim to choose the elementary rates of hybridization and dehybridization such that our system, only undergoing these two reactions, would reach chemical equilibrium. We, therefore, have to connect the chemical rate constants k_μ to the elementary rates r_μ. For this purpose, we have to derive the combinatorial factors h_μ, where we closely follow the work of Gillespie [5]. For completeness, we derive the combinatorial factors not only for hybridization and dehybridization but for all reactions present in our model as they would be needed, e.g., to compare results obtained via our stochastic simulation to results obtained via solving a set of chemical rate equations. For simple chemical reactions the combinatorial factors are illustrated in Fig. 4.2:

- For bimolecular reactions of two different reactants, $S_i \neq S_j$, the combinatorial factor is $h_\mu = N_i N_j$, and for two reactants of the same species S_i it is $h_\mu = N_i(N_i - 1)/2$.
- For monomolecular reactions of a species S_i it is $h_\mu = N_i$.

The elementary rates are connected to their corresponding (chemical) rate constants k_μ (for molar concentrations) via:

- $k_\mu = VN_A r_\mu$ for bimolecular reactions with $S_i \neq S_j$, and $k_\mu = VN_A r_\mu/2$ in case that $S_i = S_j$, where N_A is the Avogadro constant.
- $k_\mu = r_\mu$ for monomolecular reactions.

In our case, these relations must be modified to account for multiple channels leading to the same product due to the internal structure of the complexes. We will therefore derive modified relations for h_μ and k_μ in this section.

4.7.1. Rotationally symmetric duplexes

To this end, we need the notion of a rotationally symmetric complex, see Fig. 4.3, where we used the previously defined block notation but with an example sequence. In the following sections, we will omit the notation of the sequences.

The probability of the formation of a rotationally symmetric complex will generally decrease with the size of the alphabet and the length of the complex. Also, generally (for a non self-complementary choice of nucleotides, hence $|\mathcal{A}| > 1$), symmetric duplexes will have a large positive binding energy (unstable) as they must include many non-matching base pairs. The only exception being symmetric duplexes, where the strands consist of alternating sequences.

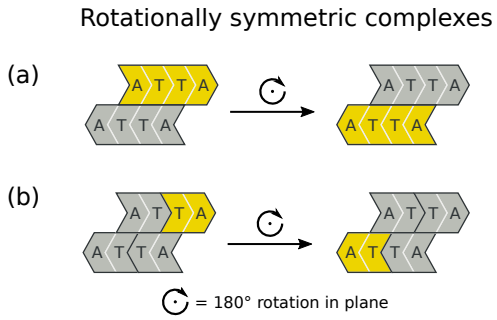


Figure 4.3.: Two example complexes that are rotationally symmetric under a 180° rotation in the plane: **(a)** symmetric duplex, **(b)** symmetric complex of four strands.

4.7.2. Hybridization

Let us briefly revise the relationship between the elementary rate and the corresponding rate constant for simple bimolecular reactions where the number of channels is $c = 1$. We first consider a bimolecular reaction where two reactants of species $S_i = S_j$ react to species S_k, i.e. $\mu = (\text{on}, i, i)$,

$$R_\mu : \; 2S_i \to S_k. \tag{4.18}$$

As illustrated in Fig. 4.2, the combinatorial factor is $h_\mu = N_i(N_i - 1)/2$. Hence the total reaction rate as defined in Eq. (4.2) reads:

$$r_\mu^{\text{tot}} = \frac{N_i(N_i - 1)}{2} r_\mu. \tag{4.19}$$

For large N_i this becomes approximately $r_\mu^{\text{tot}} \approx N_i^2 r_\mu/2$. Let us now consider an ensemble of systems, all in the current state n. The ensemble average $\delta t \langle N_i^2 \rangle r_\mu/2$ will then correspond to the average number of reactions in the system that will occur in the next time step δt. Neglecting fluctuations we can write $\langle N_i^2 \rangle = \langle (\langle N_i \rangle + \delta N_i)^2 \rangle = \langle N_i \rangle^2 + \langle \delta N_i^2 \rangle \approx \langle N_i \rangle^2$, which we can use to formulate a (chemical) rate equation

$$\langle \dot{N}_k \rangle = \frac{r_\mu}{2} \langle N_i^2 \rangle \approx \frac{r_\mu}{2} \langle N_i \rangle^2 . \tag{4.20}$$

Rewriting this equation in molar concentrations $c_i = \langle N_i \rangle / (N_A V)$, with $N_A = 6.022 \times 10^{23} \text{mol}^{-1}$ being the Avogadro constant, yields

$$\dot{c_k} = \frac{r_\mu}{2} V N_A c_i^2, \tag{4.21}$$

from which we can identify the rate constant k_μ to be related to the on-rate r_μ by

$$k_{\text{on}}(i,i) = \frac{r_{\text{on}}(i,i)}{2} V N_A. \tag{4.22}$$

Let us now consider the simple bimolecular reaction $\mu = (\text{on}, i, j)$ of two different reactants $S_i \neq S_j$

$$R_\mu : \; S_i + S_j \to S_k. \tag{4.23}$$

The combinatorial factor is given by $h_\mu = N_i N_j$. The procedure of deriving the relationship between the rate constant and the elementary rate is analogous to the above derivation, except that instead of fluctuations, correlations between the copy numbers are neglected, $\langle N_i N_j \rangle \approx \langle N_i \rangle \langle N_j \rangle$, yielding

$$k_{\text{on}}(i,j) = r_{\text{on}}(i,j) V N_A. \tag{4.24}$$

These relations must be adapted in our system because of configurations where multiple hybridization channels lead to the same product complex. This is the case if one or both of the reactants are (rotationally) symmetric, see Figure 4.3. The combinatorial factors h_μ for bimolecular reactions are replaced by $h_\mu \to 2^m h_\mu$, *cf.* Fig. 4.4. The relation between the elementary rate and rate constant for a hybridization $\mu = (\mathrm{on}, i, j, c)$ is thus generally given by

$$k_{\mathrm{on}}(i,j,c) = 2^{m-\delta_{ij}} r_{\mathrm{on}}(i,j,c) V N_A. \tag{4.25}$$

An overview of the values of the prefactor $2^{m-\delta_{ij}}$ can be found in Table 4.1.

	$m=0$	$m=1$	$m=2$
$S_i = S_j$	1/2	1	2
$S_i \neq S_j$	1	2	4

Table 4.1.: Overview of prefactor for the relation between the on-rate and rate constant, $2^{m-\delta_{ij}} r_{\mathrm{on}}$. m is the number of species that undergo the hybridize which are rotationally symmetric under a 180° rotation.

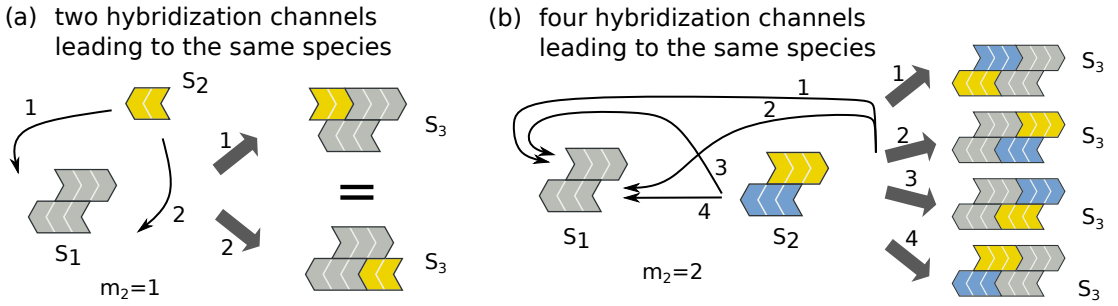


Figure 4.4.: (a) The complex of species S_1 is rotationally symmetric, whereas the the complex of species S_2 is not, hence $m = 1$. The two hybridization sites 1 and 2 lead to the same complex (chemically) of species S_3. **(b)** The complexes of species S_1 and S_2 are rotationally symmetric, hence $m = 2$. The four hybridization sites 1,2,3, and 4 lead to the same complex (chemically) of species S_3.

4.7.3. Dehybridization

For a dehybridization $\mu = (\mathrm{off}, i, c)$ of a complex of species S_i, into complexes of species S_j and S_k,

$$R_\mu : \; S_i \to S_j + S_k, \tag{4.26}$$

the combinatorial factor is given by $h_\mu = N_i$. However, a special situation occurs for rotationally symmetric complexes consisting of more than four strands $n \geq 4$ (*cf.* Fig. 4.5). In this case, there are always two dehybridizations leading to the same resulting complex species except for the hybridization site in the (geometric) center, which leads to two symmetric complexes of the same species $j = k$. Hence, for the center dehybridization, the combinatorial factor is simply given by $h_\mu = N_i$ whereas for the other dehybridizations the combinatorial factor is $h_\mu = 2N_i$. Generally, the relation between the elementary rate r_μ and the rate constant k_μ is thus given by

$$k_{\mathrm{off}}(i,c) = 2^{m_1-\delta_{jk}} r_{\mathrm{off}}(i,c), \tag{4.27}$$

where $m_1 = 1$ if the initial complex is rotationally symmetric and $m_1 = 0$ otherwise and $2^{-\delta_{jk}}$ accounts for the center dehybridization resulting in two complexes of the same species.

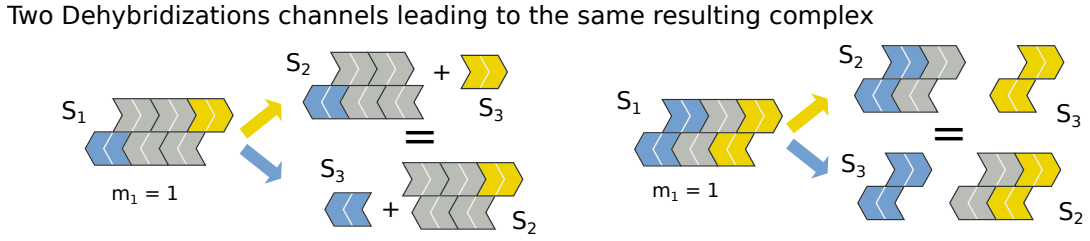


Figure 4.5.: Two elementary dehybridizations lead to the same complexes. This is the case for all dehybridizations of a rotationally symmetric complex consisting of more than four strands, $n \geq 4$, except the center dehybridization.

4.7.4. Collision

As the collision between two complexes is a regular bimolecular reaction the relation between the rate constant and the elementary rate is

$$k_{\text{coll}}(i,j) = 2^{-\delta_{ij}} r_{\text{coll}}(i,j). \tag{4.28}$$

4.7.5. Ligation

The derivation of the combinatorial factors for ligation reactions $\mu = (\text{lig}, i, c)$ is analog to the derivation of the combinatorial factors for dehybridization of the last section. The relation between the elementary rate and the rate constant is thus given by

$$k_{\text{lig}}(i,c) = 2^{m_1} r_{\text{lig}}(i,c), \tag{4.29}$$

where $m_1 = 1$ if the initial complex is rotationally symmetric and $m_1 = 0$ otherwise, see Fig. 4.6.

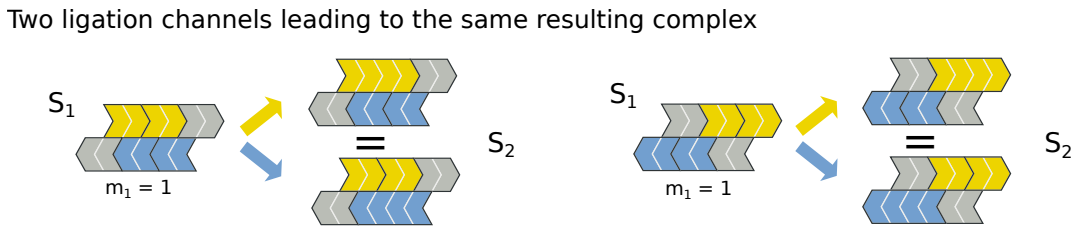


Figure 4.6.: Two ligations lead to the same complex. This is the case for ligations in rotationally symmetric complexes.

4.7.6. General rule for mapping between rate constants and rates

Like we have seen in the previous sections for hybridization, dehybridization, and ligation, the standard combinatorial factors need to be multiplied by a factor 2^m where m is the number of symmetric complexes undergoing the reaction. Additionally, we must divide by a factor $1/2$ if the reaction consumes or generates two complexes of the same species.

In order to obtain a combined expression for reactions that change the number of complexes, $\Delta N^{\text{tot}} = \pm 1$, (such as hybridization and dehybridization) and reactions that do not

change the number of complexes, $\Delta N^{\text{tot}} = 0$, (such as templated ligation), we assign the non existing second complex the species label S_0:

$$h_\mu \to 2^{m-\delta_{ij}} h_\mu, \tag{4.30}$$

where $i, j \in \mathbb{N}$ is the species of the two generated/consumed complexes and $j = 0$ if the reaction is a monomolecular reaction that does not lead to an increase in the number of complexes (*e.g.*, templated ligation, cleavage of strand in a double strand configuration).

4.7.7. Cleavage

According to Section 4.7.6 the relation between the rate constant and elementary rate for cleavage of a complex of species S_i into species S_j, S_k is:

$$k_{\text{c}}(i,c) = 2^{m-\delta_{jk}} r_{\text{cs}}(i,c). \tag{4.31}$$

We can distinguish between cleavage of strands in a double strand configuration and the cleavage of a single strands or strands at an interface to a double-strand which leads to two product strands:

$$k_{\text{cd}}(i,c) = 2^m r_{\text{cd}}(i,c) \tag{4.32}$$

$$k_{\text{cs}}(i,c) = 2^{m-\delta_{jk}} r_{\text{cs}}(i,c). \tag{4.33}$$

4.8. Gibbs Free Energies of Hybridization

In this section, we will first show how the correct choice of the off-rate leads to a Gibbs free energy of a hybridization site, which is physically meaningful. The thereby derived free energy of a hybridization site is in agreement with standard energy models for DNA or RNA [52, 32]. We then use these results to derive the Gibbs free energy of a complex and show that it is independent of the specific assembly trajectory.

4.8.1. Specific choice of kinetics

We start with an overview of the rates defined in the last sections. The elementary rate for a hybridization (Eq. 4.7), using Eq. 4.11, is given by

$$r_{\text{on}}(i,j,c) = \frac{1}{VN_A c^\circ} r_0(i,j) p_a(i,j) p_c(i,j,c). \tag{4.34}$$

Thus the rate constant (Eq. 4.25) is

$$k_{\text{on}}(i,j,c) = 2^{m-\delta_{ij}} p_a p_c r_0 \frac{1}{c^\circ}. \tag{4.35}$$

The rate constant $k_{\text{coll}}(i,j)$ of the elementary collision rate $r_{\text{coll}}(i,j)$ is given by (simple bimolecular reaction)

$$k_{\text{coll}}(i,j) := 2^{-\delta_{ij}} r_0 \frac{1}{c^\circ}. \tag{4.36}$$

We now choose the elementary rate for the dehybridization to be

$$r_{\text{off}}(i,c) = p_c r_0 e^{\beta \Delta G_b^\circ}, \tag{4.37}$$

and hence the rate constant for the dehybridization becomes (Eq. 4.27)

$$k_{\text{off}}(i,c) = p_c r_0 2^{m_1(1-\delta_{ij})} e^{\beta \Delta G_b^\circ}. \tag{4.38}$$

The justification for this choice will become apparent in the course of this and the following section. It will lead to a thermodynamically consistent energy model for the total free energy of a complex.

We further set the probability of a hybridization channels to be chosen equal to

$$p_c = \frac{1}{\Theta}, \tag{4.39}$$

where Θ is the number of possible hybridization channels, see Figure 4.7. Note that the following discussion of the energy model is independent of this choice as the hybridization and dehybridization rates are both proportional to p_c and will therefore cancel out.

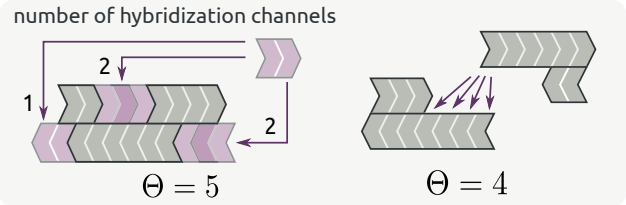


Figure 4.7.: Specific choice of the the probability of a hybridization channel to be chosen upon collision, $p_c = 1/\Theta$.

4.8.2. Free energy of a hybridization site

We now relate the ratio of the rate constants for (de)hybridization to the Gibbs free energy of a hybridization site c, $\Delta G^\circ_{\text{hyb}}(c)$:

$$\ln\left(\frac{k_{\text{off}}}{c^\circ k_{\text{on}}}\right) = \beta \Delta G_b^\circ + (\delta_{ij} - m + m_1(1-\delta_{ij}))\ln(2) - \ln(p_a) = \beta \Delta G^\circ_{\text{hyb}}(c), \tag{4.40}$$

We can further simplify the term $\delta_{ij} - m + m_1(1-\delta_{ij})$. Let us therefore consider a dehybridization of a complex of species S that leads to two complexes of species S_i and S_j.

- If S is rotationally symmetric, $(m_1 = 1)$ the dehybridization leads either to two different complexes $S_i \neq S_j$, where non is symmetric ($m = 0$, Fig. 4.8(a) top), or to two equal and symmetric complexes ($m = 2$, Fig. 4.8(a) bottom).
- If S is not symmetric ($m_1 = 0$) the dehybridization leads to two complexes of different species $S_i \neq S_j$, see Fig. 4.8(b). Either non of the resulting complexes is symmetric ($m = 0$) or one ($m = 1$).

We therefore have

$$\delta_{ij} - m + m_1(1-\delta_{ij}) = (m_1 - m) = \Delta m,\ \Delta m \in \{-1, 0, +1\}. \tag{4.41}$$

Thus, Δm appears as the difference in the number of rotationally symmetric complexes due to a hybridization. The values of Δm in dependence of m_1 and m are shown in Table 4.2. The expression of the Gibbs free energy Eq. (4.40) of a hybridization site then becomes

$$\beta \Delta G^\circ_{\text{hyb}}(c) = \beta \Delta G_b^\circ + \Delta m \ln(2) \underbrace{- \ln(p_a)}_{\nu}. \tag{4.42}$$

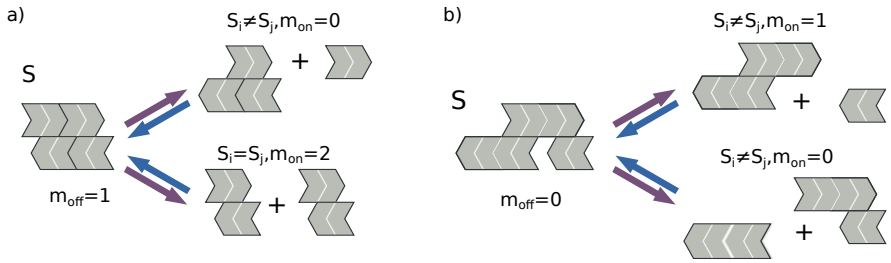


Figure 4.8.: m_1 indicates if a dehybridizing complex is symmetric ($m_1 = 1$) or non-symmetric ($m_1 = 0$). m indicates the number of symmetric complexes undergoing a hybridization, $m = 0, 1, 2$. (a) Dehybridization of a symmetric complex $m_1 = 1$: The dehybridization leads either to two different non-symmetric complexes $S_i \neq S_j$ ($m = 0$) or to two equal and symmetric complexes of species $S_i = S_j$ ($m = 2$). (b) Dehybridization of a non-symmetric complex: The dehybridization leads either to two non-symmetric complexes ($m = 0$) or one ($m = 1$) non-symmetric complex.

	$m_{on} = 0$	$m_{on} = 1$	$m_{on} = 2$
$m_{off} = 0$	0	-1	
$m_{off} = 1$	+1		-1

Table 4.2.: Possible values for Δm, the change in number of rotationally symmetric complexes due to the reaction.

The three different terms of the hybridization energy are (see [10]):

(i) $\beta\Delta G_b^\circ = \gamma l$ the standard binding free energy (in units of $k_B T$) due to base pairing and possible internal secondary structures of the hybridization (internal loops), see Chapter 3.

(ii) $\Delta m \ln(2)$ is a symmetry correction, which gives an energy reward of $-\ln(2)$, if the hybridization caused a reduction in the number of symmetric complexes, and an energy penalty of $+\ln(2)$ if the number of symmetric complexes is increased. The term is zero if there is no change in the number of symmetric complexes. This term corresponds to the symmetry correction of $+0.43 kcal/mol$ in the nearest neighbor data base *cf.* Ref. [52, 32].

(iii) p_a is the probability that a hybridization is performed upon collision. It can be thought of as the probability to form a first base pair p_{1bp} times the probability that the formation of the first base pair leads to zipping of the strands onto each other p_{zip}, (see [66]). We can therefore write $p_a = p_{1bp} p_{zip}$. Thus, we can interpret $\nu = -\ln(p_a) \geq 0$ as an energy penalty associated with the formation of the first base pairs upon hybridization.

The choice of the hybridization rate made in Eq. (4.37) led to an energy model for the hybridization site that can be interpreted in a physically meaningful manner and is conceptually equivalent to standard energy models for DNA and RNA as presented in Section 2.3.

4.8.3. Total Gibbs free energy of a complex

The total Gibbs free energy ΔG_{tot}° of a complex C of order n is the sum over the free energies of its $n-1$ hybridization sites $c \in C$ and is given by

$$\beta\Delta G_{tot}^\circ = \beta \sum_{c=1}^{n-1} \Delta G_{hyb}^\circ(c) = \sigma \ln(2) + (n-1)\nu + \sum_c \Delta G_b^\circ(c), \tag{4.43}$$

where $\sigma = 1$ if the complex is symmetric and $\sigma = 0$ if it is not, and ν is the initiation penalty of the duplex formation. $\Delta G^\circ_{\text{tot}}$ is independent of the specific assembly path (trajectory), i.e., the order in which the hybridizations are formed. It is only a function of the complex structure itself. Hence the choice of r_{off}, Eq. 4.37, led to a thermodynamic consistent energy model.

As an example, consider the two different assembly trajectories leading to the same complex configuration shown in Fig. 4.9. Even though the hybridization energies among the trajectories are different, as the number of symmetric complexes m differs after the second hybridization, the total free energies $\Delta G^\circ_{\text{tot}}$ of the resulting complexes are the same.

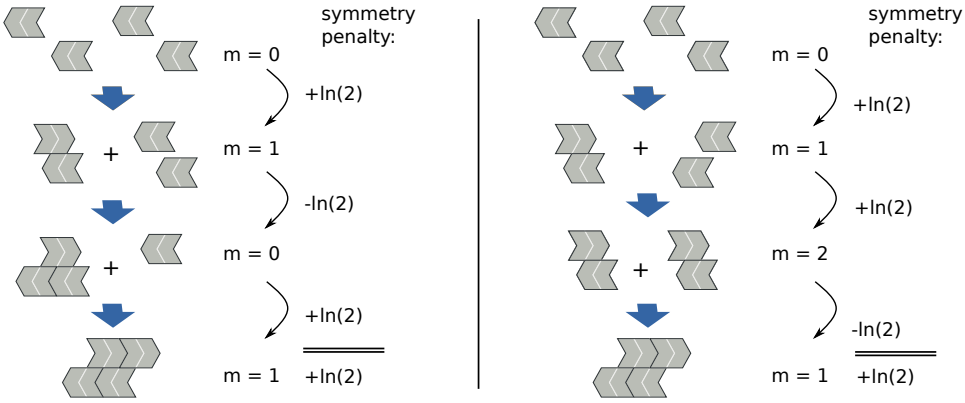


Figure 4.9.: The left and right assembly trajectory lead to the same final complex. The hybridizations among trajectories have different free energies, as the numbers of symmetric complexes after the second hybridization differs (left: $m = 0$, right: $m = 2$), but their sum yields the same $\Delta G^\circ_{\text{tot}}$. m is the number of rotationally symmetric complexes at each assembly step. Only the penalty term $\Delta m \ln(2)$ is shown, as the sum over the other two contributions ΔG°_b and ν is independent of the trajectory.

4.8.4. Thermodynamically consistent kinetics for nearest-neighbor models

For a thermodynamically consistent model, the parametrization of the (de-)hybridization kinetics is constraint by the binding energy. In particular, thermodynamic consistency implies that that the free energy $\Delta G^\circ_{\text{tot}}$ (see Eq. (4.43)) of any complex is independent of the assembly trajectory. Consequently, using an energy model that is in the spirit of common nearest neighbor models, *cf.* Eq. (4.44), introduces constraints on the available parameterizations.

More precisely, thermodynamic consistency with nearest-neighbors models requires that the energy associated with a hybridization channel, $\Delta G^\circ_{\text{hyb}}(c)$, is independent of θ. In general, $\Delta G^\circ_{\text{hyb}}$ is determined by the choice of the rate constants k_{on} and k_{off} via $\beta \Delta G^\circ_{\text{hyb}} = \frac{k_{\text{off}}}{c^\circ k_{\text{on}}}$. The rate constants and rates can be mapped onto each other via Eq. (4.25) and Eq. (4.27). Choosing a collision based Ansatz for the on-rate (Eq. (4.7)) $r_{\text{on}} = (V N_A c^\circ)^{-1} r_0 p_a p_c$, and writing the off-rate as $r_{\text{off}} = a e^{\gamma l}$, $\Delta G^\circ_{\text{hyb}}$ becomes

$$\Delta G^\circ_{\text{hyb}} = \gamma l + \Delta m \ln(2) + \ln\left(\frac{a}{p_a p_c r_0}\right). \tag{4.44}$$

In order to get a path independent $\Delta G^\circ_{\text{tot}}$, the acceptance probability p_a, the channel probability p_c and the effective collision rate must be chosen such that the third term of Eq. (4.44) becomes constant and hence independent of the (mutual) properties of the hybridizing complexes.

As discussed above we can interpret $\nu = -\ln(p_a)$ as the initiation energy. In the simplest form of the kinetics, the channel probabilities p_c must sum to unity over the θ different channels. The simplest and least biased choice is then $p_c = 1/\theta$. Hence, Eq. (4.44) becomes

$$\Delta G^\circ_{\mathrm{hyb}} = \gamma l + \Delta m \ln(2) + \nu + \ln\left(\frac{a\theta}{r_0}\right). \tag{4.45}$$

In order to have a thermodynamically consistent energy model that is still in the spirit of a nearest-neighbor model, we have to eliminate the factor θ from the binding energy. Otherwise, the total energy of a complex, Eq. (4.43), would depend explicitly on more detailed properties of all the strands involved.

There are two possible canonical choices of a and r_0 that can be made in order to set the last term to zero [2]: (i) $a = r_0/\theta$ or (ii) $a = 1$ and $r_0 = r_0^*\theta$, where r_0^* is a constant. Both choices lead to a thermodynamic consistent energy model but may be considered microscopically unsatisfactory: (i) This is the choice we used for our model. At first glance, it seems odd that the off-rate depends on the number of channels θ. However, this problem is intrinsic when using kinetics in nearest-neighbor type models. (ii) The second choice has the advantage that the θ-dependence is contained in the hybridization rate and is altogether absent from dehybridization. It has the microscopic advantage, that the hybridization rather than the dehybridization rate depends on θ. However, this is also physically questionable from the perspective of microscopic, diffusive dynamics (see below).

As we will see in Chapter 6, whatever the exact choice, as long as it is not exponential in the strand length, microscopic kinetic factors like θ only contribute subexponentially to the length-scales derived from the competition of time scales.

The primary reason for our choice (i) is computational. Having a collision rate that is independent of the exact nature of the complexes allows us to sample colliding pairs without considering their possible hybridizations *a priori*. Any other choice would massively increase the computational complexity because of the additional computation that needs to be performed for all pairs of species, *cf.* Fig. S4. After each hybridization we would have to update the whole reaction network, which scales as $\sim \langle\theta\rangle\,(N_n^{\mathrm{tot}} - 1)$, where $\langle\theta\rangle$ is the average number of hybridization channels, which is computationally impractical. In fact, implementing such a more complex kinetics along the lines of (ii) was our first attempt. However, for the above reasons, the algorithm then became prohibitively slow and complicated.

As explained in Sec. 4.6, our implementation is fast because it assumes a constant collision rate, which allows us to reduce all collision events to a single total collision reaction. Only after the colliding pair is drawn, we subsequently choose a channel and then update the possible hybridization and dehybridization reactions just in time.

Moreover, a closer look at the kinetics (ii) reveals that it is not necessarily more physical, as long as the underlying transport process is not further specified. For example, $r_0 = r_0^*\theta$ neglects the decrease in mobility with strand length as would be the case for regular diffusion. To illustrate this aspect more thoroughly, let us consider a simple model of two diffusing strands with diffusion coefficients D_1, D_2. We assign to the strands the hydrodynamic radii R_1, R_2 and assume that the strands undergo a reaction as soon as their distance becomes smaller than $R_1 + R_2$. The collision rate is then obtained via the Smoluchowski rate coefficient [1, 9, 27]

$$r_{\mathrm{coll}} = \frac{1}{V} 4\pi (R_1 + R_2)(D_1 + D_2). \tag{4.46}$$

[2] a remaining constant could be absorbed into ν

Using that the hydrodynamic radius is inversely proportional to the diffusion coefficient [44], Eq. (4.46) becomes

$$r_{\text{coll}} \sim \frac{1}{V} 4\pi \frac{(D_1 + D_2)^2}{D_1 D_2}. \tag{4.47}$$

Experimentally, the following relation between diffusion coefficient and length for single (double) stranded DNA has been found:

$$D \sim L^{-\nu}, \tag{4.48}$$

where $\nu = 0.45$ ($\nu = 0.67$) [28, 50]. Thus the collision rate is proportional to

$$r_{\text{coll}} \sim \frac{(L_1^\nu + L_2^\nu)^2}{L_1^\nu L_2^\nu} =: f(L_1, L_2, \nu). \tag{4.49}$$

This expression scales differently with L_1 and L_2 than $\theta = L_1 + L_2 - 1$, compare Fig. 4.10(left) with Fig. 4.10(right).

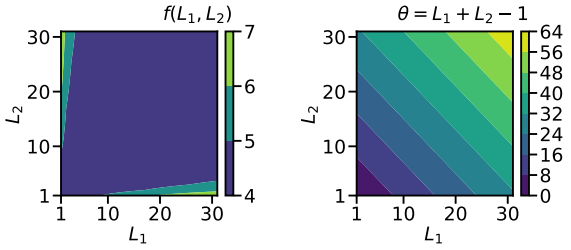


Figure 4.10.: (left) $r_{\text{coll}} \sim f(L_1, L_2)$ for $\nu = 0.45$. (right) number of channels $\theta(L_1, L_2)$. $f(L_1, L_2)$ and $\theta(L_1, L_2)$ have different scaling.

One potential solution to this apparent dilemma would be the introduction of intermediate states. This state would be characterized by two molecules that have collided but have not yet formed a hybridization complex. From this intermediate state, molecules can then either hybridize along a channel or go back into solution. In accordance with microscopic reversibility, the dehybridization has to pass through this intermediate state, which effectively allows the constituent strands of a complex to reassemble.

In that case, the factors weighing the hybridization into different channels could be completely arbitrary since the corresponding probabilities would not need to sum to unity but only determine the (average) lifetime of this intermediate state. Acceptance probabilities with a more general parameter dependence can be formulated independently from an (optional) "initiation energy" in a nearest-neighbor model. However, such a simulation would require many more (unknown) parameters and a different implementation.

5. Implementation of the simulation framework

5.1. Implementation of the simulation framework

In this section, we briefly discuss the implementation of the simulation framework and give some insights into the technical details. The simulation implements a Gillespie algorithm on the Markov chain of copy numbers as introduced in Sec. 4.1. The codebase is written in C++. It was jointly developed between Tobias Göppel and myself, where I took stronger responsibility for the architecture during the first years, and he took stronger responsibility implementing important speedups such as. e.g., the generation of a constant background. Further, I was mostly responsible for developing the software necessary to run and manage the data acquisition on the cluster.

5.1.1. Data structure of complexes

The main requirements of the data structure of the complexes are twofold: (i) The ability to quickly calculate all possible hybridizations between two complexes, and (ii) to allow for a simplistic update of the complex structure upon hybridization, dehybridization, or ligation. It turned out that these requirements are fulfilled well by splitting strands into segments as already presented in Section 3.2. A new segment starts whenever a new hybridization site in a complex starts or ends, *cf.* Fig. 5.1(a) (red dotted lines). A network of linked segments then represents a complex, *cf.* Fig. 5.1(b). A complex has two pointers onto the two end-segments of this quasi-linked list (seg1, seg2 in 5.1(b)), which define the entry points of custom iterators. We can then traverse the network by going to the left, right, or hybridized neighbor of a segment. Each end-segment is marked by a virtual end-segment token. For example, a single strand would correspond to a complex with one segment, whereas a duplex has minimally two and maximally four segments.

This data structure allows for a fast calculation of the possible hybridizations. For example, for the hybridization of a single strand onto a duplex, only the single-stranded end segments, if there exists one, must be considered.

Each segment class has a string member variable that stores its sequence information. It also allows for the labeling of individual complexes by assigning them a specific sequence, as used in Section A.5 for the sampling of trajectories.

Like already defined in Chapter 2, we call the set of allowed nucleotides alphabet and notate it by the letter $\mathcal{A}$.

The null model, which will be presented in Chapter 6 uses an energy model that is only length-dependent. In this case, it had been sufficient to store the length of each segment in an integer member variable of the latter. However, in order to keep the algorithms generic, we instead query the length of the sequence. Hence, the choice of alphabet does not matter in this case and we simply set $\mathcal{A} = \{A\}$.

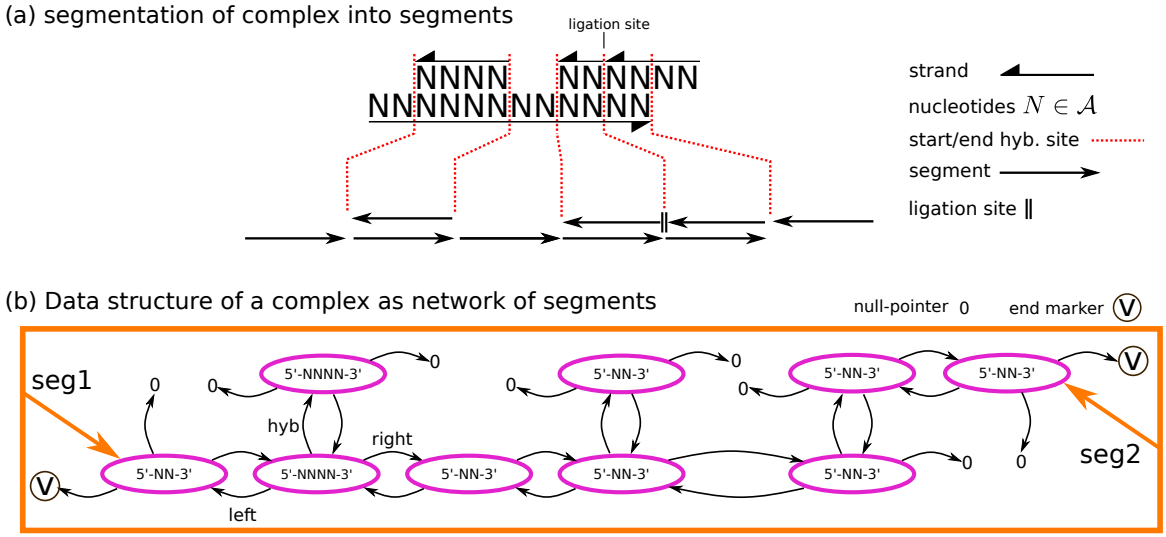


Figure 5.1.: (a) Mapping a complex consisting of hybridized strands onto segments. A new segment starts whenever a new hybridization site starts or ends (red dotted lines). $\mathcal{A}$ is the alphabet of the nucleotides. A ligation site is represented by two vertical lines between two adjacent segments. (b) A complex is basically a linked list of segments. Each segment has pointers to its neighboring segments: hyb, left and right. The pointer is 0 if the neighboring position is not occupied. The segments specifying the ends of the linked list, called seg1 and seg2, point to a unique segment called virtual end segment (v). The complex class itself has only pointers to seg1 and seg2. A custom iterator uses them as an entry point in order to iterate over the segments of the complex.

5.1.2. Container of species

If several complexes belonging to the same species are present in the system, we do not store them separately. Instead, we store one instance of the complex together with its copy number. Hence, a species is characterized by a complex and its associated copy number. We store all species present in the system in an unordered map, to which we refer to as species container. The unordered map allows for fast insertion and deletion of species via a key created by uniquely mapping a complex structure onto a string. When inserting a complex of species S_i into the species container, it is first checked if the species S_i already exists. If this is the case its copy number is increased, $N_i \to N_i + 1$. If the species is not yet contained in the species container, the complex is inserted with copy number $N_i = 1$. Drawing an individual complex from the species container because it is involved in a reaction means creating a copy of the species and reducing the copy number, $N_i \to N_i - 1$. If the copy number of the complex, which is chosen for the next reaction, is one, the complex is removed from the container.

5.1.3. Process flow of the simulation

We send an ensemble of jobs (simulations) to the computer cluster, managed by a Sun Grid Engine (SGE) queuing system. At the start of a simulation, the initial species are read in. All reactions are created and stored in a reaction container. The reaction container has an interface to a binary search tree, which is used to select a reaction based on its weight according to the Gillespie Algorithm (*cf.* Sec. 4.2). First, all monomolecular reactions (dehybridizations and ligations) are calculated and inserted into the reaction-container. The bimolecular reactions (hybridizations) do not need to be calculated individually. As described in Sec. 4.6, we condense them into a single total collision reaction, and therefore

only need to track the total number of complexes N_n^{tot} in the system. We also insert the total collision reaction into the reaction-container. Its rate is updated whenever the total number of complexes changes. The number of monomolecular reactions scales with the number of species present S_n^{tot}, whereas the number of hybridizations scales like $(S_n^{\text{tot}})^2$. By contracting the bimolecular reactions onto one single collision reaction, we reduce the computational complexity of the reaction selection step from $\approx \mathcal{O}(\ln(S_n^{\text{tot}} + S_n^{\text{tot}^2}))$ to $\approx \mathcal{O}(\ln(S_n^{\text{tot}}))$.

When the simulation is set up, one reaction is chosen from the container via the binary search tree based on the weight of its rate, and the system time is updated (*cf.* Sec. 4.2). If it is a monomolecular reaction, it can be performed directly. If the total-collision-reaction is chosen, a specific hybridization reaction must be generated. For that, two complexes are chosen from the species container, to perform the actual collision event. Next, the number of possible hybridizations Θ is calculated. If $\Theta = 0$, the algorithm returns to the selection step of a reaction. If $\Theta > 0$ a specific hybridization channel is chosen. We thereby implement the acceptance probability p_a and channel probability p_c as described in Sec. 4.4.

Independent of whether the picked reaction is a collision leading to a hybridization or a monomolecular reaction, the resulting complex structure must be obtained. In the following, we discuss the case of a monomolecular reaction. The bimolecular reaction is handled analogously: First, the complex undergoing the reaction is drawn from the species container. We then form the new complex-structure specified by the reaction and insert it into the species container as described in Sec. 5.1.2. If it belongs to a new species, we create all ligations and dehybridizations. As discussed in Sec. 4.8.1, the dehybridization rate for a hybridization site c is given by

$$r_{\text{off}}(i, c) = \frac{1}{\Theta_c} e^{\gamma l_c} \frac{1}{t_0}. \tag{5.1}$$

Hence the hybridization channels Θ_c need to be calculated for all hybridization sites. This calculation is implemented via a virtual dehybridization of the complex, see Fig. 5.2. The

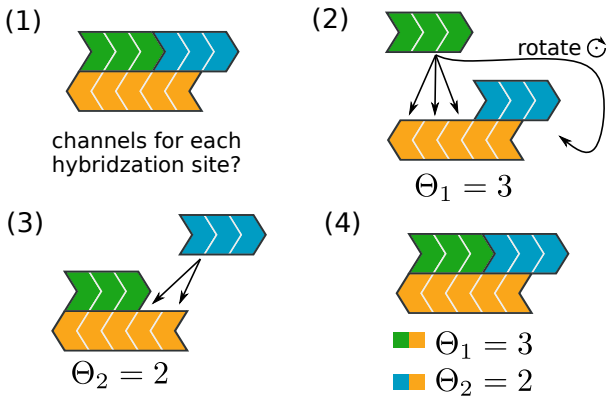


Figure 5.2.: Example of the calculation of the channel factor Θ_c for each hybridization site: (1) The complex we want to calculate the channel factors for is a triplex with two hybridization sites. (2) We virtually open the hybridization site between the green an the orange strand and calculate all possible hybridization channels. (3) Equivalent to (2) but for the blue-orange hybridization site. (4) summary of the calculated channel factors.

reaction is completed, and the program returns to the selection step of a reaction. The described process starts to repeat itself. After a set time interval, observables such as, e.g., species and length distributions are saved. As the SGE sets a time limit on the runtime of jobs, jobs resubmit themselves automatically until the simulation finishes. Therefore, when the runtime exceeds a specific value (1.5h), the state of the simulation is saved, and a new job referring to the unfinished simulation is passed to the SGE-queue. When the newly submitted job starts, the simulation's previous state is loaded, and the algorithm continues.

5.1.4. Speed up by the introduction of background species

We achieve a speedup of the simulation by the introduction of so-called background species. We calculate the hybridization equilibrium of the short building blocks of the reservoir (monomers, dimers) and insert them into a container of background species. Species of the background can not collide among each other, though they can collide with species in the regular species container. The concentrations (copy numbers N_i) of the background species are kept constant.

6. The null model*

6.1. Introduction

Like described in Chapter 1 a bottom up study of emergent phenomena of oligonucleotide self-assembly and growth via templated ligation was missing so far in the literature.

The goal of this work is to close this gap. To this end, we investigate a simple model that contains all the elementary processes important for growth processes governed by templated ligation. To focus on the self-assembly process alone, we ignore the sequence-dependence of hybridization in the present study. Instead, the binding energy of a hybridization site is proportional to its length, where the binding energy per nucleotide is negative and of the order of the thermal energy. As such, this model serves as a *null model* for other models which might include sequence-dependent binding.

Our main finding is that the competition of time-scales between (length-dependent) dehybridization, extension and a third global time scale leads to the emergence of a non-monotonous strand-length distribution. We show that this feature is generic and appears in experiments, *cf.* Fig. 6.1(a), and our simulation, *cf.* Fig. 6.1(b). Furthermore, we can predict the characteristic strand lengths that lead to strong catalytic behaviour and thus shape the non-monotonous strand-length distribution.

Figs. 6.1(c-f) schematically illustrate these ideas: Hybridization and bare ligation are combined into extension reaction which occurs at an effective rate r_{ext}, *cf.* Fig. 6.1(c). Above a characteristic strand length L^*, the rate of extension becomes larger than the rate of dehybridization. The extended duplex binds stronger and dehybridization is further suppressed. Another extension then becomes even more likely leading to a fast process we refer to as an *extension cascade*, *cf.* Fig. 6.1(d). Generically, extension cascades only stop when a persisting configuration is reached where no further extension is possible. When long duplexes eventually dehybridize, the released single strands recombine and trigger further extension cascades. The time-scale of combined extension and reassembly is limited by the dehybridization rate. When strands are constantly removed, a non-equilibrium steady state is established and the dehybridization rate competes with the outflux rate r_{out}. In contrast, in a transient non-equilibrium situation, dehybridization times compete with the observation time τ_{obs}. In both cases, the length dependence of dehybridization yields an associated length scale $L^\dagger$, which corresponds to the local maxima in strand-length distributions. Together, the characteristic length scales L^* and $L^\dagger$ define different dynamical regimes in the strand-length distribution, *cf.* Fig. 6.1(e). Applying the same arguments in an experimental scenario using random DNA sequences, we are able to predict and observe the emergence of a non-monotonous length distribution.

Our work is structured as follows: In Section 6.2 we cover the specific thermodynamic and kinetic model used. Section 6.3 first presents the core simulation results. In parallel, we develop the analytical theory necessary to understand the observed phase transitions and shape of the strand-length distributions. In Section 6.5 we present the results of a DNA-ligation experiment and interpret it using our theory. The implications of our results in the context of the origins of life are discussed in Section 6.6.

*This chapter got adapted from [106]

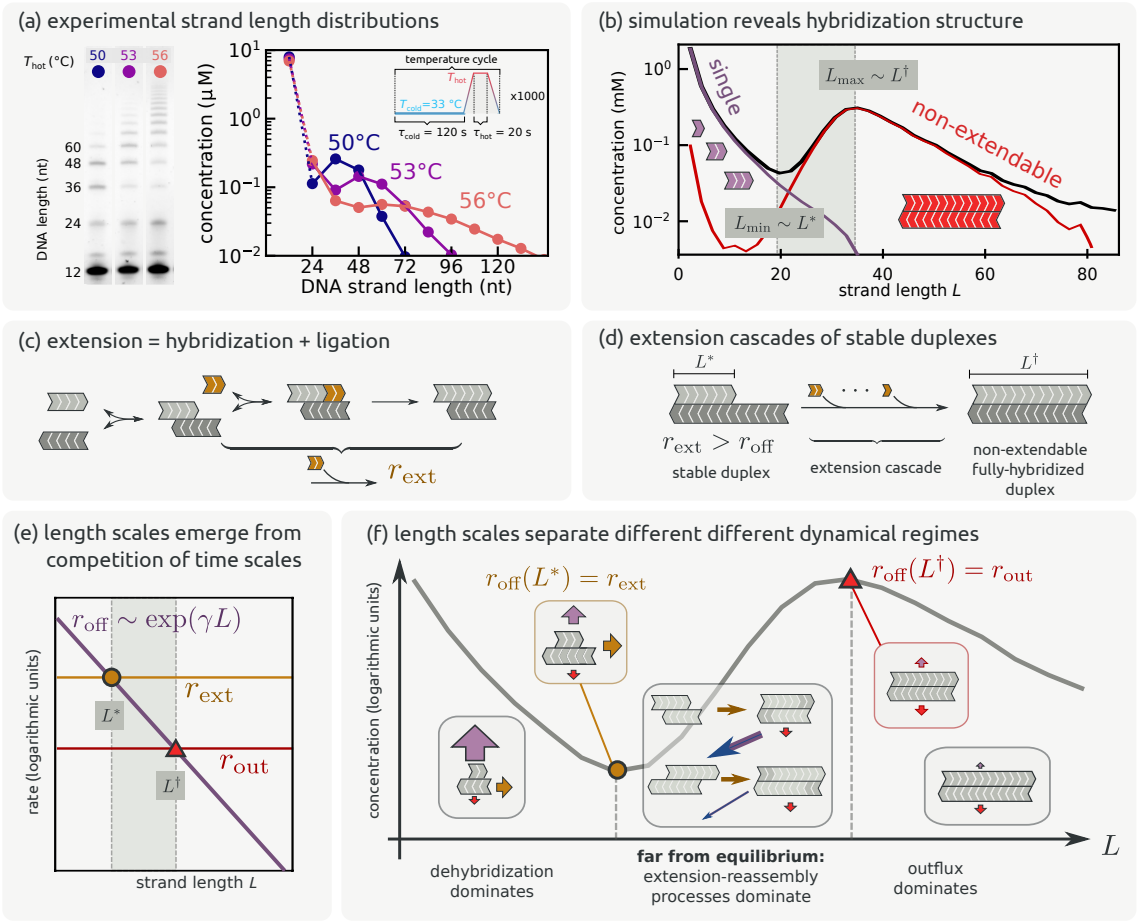


Figure 6.1.: An overview about this work. (a) Random DNA sequences with a ligase exhibit a non-monotonous strand-length distribution when subjected to temperature cycles. The positions of the local minimum and maximum depend on the temperature in the hot phase T_{hot}. (b) Stochastic simulations of our model reproduce this behavior. Strands to the left of the minimum ($L_{min} \sim L^*$) are dominantly single-stranded, while longer strands are fully hybridized and thus non-extendable. These double strands cause a local maximum at $L_{max} \sim L^\dagger$. (c) An important parameter of the dynamics is the emergent extension rate r_{ext}, which combines hybridization and ligation reactions. (d) A duplex is stable when the extension rate r_{ext} exceeds the dehybridization rate r_{off}. Extension cascades lead to non-extendable, fully-hybridized duplexes. (e) The dehybridization rate $r_{off}(L)$ relates length and time-scales: The minimal length scale for stable duplexes, L^*, is set by $r_{off}(L^*) = r_{ext}$. At the typical length scale for the fully-hybridized duplexes, $L^\dagger$, the dehybridization rate equals the global outflux rate, $r_{off}(L^\dagger) = r_{out}$. (f) Different regions in the strand length distribution exhibit different dynamical regimes. In the region $L^* \leq L \leq L^\dagger$ extension-reassembly dynamics dominate a dynamical regime that is far from equilibrium. The size of arrows scales with the magnitude of the associated rates r_{off} (purple, arrow pointing to top), r_{ext} (brown, arrow to the right), and r_{out} (red, arrow to the bottom).

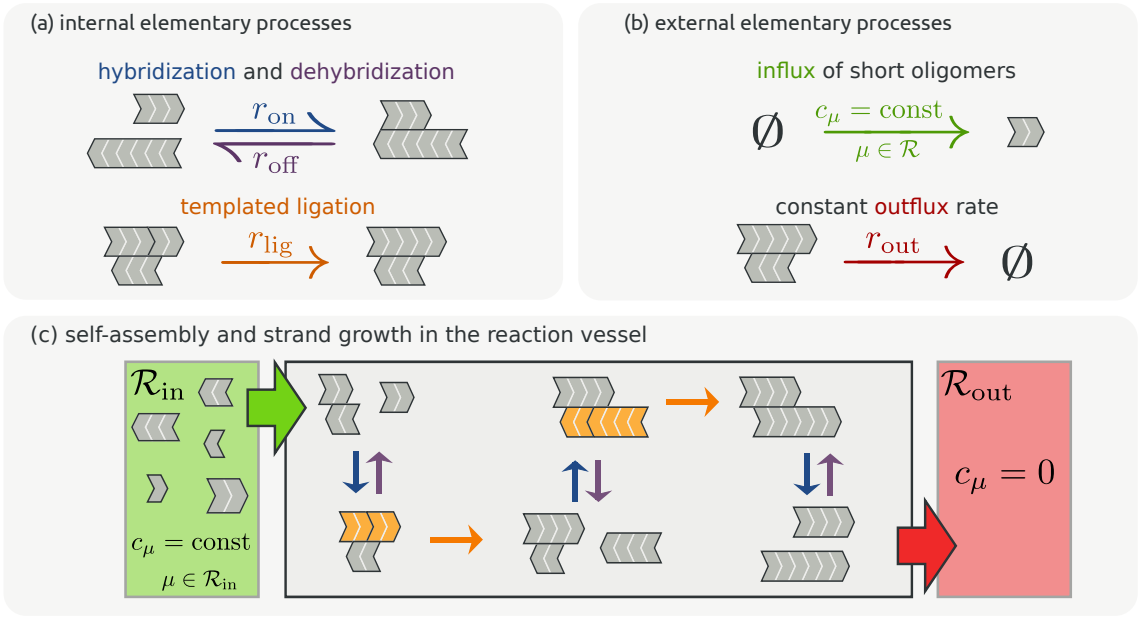


Figure 6.2.: (a) The internal elementary processes are hybridization, dehybridization, and templated ligation with corresponding rates r_{on}, r_{off}, and r_{lig}. (b) The external elementary processes couple the system to its environment. Short strands of length $L = \mu$ for $\mu \in \mathcal{R}$ are chemostated via the coupling to an external reservoir of initial building blocks at fixed concentrations c_μ. All complexes leave the system at a constant rate r_{out}. (f) Short strands that enter the reaction vessel from the reservoir $\mathcal{R}_{\text{in}}$ are the initial building blocks of the system. Inside the vessel, strands form various complexes via hybridization and dehybridization. Subsequent ligation leads to longer strands. All complexes can leave the system by a constant outflux rate r_{out}, which can be interpreted as a coupling of the system to an infinite empty reservoir $\mathcal{R}_{\text{out}}$.

6.2. Model and simulation method

In this section, we formulate the specific model used in this chapter to study the self-assembly and growth of informational polymers via templated ligation.

We start building the model in a bottom-up approach: Oligonucleotides (or "strands") are either free in solution or part of a hybridization complex. In order for two strands to become ligated, they need to be hybridized next to each other on a third strand. The most simple configuration allowing for templated ligation is thus a triplex see Fig. 3.1 (b)(left). However, neither do all triplex configurations allow for templated ligation, nor does templated ligation only occur in triplexes.

The internal reactions in our model are hybridization and dehybridization as well as templated ligation, *cf.* Fig. 6.2 (a). While the bare ligation rate r_{lig} is assumed to be constant, the dehybridization rate r_{off} depends on the binding energy of a complex.

The binding energy of two strands is proportional to the length of the hybridization site and characterized by the dimensionless binding energy per nucleotide γ. We also include a "bounded" variant of our model, where the dehybridization rate cannot become smaller than some prescribed "cutoff" rate r_{cut}.

The system is coupled to an environment which regulates the in- and outflux of strands and complexes, *cf.* Fig. 6.2 (b). Since the goal is to study the self-assembly of longer strands and complexes from short oligomers, we couple the system to an external reservoir $\mathcal{R}_{\text{in}}$ which keeps the concentration of initial building blocks constant. All complexes are subjected to a stochastic outflux prescribed by a constant rate r_{out}, which can be interpreted as the coupling to an infinite empty reservoir $\mathcal{R}_{\text{out}}$.

Consequently, the stochastic dynamics of self-assembly and growth are simulated *ab initio* with the dynamics schematically illustrated in Fig. 6.2 (c).

6.2.1. Complexes and strands

As described in Chapter 3 the basic element of our dynamics is a directed oligomer called a strand, which consists of a number of covalently bound nucleotides. Strands have distinct upstream and downstream directions pointing towards the 5′ and 3′ oligonucleotide ends. The only form of secondary structure taken into account here is the hybridization of strands to form hybridization complexes. Self-folding is excluded from our model. Complexes generally consist of an arbitrary number of strands and are called complexes of order n. A complex of order $n = 1$ is a single strand and will also be called an m-mer, where we explicitly refer to monomers, dimers, trimers and tetramers for the cases of $m = 1, 2, 3, 4$, respectively. The simplest nontrivial complex is a duplex consisting of $n = 2$ strands. Because of the linear topology, an nth-order complex has exactly $n - 1$ distinct hybridization sites. Each hybridization site involves exactly two strands and is characterized by an overlap length l, see Fig. 6.3. Thereby internal loops structures are excluded from the model, and there are no other secondary structures than helices.

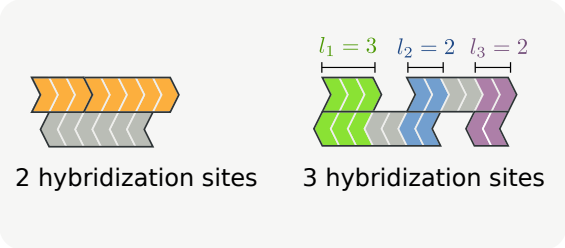


Figure 6.3.: Examples of higher order complexes with multiple hybridiziation sites: (left) A triplex with a templated ligation site. (right) A complex of order 4 with not ligation site.

6.2.2. Elementary processes and parameters

Fig. 6.2 (c)(f) gives an overview of the model dynamics: Short single strands enter a well-mixed reaction vessel of volume V. Within the reaction vessel, strands hybridize to form complexes. Strands in suitable complexes may undergo templated ligation. All complexes leave the reaction vessel at a constant rate, mimicking a flow-reactor or (fast) serial dilution.

The internal elementary processes (*i.e.*, reactions), are hybridization, dehybridization and templated ligation, see Fig. 6.2 (a). Hybridization and dehybridization are assumed to be elementary and reversible reactions occurring with rates r_{on} and r_{off}, which are defined for single hybridization sites.

Thermodynamic consistency [54, 80] connects the elementary rates r_{on} and r_{off} for hybridization and dehybridization to the standard binding free energy ΔG_b° of a hybridization site:

$$\frac{r_{\text{off}}}{r_{\text{on}}} = (V N_A c^\circ) e^{\beta \Delta G_b^\circ}, \tag{6.1}$$

where $\beta = (k_{\text{B}} T)^{-1}$, k_{B} is Boltzmann's constant and T denotes the (absolute) temperature, N_A is the Avogadro constant and $c^\circ = 1\,\text{mol/l}$ is the standard concentration.

When two strands of length L_1 and L_2 are hybridized adjacently on a third strand, they can ligate and become a new strand of length $L_1 + L_2$. This process is called templated-ligation, with the third strand understood as the template. In this model, templated ligation is the only process through which strand length is increased. We assume this process to be irreversible and to occur with a bare ligation rate r_{lig}. While no process in nature is truly irreversible, this simplification is justified under the assumption that either (i) single strands

enter the system in an activated form providing chemical energy $\Delta G^\circ_{\text{lig}} \gg k_B T$ (e.g. [69]) or (ii) if templated ligation is catalyzed under the consumption of energy by an enzyme (a ligase) [100, 63, 48, 73]. The irreversibility of the ligation reaction is a hallmark of a non-equilibrium system, with an energy source provided by reactive building blocks or the chemical fuel consumed by a ligase. Since template-free random ligation is a much slower process than templated ligation [64], the former process is ignored in this simple model.

In addition to the internal reactions, two external reactions account for the connection of the system with its environment, *cf.* Fig. 6.2 (b): (i) Coupling the system to a large reservoir keeps the concentrations c_m of single strands with lengths $m \in \mathcal{R}$ constant. (ii) Strands and complexes are removed from the system with a constant rate denoted as r_{out}, which is the same for all complexes. Thermodynamically, our system is open and coupled to different external reservoirs. It and thus allows for non-equilibrium stationary states.

6.2.3. Thermodynamics and kinetics of hybridization

For real oligonucleotides, the binding energy ΔG°_b of hybridization depends on the number and nature of paired nucleobases at and adjacent to the hybridization site. Typically, a nearest-neighbor model is used for calculating binding energies [52, 32] as presented in Chapter 2.3. The possible structures that can potentially be represented by our simulation were presented in Chapter 3:

In this chapter, however, a hybridization site and thus its binding energy is only characterized by the overlap length l, *cf.* Fig. 6.3. The binding energy of a hybridization site is thus given by

$$\beta \Delta G^\circ_b(l) = \gamma l, \tag{6.2}$$

where $\gamma < 0$ is a parameter that gives the (negative) binding energy per unit length in units of the thermal energy $k_B T$.

Via Eq. (6.1), the binding energy determines only the ratio of r_{on} and r_{off}. An additional kinetic parameter is needed for a full parametrization of these rates. We therefore introduce the rate of collision between two complexes $r_{\text{coll}} = (V N_A c^\circ t_0)^{-1}$, where $t_0 = (r_0)^{-1}$ is a microscopic, intensive collision time scale, see Sec. 4.5.2. In what follows, all times are measured units of t_0.

In general, two colliding complexes can form multiple hybridization configurations via Θ distinct hybridization channels (see Fig. 4.7). The conditional probability of choosing one of these channels is

$$p_{\text{hyb}} = \begin{cases} 0, & \Theta = 0 \\ 1/\Theta, & \Theta > 0 \end{cases}. \tag{6.3}$$

. We thereby set the acceptance probability $p_a = 1$ and combined it with the channel probability $p_c = \frac{1}{\Theta}$ to p_{hyb}. With that, the hybridization rate for a given hybridization channel reads

$$r_{\text{on}} = r_{\text{coll}}\, p_{\text{hyb}}, \tag{6.4}$$

whereas the dehybridization rate

$$r_{\text{off}} = \frac{1}{\Theta} e^{\gamma l} \tag{6.5}$$

is given by Eqs. (6.1) and (6.2).

In reality, the collision rate depends on the properties of the colliding complexes, the properties of the solvent and temperature. This choice of kinetics can be interpreted as an activation-controlled regime, where the activation barrier is assumed to be constant, *cf.*

Ref. [65]. A parametrization where the binding energy γl is attributed to the dehybridization rate r_{off} is a common kinetic assumption, which has been qualitatively confirmed by experiments, *cf.* Refs. [4, 67, 93]. Implementing the kinetic sampling in the way described above reduces the computational complexity due to hybridization massively while still sampling all complex configurations in a thermodynamically consistent way.

With these elementary rates, the total binding energy of a complex C is found to be

$$\beta\Delta G^{\circ}_{\text{tot}}(C) = \gamma \sum_{i\in C} l_i + \sigma \ln(2), \tag{6.6}$$

where we sum over all hybridization sites in the complex, *cf.* Eq. (4.43). Recall that the final term $\sigma\ln(2)$ is a "symmetry penalty" that occurs if the complex is rotationally symmetric ($\sigma = 1$) and is zero ($\sigma = 0$) otherwise. It also appears in the standard databases for oligonucleotide binding energies [32, 52] and was derived in detail in Section 4. Thermodynamically, it is equivalent to a decrease in the (standard internal) entropy by a factor of $\ln(2)$ due to the rotational symmetry.

In addition to our standard model (where the binding energy is strictly proportional to the overlap length, Eq. (6.2)), we also consider a "bounded" variant of our model. In this bounded model, the dehybridization rate cannot become smaller than a minimal rate r_{cut}, such that $r_{\text{off}} = r_{\text{cut}}$ if $e^{\gamma l}/\Theta < r_{\text{cut}}$. The bounded model can be thought of as an effective implementation of a system that is subjected to an external mechanism that causes dehybridization of *all* complexes with a timescale of $\tau \sim (r_{\text{cut}})^{-1}$. Such a situation can be realized, for example, in a thermocycler, a "thermal trap" situated in the vicinity of a hydrothermal vent or be the consequence of day–night or other naturally occurring cycles, *cf.* Refs. [74, 97, 91, 48, 73, 102].

6.2.4. Standard choice of parameters

In what follows, we first discuss a model where the initial building blocks entering from the reservoir are dimers only. If not indicated otherwise, the dimer concentration is fixed at $c_2 = 2\,\text{mM} = 2{\times}10^{-3}c^{\circ}$. In what follows, values of the concentrations will always be stated in molar units. With this concentration, the reaction volume is chosen such that this corresponds to 10^4 single-stranded dimers constantly present, which is a much larger system than used in previous studies [58, 36, 82, 70].

This dimer-only model is the simplest model that allows for templated ligation and makes analytical considerations easier. As we show below, all the features that are important in the context of this work are generic with regard to the composition of the reservoir. That being said, the dimer-only model has a special symmetry since all possible strands have an even length. This leads to a particular shape of the tail of the distribution, which we will discuss in more detail in Section 6.4.5.

Our remaining parameters are thus the binding energy per unit length, γ, the ligation rate r_{lig}, and the outflux rate r_{out}, with values specified in units of $r_0 = (t_0)^{-1}$. If not otherwise stated, their standard values are $\gamma = -0.5$, $r_{\text{lig}} \approx 2.5 \cdot 10^{-3}$ and $r_{\text{out}} = 5{\times}10^{-9}$. In the bounded model, the cutoff rate r_{cut} is a further optional parameter.

6.3. Simulation results and analysis

The main observable in this work is the strand-length distribution $\rho(L)$, shown in Fig. 6.1(a,b). It expresses the concentration of a strand of length L, irrespective whether it is part of a

complex or free in solution. Note that $\rho(L)$ should not be confused with the concentration c_m of single strands of length m.

Unlike the concentrations of all possible chemical species, the strand-length distribution is an observable that is much more tractable. Quantitative experimental methods (like gel electrophoresis used in this work) provide access to the strand-length distribution, *cf.* Refs. [85, 100]. Notice, that the strand-length distribution on its own does not contain information about the structure of the complexes. However, within our simulation framework we have full access to this information.

6.3.1. Self-enhancing catalysis leads to long-tailed distributions

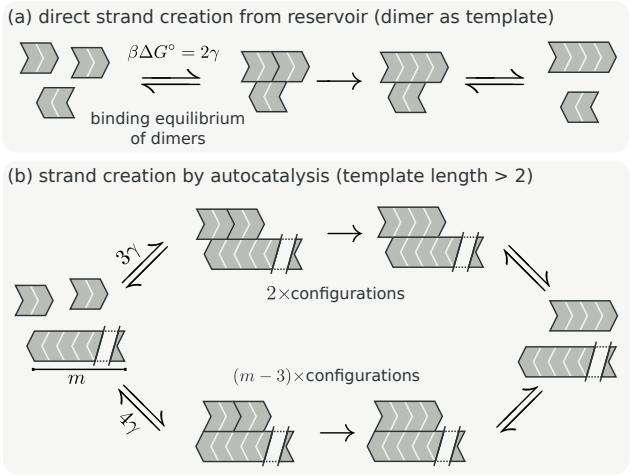


Figure 6.4.: (a) Formation of a tetramer from the dimer background. A total overlap of two leads to a total binding energy of $\beta\Delta G^\circ = 2\gamma$. (b) Templated ligation of dimers on an m-mer. There are two overhanging configurations with $\beta\Delta G^\circ = 3\gamma$ and $m-3$ configurations with $\beta\Delta G^\circ = 4\gamma$.

Self-assembly via templated ligation is a self-enhancing process, where long strands facilitate their own formation. In general, this process competes with a mechanism that leads to their destruction by strands breaking apart ("cleavage", [74, 58, 82]) or being removed from the system ([100, 92, 96]). Our model, featuring an outflux with a constant rate r_{out}, implements the latter scenario.

For large outflux rates, strands remain inside the reaction volume only for short amounts of time. Thus any strand participates in only a few or even no templated ligations. The resulting stationary length distribution is therefore expected to be short-tailed with few long strands present.

In contrast, for a small outflux rate, strands spend more time inside the system and thus have a higher chance to serve as a template or to get ligated to another strand, eventually leading to the formation of longer strands. Longer strands again allow for larger hybridization sites and in turn, become better templates.

Consequently, we expect the existence of a crossover value for the outflux rate $r_{\text{out}} = r_{\text{out}}^{\text{c}}$, where the formation of longer strands is dominantly self-enhancing, leading to a qualitatively different strand-length distribution.

Under the assumption that (i) short-tailed distributions are dominated by the smallest building blocks and (ii) that the time scales of the dehybridization of these small building blocks are small compared to the time scale of ligation we can derive the value of this cross-over rate:

Consider the total concentration $\rho_>$ of strands with a length larger than two, *i.e.*, strands that are not provided as building blocks by the reservoir. In a steady state we have

$$0 = \partial_t \rho_> = \phi - \rho_> r_{\text{out}}, \tag{6.7}$$

where ϕ is the concentration flux indicating processes by which $\rho_>$ grows, namely the formation of tetramers from dimers. Notice that the formation of strands with $L \geq 4$ does not change $\rho_>$. In general, this templated ligation can happen in all complex configurations with two dimers that are adjacently hybridized on a third strand. Ignoring higher-order complexes, we assume that the most important contribution to the production of longer strands arises from a ligation reaction happening at triplexes consisting of two dimers and another templating strand of length $L \geq 2$, see Fig. 6.4.

As the hybridization dynamics of dimers are fast, we assume it to be approximately at equilibrium. This means that the ratio of the concentration of a triplex and its constituents is determined by its binding energy. Under that assumption, the ligation flux for triplexes consisting of dimers only is $\phi_2 = (c_2)^3 e^{-2\gamma} r_{\text{lig}}$, see Fig. 6.4 (a). In contrast, the ligation corresponding to templates of length $m > 2$ is

$$\phi_m = (c_2)^2 (2e^{-3\gamma} + (m-3)e^{-4\gamma}) c_m r_{\text{lig}}, \tag{6.8}$$

where we took into account the different configurations of the relevant triplexes, see Fig. 6.4 (b).

We separate the ligation flux into two components, $\phi = \phi_2 + \phi_>$. The first term, ϕ_2, only involves the building blocks provided by the reservoir. In contrast, the second term $\phi_> := \sum_{m>2} \phi_m$ involves the concentrations of whose formation is catalyzed by longer strands and is thus a self-enhancing process. Assuming that the length distribution is still short tailed and that most of the strands are single stranded configuration we can approximate $\rho_> \sum \rho_m \approx \rho_4 \approx c_4$. Thus the steady state condition approximately

$$0 = \phi_2 + \phi_4(c_4) - c_4 r_{\text{out}}. \tag{6.9}$$

We assume the transition to occur when the concentration of tetramers becomes large enough such that the auto-catalytic term becomes equal to the strand growth from the background, $\phi_>(c_4^c) = \phi_2$, which yields the condition

$$0 = 2\phi_>(c_4^c) - c_4^c r_{\text{out}}, \tag{6.10}$$

for the transition to occur. Plugging in the expression for ϕ_4, Eq. (6.8), the latter condition becomes

$$0 = 2(c_2)^2 (2e^{-3\gamma} + e^{-4\gamma}) r_{\text{lig}} - r_{\text{out}}, \tag{6.11}$$

which we can use to determine the critical outflux rate $r_{\text{out}}^{\text{c}}$ for the transition

$$r_{\text{out}}^{\text{c}} = 2(c_2)^2 \left(e^{-4\gamma} + 2e^{-3\gamma}\right) r_{\text{lig}}. \tag{6.12}$$

We now probe the stationary distribution of the bounded ($r_{\text{cut}} = e^{-6}$) and unbounded model using simulations for various values of the outflux rate r_{out}. Simulation results for the standard model are shown in Fig. 6.5(a). For comparison, Fig. 6.5(b) shows the analogous results for the bounded model, where the dehybridization rate cannot become smaller than some r_{cut}. As a consequence, the (slow) time scales of the effective extension

reactions (characterized by an inverse rate r_{ext}^{-1}) are well separated from the time scales of hybridization and dehybridization of any complex.

Since the derivation of Eq. (6.12) does not rely on the dynamics of long strands, we expect the same transition from short to long-tailed strands to occur in both situations. The first column of Fig. 6.5 shows the stationary strand-length distribution in both models. For sufficiently large outflux rates the resulting short-tailed strand-length distribution looks qualitatively and quantitatively similar between the two models. In both cases, the curve for t8he crossover outflux rate $r_{\text{out}}^{\text{c}} = 3.24\times10^{-7}$ obtained from Eq. (6.12) is indicated in orange.

The long-tailed distributions obtained for small outflux rates differ significantly: In the standard model, Fig. 6.5(a), we see the emergence of a characteristic hump leading to a local minimum and maximum. In contrast, the long-tailed distributions in the bounded model, Fig. 6.5(b), still decay monotonously.

The reason for this behavior becomes clear from the second column, where we sketch the dependence of the (effective) rates of the important processes on strand length. In the unbounded model, the dehybridization rate r_{off} intersects the horizontal lines corresponding to constant extension and outflux rates at two distinct length scales L^* and $L^\dagger$. This behavior already hints at the two emergent length scales L_{min} and L_{max} in the strand-length distribution. In contrast, this intersection does not occur for the bounded model, where there are no distinct scales visible in the strand-length distribution.

An analogous argument to Eq. (6.12) for the transition from long to short tales was made by Maslov and Tkachenko in Ref. [74]. In that work, the authors studied templated ligation in a different model, where long strands break by cleavage. In contrast to a stochastic simulation, they numerically solved an effective set of ordinary differential equations showing a phase transition from a short to long-tailed length distribution.

The crucial difference between their work and our standard model is that in their model ligation is always the slowest process. Similar to the situation used in the bounded model, they motivated an effective slowest dehybridization rate by means of a cyclic process that separates all strands on some time-scale $\tau_{\text{cycle}} \sim r_{\text{cut}}^{-1}$. Importantly, the long-tailed distributions obtained in their model were also monotonically decaying, similar to the results obtained in the bounded model.

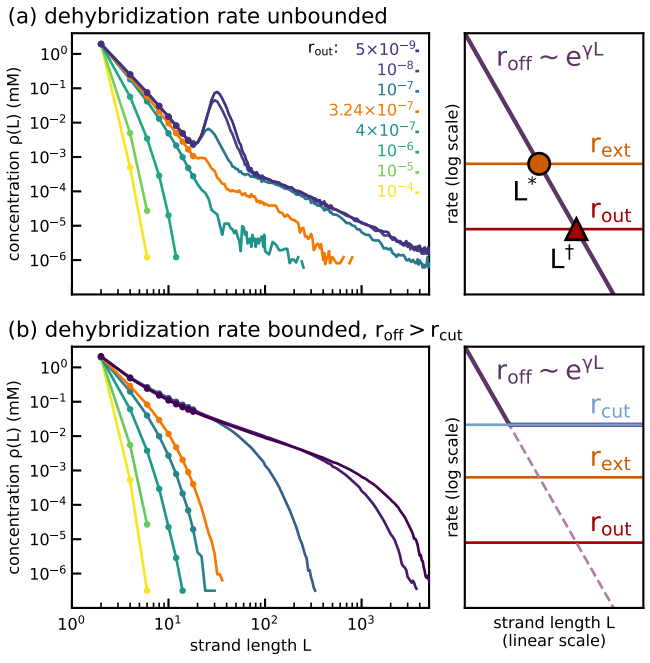


Figure 6.5.: Stationary strand-length distributions for the standard (unbounded) model (a) and its bounded variant (b) for different values of the outflux rate r_{out}. In the bounded model, dehybridization cannot become smaller than $r_{\text{cut}} = 0.05$. Dehybridization is thus faster than ligation ($r_{\text{lig}} = 2.5\times10^{-3}$) for all lengths. In both models, the length distributions (left) develop long tails when decreasing the outflux rate r_{out}. The orange curves corresponds to a system where the outflux rate takes the crossover value $r_{\text{out}} = 3.24 \times 10^{-7}$, *cf.* Eq. (6.12). For outflux rates below the transition value, the unbounded model exhibits a non-monotonous strand-length distribution with a local minimum at L_{min} and local maximum L_{max}. Decreasing the outflux rate does not affect the minimum but increases both the position and the value of maximum.

6.4. Estimation of the outflux rate at the transition from short- to long-tailed distributions

6.4.1. Competition of time scales enables extension cascades and persisting complexes

For the rest of this work, we focus on our standard (unbounded) model with sufficiently low outflux rates, such that the strand-length distribution exhibits the non-monotonous behavior shown in Fig. 6.5(a). From the discussion, it already became clear that this behavior occurs for complexes for which dehybridization is not necessarily the fastest process.

To be more precise, consider the fates of duplexes illustrated schematically in Fig. 6.1(f). If its binding energy is close to zero, it will dehybridize quickly. In contrast, if the binding energy has a large (absolute) value, the complex is stable, and extensions with a third strand become more probable. The extended complex is then even more stable, and another extension becomes even more probable. We call this phenomenon an *extension cascade*. Extension cascades originate from the fact that time scales of dehybridization and extension are not clearly separated.

Disregarding dehybridization and outflux for now, an extension cascade only stops when a configuration is reached where no further extension is possible. In our model, this is only the case for a fully-hybridized duplex, *i.e.*, a duplex consisting of two maximally overlapping strands with the same length. Since extensions cannot occur in this configuration, fully-hybridized duplexes persist for long times. Eventually, the fate of such a long-lived complex is determined by either dehybridization or outflux. The competition between these two processes leads to non-monotonous strand-length distributions.

6.4.1.1. A closer look at the structure of complexes

We partition complexes into different classes by distinguishing between single strands, duplexes, and higher-order complexes, *cf.* Fig. 6.6. We then further subdivide duplexes according to their overhangs, *i.e.*, the length of the single-stranded segments at their ends. In the dimer-only model, where all strands are of even length, we consider the parity of an overhang. Zero-parity duplexes are fully hybridized and have no overhang. In contrast, duplexes with odd and even parity have overhangs with an odd and even length, respectively.

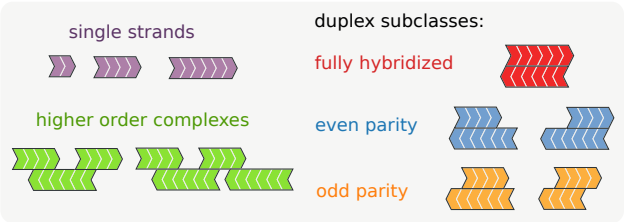


Figure 6.6.: Strands are grouped according to the order of the hybridization complex they belong to. In addition, duplexes are distinguished by their parity: Fully-hybridized duplexes have zero parity, whereas duplex with odd and even overhangs have odd and even parity, respectively.

Extension cascades only reach a terminal fully-hybridized duplex when they start from even duplexes. Duplexes with odd parity will undergo quasi-infinite extension cascades where the role of "primer" and "template" changes with each extension.

For our standard set of parameters, Fig. 6.7(a,b) shows the strand-length distribution partitioned into these sub-classes. As expected, short strands are predominantly single-stranded. In contrast, the peak of the length distribution is dominated by fully-hybridized duplexes. The effect of infinite extension cascades is indeed visible in the long tails. We call

the complexes of the long tail with a length C much larger than $L_{\max}$, $C \gg L_{\max}$ elongators. Also, we see that higher-order complexes are much less abundant and do not contribute significantly to the shape of the strand-length distribution.

The minimum of the distribution appears due to the increase of the concentration of fully-hybridized duplexes at a characteristic length scale $L^* < L_{\min}$. Fig. 6.7 (c) shows that this is the typical length-scale where duplexes become stable enough that extension cascades start playing a role.

6.4.1.2. Characterization and kinetics of duplexes

Since the dehybridization rate depends on the length of the hybridization region, it connects time scales with length scales. As such, the characteristic scales L^* and $L^\dagger$ also divide the strand-length distribution into the different dynamical regimes depicted in Fig. 6.1(f). Since the strand-length distribution is dominated by single strands and duplexes, we consider the kinetics of duplexes in more detail.

A duplex consisting of two strands S_1 and S_2 with corresponding lengths L_1 and L_2 is fully characterized by the three-tuple $D := (L_1, L_2, o_1)$. The number o_1 is the (positive or negative) overhang of strand S_1 on its 3′ end, see Fig. 6.8. The overhang o_2 at the 3′-end of S_2 then obeys $0 = L_1 - L_2 - o_1 + o_2$. From these numbers, the overlap which determines the binding energy, is given as $l = \frac{1}{2}(L_1 + L_2 - |o_1| - |o_2|)$. When the two strands collide, they can form $\Theta = L_1 + L_2 - 1$ different configurations. Applying this to Eq. (6.5), the dehybridization rate of a duplex is given as

$$r_{\text{off}}^{\text{dupl}}(D) = \frac{1}{L_1 + L_2 - 1} e^{\gamma l(D)}. \tag{6.13}$$

An extension with an m-mer is the combined process of an m-mer hybridizing next to a strand of a duplex D. Assuming that this happens at the non-zero overhang o_i of the duplex, a triplex T_i is formed as an intermediate where the subsequent templated ligation may occur. The length of the hybridization site of the m-mer with the duplex is given by $z_i = \min(|o_i|, m)$. In order to calculate an effective rate for this process, we assume that the dynamics of the second hybridization *is fast compared to the bare ligation rate*. Moreover, we assume that the length of the hybridization site z_i is small enough, such that the dehybridization rate of the m-mer is much larger than the dehybridization rate of the duplex itself.

Under these conditions, we may assume the hybridization of a duplex and a (short) m-mer into a triplex T_i to be in equilibrium. The index $i \in (1, 2)$ distinguishes the two ligatable triplexes that can be formed by the m-mer attaching to the overhang o_i of the duplex. Thus we obtain

$$c_{T_i} = c_D c_m e^{-\gamma z_i}. \tag{6.14}$$

With this we define the effective extension rate with an m-mer as the ratio of the rate of ligations from that triplex and the duplex concentration, *i.e.*, $r_{\text{ext},m} = r_{\text{lig}} c_T / c_D$. Using Eq. (6.14) and taking into account that in general there are two ligation sites ($i = 1, 2$), the extension rate with a m-mer reads

$$r_{\text{ext},m}(D) = r_{\text{lig}} c_m \sum_{\substack{i \in \{1,2\} \\ o_i \neq 0}} e^{-\gamma z_i}. \tag{6.15}$$

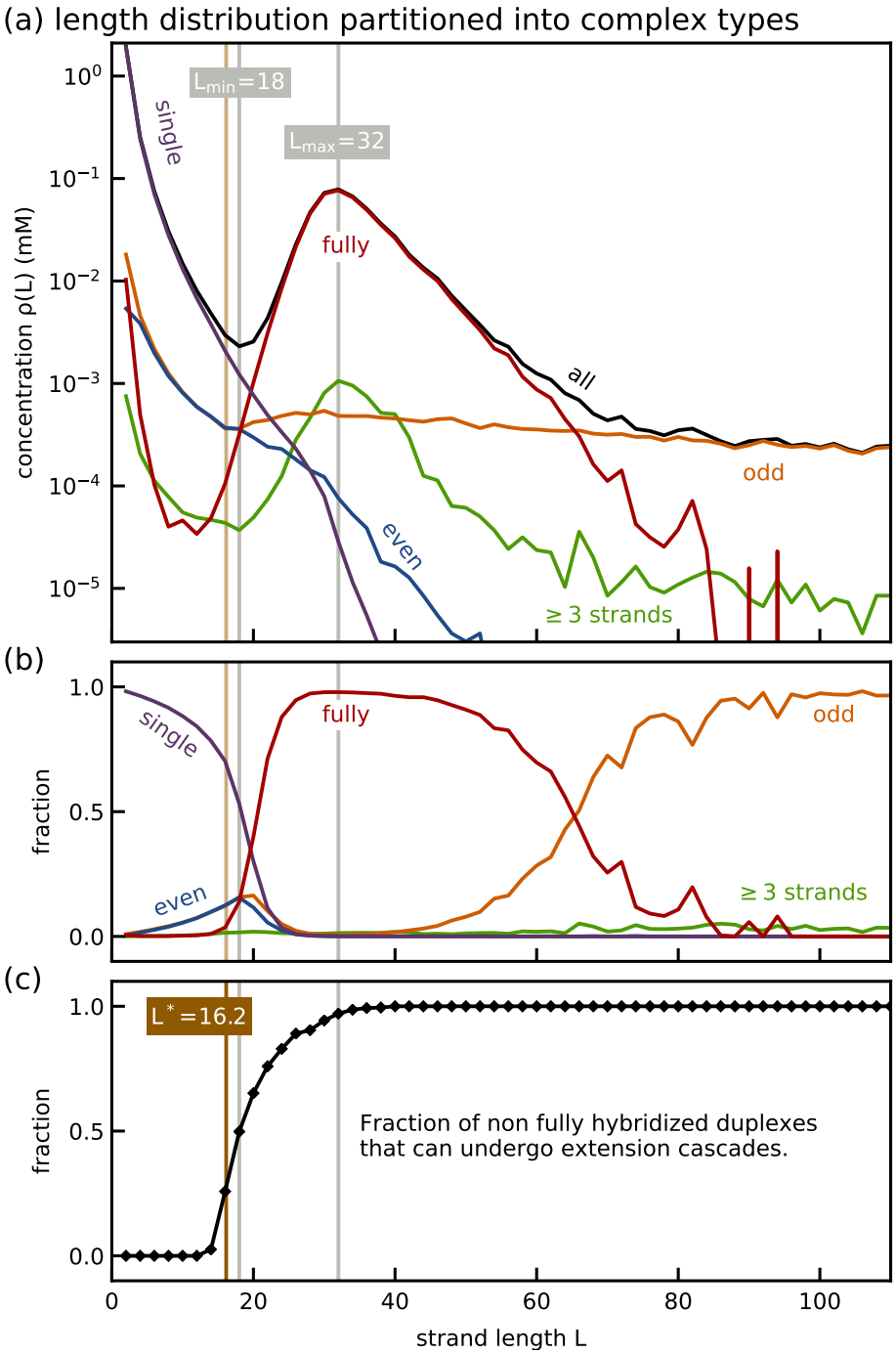


Figure 6.7.: Partitioning the contributions of the different subgroups (*cf.* Fig. 6.6) to the strand-length distribution reveals the dominant configurations: Short strands are mostly single-stranded. Strands with lengths around the peaks are in the persistent fully-hybridized zero-parity configuration. In the dimer-only model, all strands are of even length. Odd duplexes thus never reach a fully-hybridized state and cause the long tail of the distribution. (c) The probability of different complex types conditioned on strand length. (d) The probability that a duplex with non-zero parity is stable conditioned on strand length. Around $L = L^*$ (*cf.* Eq. 6.18) this probability increases rapidly.

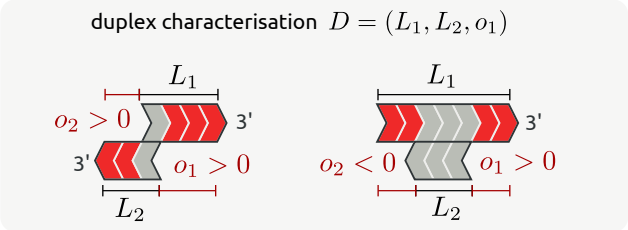


Figure 6.8.: Duplexes are uniquely characterized by the strand lengths $L_1, L_2 \in \mathbb{N}$ and the overhang $o_1 \in \mathbb{Z}$ of strand S_1 at its 3′ end. The overhang $o_2 \in \mathbb{Z}$ is defined analogously. Overhangs o_i can be negative, as for the case of o_2 in the right example. Only three of these numbers are independent since $0 = L_1 - L_2 - o_1 + o_2$.

Consequently, the rate of extension with a short m-mer is given by

$$r_{\text{ext}}(D) = \sum_m r_{\text{ext},m}(D). \tag{6.16}$$

The ratio of the extension rate r_{ext} and the dehybridization rate r_{off} gives the condition for the onset of extension cascades for the duplex D, $1 < r_{\text{ext}}(D)/r_{\text{off}}^{\text{dupl}}(D)$. As dimers are the most abundant species, we approximate $r_{\text{ext}}(D) < r_{\text{ext},2}(D)$ which yields a lower bound for the latter condition:

$$1 < \frac{r_{\text{ext},2}(D)}{r_{\text{off}}^{\text{dupl}}(D)}. \tag{6.17}$$

Up to now, we have focused on a specific duplex D. In order to obtain a more systemic view, we now consider a system containing only strand lengths smaller or equal to some fixed value L_0. We then determine the minimal L_0 such that duplexes appear, which can undergo extensions cascades.

Using Eq. (6.13) and Eq. (6.15) we write the ratio appearing on the right-hand side of Eq. (6.17) as

$$\frac{r_{\text{ext},2}(D)}{r_{\text{off}}^{\text{dupl}}(D)} = (L_1 + L_2 - 1)c_2 r_{\text{lig}} \sum_{\substack{i\in\{1,2\} \\ o_i \neq 0}} e^{-\gamma(l+z_i)}.$$

This ratio is largest for symmetric duplexes with $L_1 = L_2 = L_0$ where $l(D) + z_i = L_0$. The two duplex configurations maximizing the ratio are thus the odd duplex $D_{\pm 1} = (L_0, L_0, \pm 1)$ and the even duplex $D_{\pm 2} = (L_0, L_0, \pm 2)$. In our system, the former corresponds to a duplex that will undergo an infinite extension cascade, whereas the latter will reach a fully-hybridized configuration.

The smallest L_0, for which extension cascades become possible, defined as L^*, can then be found by solving

$$1 = 2(2L^* - 1)c_2 r_{\text{lig}} e^{-\gamma L^*}, \tag{6.18}$$

which yields $L^* \approx 16.2$. As the shortest building blocks are dimers, $L_\bullet^* = \lceil L^* \rceil$ is calculated by ceiling L^* to the next even integer, which yields $L_\bullet^* = 18$.

Notice that for strong binding, *i.e.*, $\gamma < -1$, the (kinetic) subexpontial length-dependence

which enters via the channel number Θ can be neglected. To leading order one then has

$$L^* \approx \ln\left(c_2 \frac{r_{\text{lig}}}{r_0}\right) \gamma^{-1}, \tag{6.19}$$

where we made the dependence of the microscopic kinetic parameter r_0 explicit.

As shown in Fig. 6.7, the distinct hump in the strand-length distribution is caused by persisting, fully-hybridized duplexes $(L, L, 0)$. We also know that these duplexes are the endpoints of extension cascades and that they will persist for long times, unless they eventually dehybridize or leave the system. This gives rise to two different fates of fully-hybridized duplexes depending on their length.

On the one hand, for L smaller than some critical value $L^\dagger$, $r_{\text{off}}^{\text{dupl}}(L, L, 0) > r_{\text{out}}$, such that their production in stationary state is for the most part balanced by dehybridization. On the other hand, for long duplexes with $L > L^\dagger$, we have $r_{\text{off}}^{\text{dupl}}(L, L, 0) < r_{\text{out}}$ and hence the stationary concentration is determined for the most part by a balance of their production with the outflux. Importantly, the outflux rate is independent of L, whereas the dehybridization rate decreases exponentially with L. We thus expect the existence of two different regimes where the stationary concentration of the fully-hybridized duplexes exhibit a different dependence on L.

Formally, we can find the length where the dehybridization-rate becomes smaller than the outflux-rate by

$$r_{\text{off}}(L^\dagger, \gamma) = \frac{e^{\gamma L^\dagger}}{2L^\dagger - 1} = r_{\text{out}}. \tag{6.20}$$

Solving this equation numerically for our standard parameters, we obtain $L^\dagger = 30.07$. Ceiling to the next even integer yields $L^\dagger_\blacktriangle = \lceil L^\dagger \rceil = 32$.

As above, for strong binding energies we can ignore the logarithmic kinetic dependence on the length and obtain

$$L^\dagger \approx \ln\left(\frac{r_{\text{out}}}{r_0}\right) \gamma^{-1}, \tag{6.21}$$

where again, we made the dependence of the collision rate explicit.

6.4.1.3. Understanding the shape of the strand-length distribution

We briefly summarize our main findings: Strands with $L < L^*$ typically dehybridize rapidly, such that they are most often single stranded. For $L > L^*$ duplexes dominate. The strand-length distribution exhibits a characteristic hump that is caused by fully-hybridized and thus persistent species, whose stationary concentration is either balanced by dehybridization or outflux. The strand length determining the transition between these regimes is estimated by the characteristic length $L^\dagger$.

From Fig. 6.7 (a) we already see that $L^\dagger$ coincides with the position of the maximum L_{max} in the strand-length distribution. On the two different sides of the maximum, the stationary balance equation for fully-hybridized duplexes has a different dependence on the strand length. Thus, the emergent length scale $L^\dagger$ separates two different dynamical regimes resulting in the peak at $L_{\text{max}} \sim L^\dagger$.

The other scale L^* serves as a proxy for the position of the minimum L_{min} in the strand-length distribution. While $L^\dagger$ formally marks the onset of extension cascades, *cf.* Fig. 6.7 (c). The minimum in the strand length distribution is caused by the consumption of single strands of lengths $L \gtrsim L^*$ at the start of extension cascades.

We will discuss the strong nonequilibrium dynamics of strands with lengths between L^* and $L^\dagger$ in more detail below.

6.4.2. Exploration of the parameter space

In order to show that our results are indeed generic, we performed an exhaustive screening of the parameter space of our model. As above, we empirically determine $L_{\max}$ and $L_{\min}$ from our simulations and compare them with the analytical expressions for L^* and $L^\dagger$.

The results are shown in Fig. 6.9. In each row, a single parameter is varied while all the other parameters are fixed at their standard values. The left column in Fig. 6.9 shows simulated stationary strand-length distributions. The right column presents the analytical expressions for the (ceiled) values of L^* and $L^\dagger$ with the characteristic lengths $L_{\min}$ and $L_{\max}$ of the simulated strand-length distribution. A colored curve in the left panel always corresponds to the accordingly colored marker in the right panel.

Fig. 6.9(a) shows the result for a variation of the outflux rate r_{out}. The transition from a short to a long-tailed length distribution was already discussed in Section 6.3.1. As the outflux rate should not influence the onset of extension cascades, we expect the position of the minimum to remain constant, which the simulation confirms. Increasing the outflux rate shifts $L_{\max}$ to lower lengths in an approximately logarithmic way in accordance with Eq. (6.21).

In Fig. 6.9(b) we vary the binding energy γ. We observe that increasing the binding energy displaces the characteristic hump towards shorter strands. The behavior of both curves is roughly inverse proportional $L \propto -\gamma^{-1}$.

Next, we vary the bare ligation rate r_{lig}, see Fig. 6.9(c). The position of the maximum remains unchanged, since the transition determining the fate of a fully-hybridized state is not affected by the ligation rate, see Eq. (6.20). In accordance with Eq. (6.18), decreasing r_{lig} logarithmically shifts the onset of the extension cascade and thus the position of the minimum to larger lengths. For the lowest ligation rate plotted, we cross the transition towards short-tailed distributions described in Section 6.3.1 and the characteristic hump in the length distribution disappears.

Fig. 6.9(d) shows the effect of varying the dimer concentration c_2. Since reducing c_2 logarithmically reduces the effective rate of extension with a dimer, higher concentrations enable extension cascades already for duplexes consisting of shorter strands, shifting the minimum to the left. Again, the position of the hump remains constant. For the smallest concentration shown we cross the transition towards a short-tailed distribution.

In summary, both the phenomenological position of the minimum $L_{\min}$ as well as the position of the hump, $L_{\max}$ are well described by the expressions for L^* and $L^\dagger$, Eqs. (6.18) and (6.20). As such, the characteristic features of the length-distribution can really be understood in the context of the onset of extension cascades that lead to persisting, fully-hybridized duplexes.

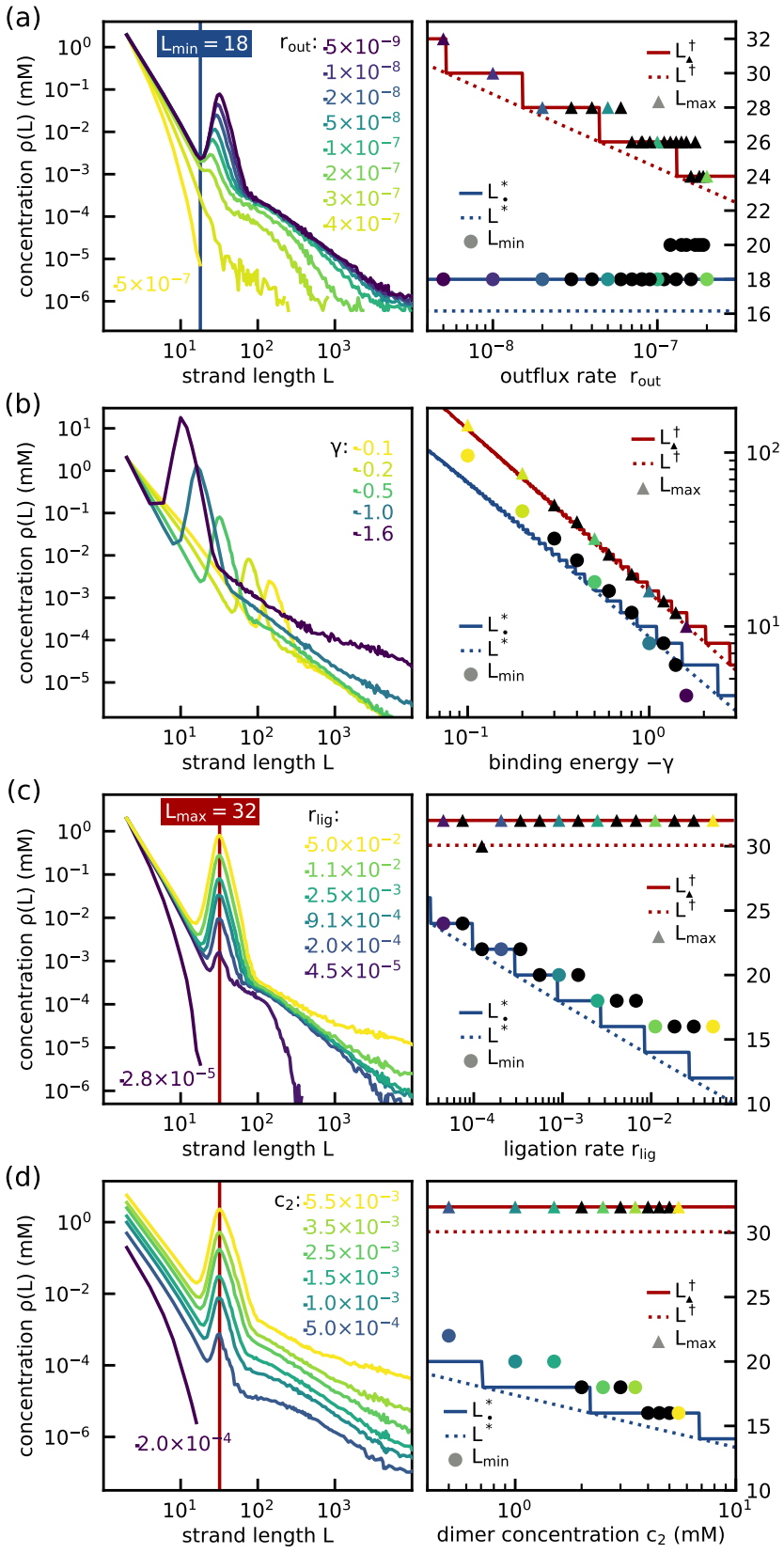


Figure 6.9.: Probing the parameter space of the dimer-only model. Left column: stationary strand-length distributions. Right column: Comparison of the observed values $L_{\min}$ and $L_{\max}$ and the predictions for L^* and $L^\dagger$ via Eqs. (6.18) and (6.20). Variable parameters are (a) the outflux rate r_{out}, (b) the dimensionless binding energy per nucleotide γ, (c) the bare ligation rate r_{lig} and (d) the concentration of chemostated single-stranded dimers c_2.

6.4.3. Sweep of the cutoff rate r_{cut}

In this section, we consider a variation in the cutoff rate r_{cut}. We separated it from the other parameter sweeps presented above as the associated transformation of the length distribution is more complex, see Fig. 6.11(left). We use the parametrization $r_{\text{cut}} = e^{\gamma l_{\text{cut}}}$. The minimum dehybridization rate is then given by $r_{\text{off}} = e^{\gamma l_{\text{cut}}}$ and thus, the maximal time scale that strands are hybridized onto each other is $t_{\text{cut}} = e^{-\gamma l_{\text{cut}}} = e^{0.5 l_{\text{cut}}}$.

6.4.3.1. Average strand length as a function of l_{cut}

Apart from the transformation of the length distribution, we want to analyze the average strand length in steady-state, $\langle \bar{L} \rangle$, where the bar notates the time average and the brackets notate the ensemble average.

We say that a steady state is reached if the time-dependent ensemble average of the strand length $\langle L \rangle (t)$ becomes constant. But as can be seen from Fig. 6.11(right) not for all values of l_{lig} a steady state was reached.

We therefore consider $\langle L \rangle (t_{\text{max}})$, where t_{max} is the maximum simulation time reached for the respective l_{cut}. The ensemble size used to calculate $\langle L \rangle (t_{\text{max}})$ varies, as not all runs belonging to a l_{cut} necessarily reached t_{max}. The value of $\langle L \rangle (t_{\text{max}})$ can be read of from Fig. 6.11(right), as it is given by the last point of each line, and are plotted as a function if l_{lig} in Fig. 6.10.

However we deduce from Fig. 6.11(right) that the ordering of $\langle L \rangle (t_{\text{max}})$ as a function of l_{cut} is the same as $\langle \bar{L} \rangle$. Or in other words, $\langle L \rangle (t_{\text{max}})$ is monotonously increasing/decreasing if and only if $\langle \bar{L} \rangle$ is monotonously increasing/decreasing, and therefore, the positions of the minima and maxima are the same. We will therefore, in the following discussion, treat both quantities equivalently.

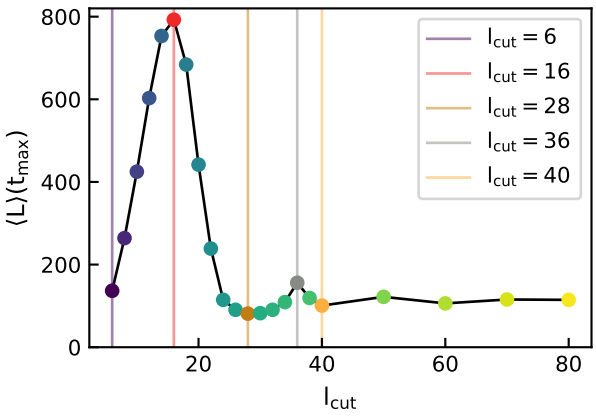


Figure 6.10.: $\langle L \rangle (t_{\text{max}})$ is the ensemble average length at the maximum simulation time that got reached by all runs belonging to a particular l_{lig} (see Figure 6.11,right). If a steady state was reached $\langle L \rangle$ is equivalent to $\langle \bar{L} \rangle$. As can be seen from Fig. 6.11(right), $\langle L \rangle (t_{\text{max}})$ and $\langle \bar{L} \rangle$ have the same ordering, hence the position of the maxima ($l_{\text{cut}} = 16, 40$) are the same in both cases.

Let us now follow the transformation of the length distribution when changing the cutoff l_{cut}. We start with a short cut off length $l_{\text{cut}} = 6$, which only allows for short-lived duplexes. Upon increasing l_{cut} to $l_{\text{cut}} = 16$ (red line) the length ´distribution extends to longer lengths at the cost of a decrease in concentration of strands of lengths $L \approx 10 - 1000$, which leads to an increase in the average length, see Fig. 6.10. A further increase of l_{cut} leads to a decline of the average length, thus the average length has a maximum at $l_{\text{cut}} = 16$, which is approximately the point where stable duplexes become possible. It marks the sweet spot between extension and dehybridization: A (almost) stable duplex gets formed and extended by an other strand. The strands dehybridize quickly thereafter such that the extended strand and the template are again available to catalyze further extensions.

Increasing l_{cut} from $l_{\text{cut}} = 16$ to $l_{\text{cut}} = 28$ leads to a decrease of the weight in the tail of the strand length distribution causing a decline in the average strand length $\langle \bar{L} \rangle$. In this range stable duplexes arise that can undergo extension cascades. The lifetime of fully-hybridized duplexes of length $L_{\max}$ is still to short compared to $1/r_{\text{out}}$ such that they do not become stalled in the fully-hybridized configuration. Hence, there is no accumulation of strands around $L_{\max}$. Further increasing their stability by changing l_{cut} from $l_{\text{cut}} = 28$ to $l_{\text{cut}} = 36$ leads to the formation of a saddle point. We conjecture that no minimum has formed yet, as strands longer then $L_{\max}$ are still destabilized due to the cutoff and hence are available as templates for the ligation of strands $L < L_{\max}$. Also $\langle L \rangle (t_{\max})$ increases again until the saddle point in the length distribution is formed at $l_{\text{cut}} = 36$. But the associated maximum in the mean length distribution is much smaller than the previous one at $l_{\text{cut}} = 16$.

Upon further increase, $l_{\text{cut}} = 36$ to $l_{\text{cut}} = 40$ the average strand length $\langle \bar{L} \rangle$ decreases while the characteristic maximum and minimum in the strand length distribution emerge.

For $l_{\text{cut}} \geq 50$, the average strand length and the whole shape of the length distribution become independent of l_{lig}. Hence for our system $l_{\text{cut}} \geq 50$ is equivalent to a system without cutoff $l_{\text{cut}} = \infty$.

6.4.3.2. Further system properties as a function of l_{cut}

We further analyzed the total number of complexes $\left\langle \overline{N_C^{\text{tot}}} \right\rangle$, the average complex mass $\langle \overline{m} \rangle$ and the total number of strands $\left\langle \overline{N_S^{\text{tot}}} \right\rangle$. As there was no steady-state reached for some l_{cut}, we proceed equivalent as for the average strand length and consider the average quantities at the latest sampled time point $t_{\max}$, see Fig. 6.12.

Fig. 6.12(left) shows the number of complexes $\left\langle \overline{N_C^{\text{tot}}} \right\rangle (t_{\max})$ which decreases with increasing l_{cut}. This behavior can be expected as strands can cluster into higher-order duplexes due to the higher possible stability of hybridizations. There is a local minimum of $\left\langle \overline{N_C^{\text{tot}}} \right\rangle (t_{\max})$ at $l_{\text{cut}} = 36$. The increase of $\left\langle \overline{N_C^{\text{tot}}} \right\rangle (t_{\max})$ for $l_{\text{cut}} = 40$ can be understood the following way: At this point $1/r_{\text{cut}}$ becomes larger than the simulated system time and hence the number of elongators that do not dehybridize before leaving the system increases. Hence these strands do not serve as a template for the ligation of short single strands, which would lead to a decrease in $\left\langle \overline{N_C^{\text{tot}}} \right\rangle (t_{\max})$.

From Fig. 6.12(middle) we can deduce that $\langle m \rangle$ has a maximum around $l_{\text{cut}} \approx 28$. Hence for this, value complexes are largest. The question is then if they are formed by the hybridization of a multitude of strands (higher-order complexes) or if they consist of few but long strands?

As the average strand length decreases for $l_{\text{cut}} > 16$, we conjecture that the increase of $\langle m \rangle$ is due to the formation of complexes of higher-order complexes, while the strand length decreases. The formation of higher-order complexes goes along with an increase in the total number of strands in the system, which also peaks around $l_{\text{cut}} \approx 28$.

We can further use $\langle N_C^{\text{tot}} \rangle$, $\langle m \rangle$ and $\langle L \rangle$ to calculate the energy and mass flux per unit volume through the system:

$$\Phi_M = \frac{\langle N_C^{\text{tot}} \rangle}{V} \langle m \rangle \, r_{\text{out}}, \tag{6.22}$$

$$\Phi_E = \frac{\langle N_C^{\text{tot}} \rangle}{V} \frac{\langle m \rangle}{\langle L \rangle} \frac{1}{2} \left(\langle L \rangle - 1 \right) r_{\text{out}} \tag{6.23}$$

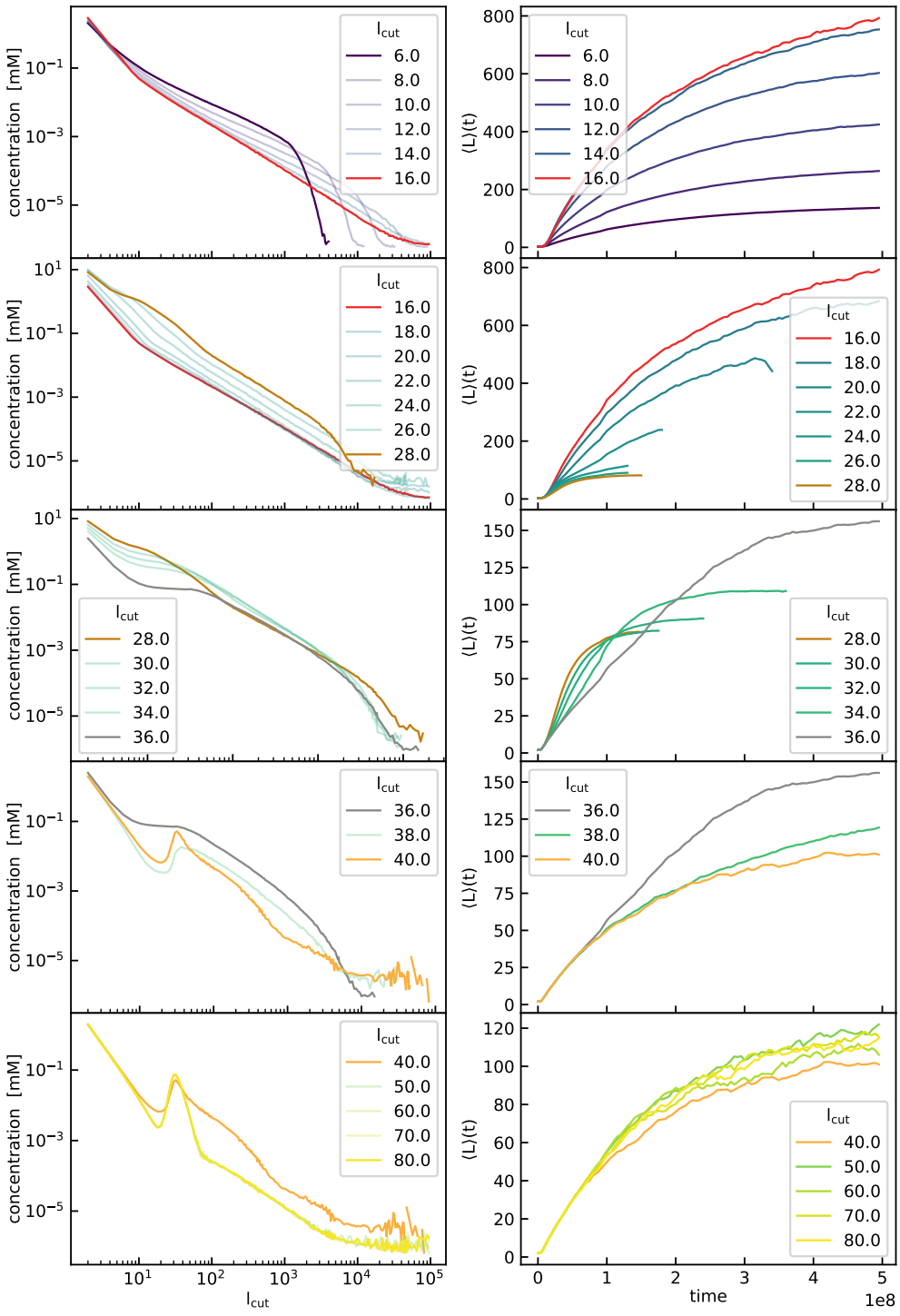


Figure 6.11.: Sweep $r_{\text{cut}} = e^{\gamma l_{\text{cut}}}$ (left) transformation of the strand length distribution by varying l_{cut} from $l_{\text{cut}} = 6$ to $l_{\text{cut}} = 80$. (right) average strand length vs l_{cut}. For some values of l_{cut} no steady state was reached (using a constant average strand length as the indicator for steady state).

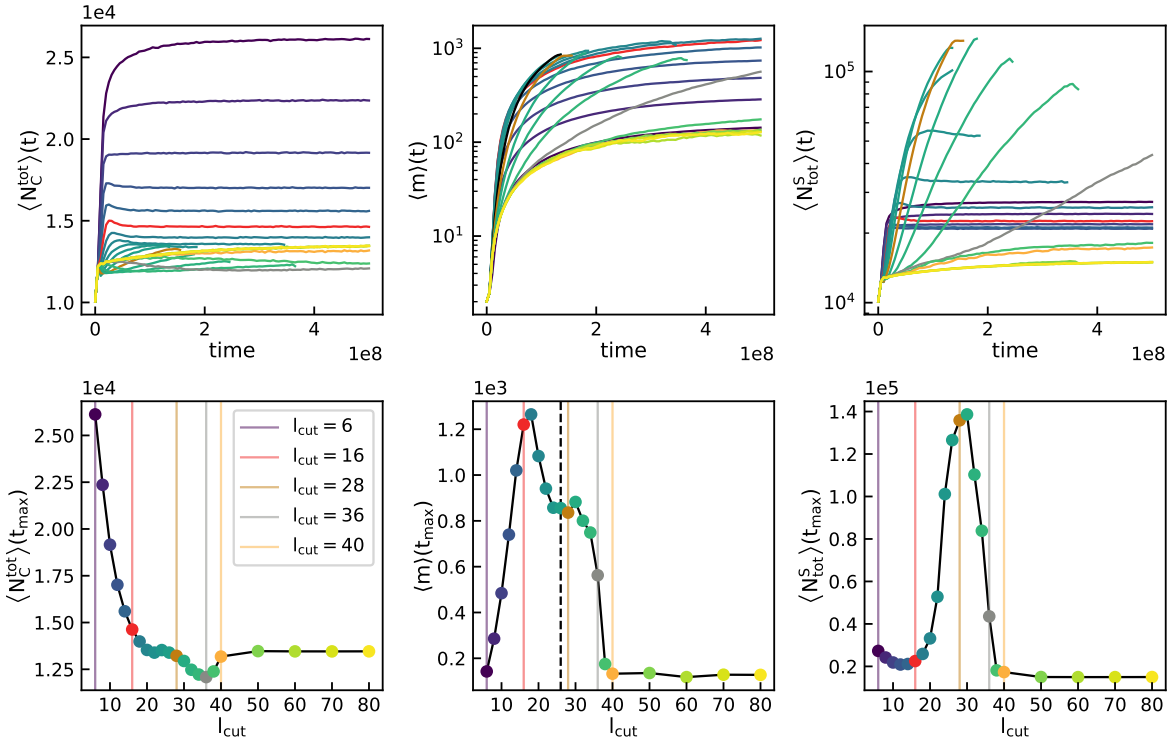


Figure 6.12.: (top) Temporal development of the total number of complexes $\langle N_C^{\text{tot}} \rangle$, the average mass of a complex $\langle m \rangle$ and the number of strands $\langle N_S^{\text{tot}} \rangle$. (bottom) The value of these quantities is evaluated at the latest simulation time t_{max} for each l_{lig}. The color code is the same as in Fig. 6.11 reaching from $l_{\text{cut}} = 6$ (dark purple colors) to $l_{\text{cut}} = 6$ (bright yellow colors). (middle) The line corresponding to $l_{\text{cut}} = 28$ was colored black as we (by eye) identify it as the most likely position of the maximum of $\langle m \rangle$ (t_{max}).

where we used that the order of the complex is approximately $\langle n \rangle \approx \frac{\langle m \rangle}{\langle L \rangle}$ and that $\frac{1}{2}(\langle L \rangle - 1)$ give the number of ligations necessary to build a strand of length $\langle L \rangle$ by ligation of dimers. The fluxes $\langle \Phi_M \rangle (t_{\max})$, $\langle \Phi_M \rangle t_{\max}$ are shown in Figure 6.13. The unbound model is apparently the model which requires the smallest influx of energy and mass in order to be operational. Those fluxes can, in principle, be used to estimate the required influx of building blocks through a membrane of a potential protocell in order to sustain the growth modes discussed in this work. Also, the impact of a limited supply of nucleotides, implemented via an influx proportional to the difference in concentration between the reactor and its surrounding, would be interesting to study.

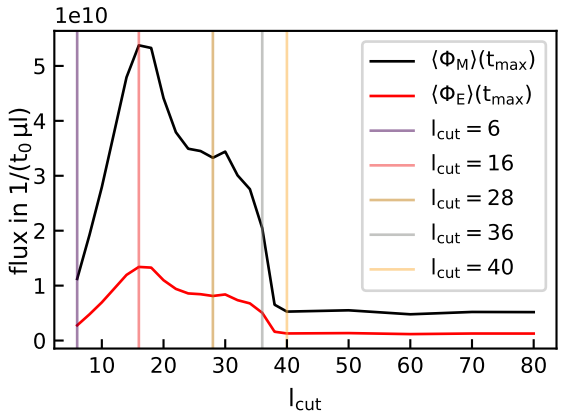


Figure 6.13.: The mass and energy flux per unit volume. Note that the peculiar shape of the curves is due to the system not having reached steady state. Never the less it is certain that the curves increase with l_{cut} for $l_{\text{cut}} \leqslant 16$, and decrease for $l_{\text{cut}} \geqslant 36$. The final distribution show presumably a single maximum.

6.4.3.3. Summary l_{cut} sweep

In summary, for $l_{\text{cut}} = 6$, we start in a state where the system has a maximum amount of complexes. However, these complexes are all short single strands. Upon an increase of l_{cut}, the average strand length increases until a sweet spot for the extension of strands is hit at $l_{\text{cut}} = 16$, while the number of complexes decreases. Further increase of l_{cut} leads again to shorter strands while the order of complexes increases until $l_{\text{cut}} = 28$. Thereafter the order of complexes decreases again until single strands and duplexes are the only relevant complexes, while the average strand length stays roughly constant, and the maximum at $L_{\max}$ emerges.

The behavior of the system due to a variation of l_{cut} can be highly relevant when conceptually thinking about the mode of operations of a hypothetical protocell. By variation of r_{cut}, the strand and complex properties can be radically changed. For example, the maximum in the mean strand length distribution produces long strands, which could potentially fold into functional entities (ribozymes). Then the increase in l_{cut} could lead to the formation of even more complex ribozymes consisting of several subunits. The further increase of l_{cut} could bring the system into a state that is well suitable for the copying of information.

6.4.4. Transient behavior in closed systems

We further investigated closed systems without in- or outflux. We prescribed the concentration of initial building blocks and let the system evolve transiently. Due to the irreversibility of the ligation reaction, closed systems are not ergodic: Short building blocks will deplete, and the final configuration contains only two very long strands. However, this stationary state will never be reached on practical time scales both in simulations and experiments.

Thus, instead of analyzing the stationary state of the system, we consider a transient state at intermediate times. More precisely, we focus on the situation where long strands have

already formed, and extension cascades are possible, but there is still a sufficient amount of short building blocks. Then, the system behaves similarly to the stationary situation in an open system with small outflux rates.

Considering the limiting behavior $r_{\text{out}} \to 0$ provides us with an intuition about the transient length distribution in this case. From Eqs. (6.20) or (6.21) we see that formally $L^\dagger \to \infty$ and the time to reach the stationary state also diverges.

Fig. 6.14(a) shows the strand-length distribution for the standard choice of parameters for various values of the transient observation time $t = \tau_{\text{obs}}$. As in the stationary case, we observe a distinct minimum and maximum in the strand-length distribution. Fig. 6.14(b) shows that the position of the maximum increases logarithmically with the observation time.

In order to get an intuition for this behavior, we again use an argument involving the competition of time scales. As in the open systems, strands longer than L^* will dominantly occur in fully-hybridized configurations. In contrast, the second time scale is not determined by a global outflux rate, and fully-hybridized duplexes eventually dehybridize with a length-dependent rate $r_{\text{off}}(L)$. Yet, dehybridization of duplexes of length L can only play a role for observation times longer than $\tau_{\text{obs}} \sim r_{\text{off}}(L)^{-1}$.

With respect to the typical feature of the strand-length distribution, we thus expect global transient observation time τ_{obs} to play the same role as the time scale r_{off}^{-1} in an open system. The length-scale $L = L^\dagger$ that determines the peak in a closed system can then be obtained by replacing r_{off} with τ_{obs}^{-1} in Eqs. (6.20) or (6.21). In that case, the position of the peak should increase logarithmically with time, consistent with the results shown in Fig 6.14(b).

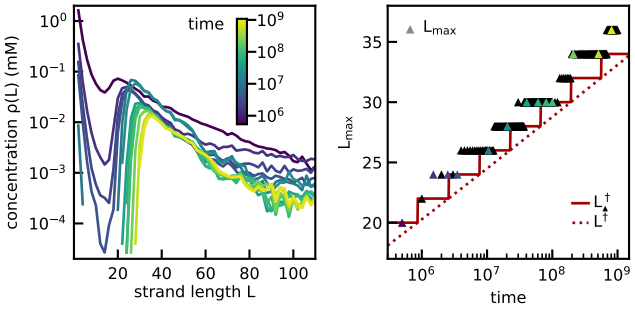


Figure 6.14.: Transient strand-length distributions: (left) Temporal development of the length distribution in a closed system. Over time the concentration of short strands decreases and the minimum develops into depleted region. (right) The position of the maximum L_{max} shifts logarithmically with time towards longer lengths.

6.4.5. Building block mixtures

So far, we have studied systems that used dimers only as their initial building blocks. While this made our calculations more convenient, only strands of even length appear in the system. This enabled "infinite" extension cascades starting from duplexes with odd parity and resulted in a heavy tailed strand-length distribution. In this section, we will consider monomer-dimer and dimer-trimer building block mixtures.

6.4.5.1. Monomer-dimer mixture of initial building blocks

Fig. 6.15 (a) shows the strand-length distribution for a reservoir where the total initial building block concentration $c_{\text{tot}} = c_1 + c_2 = 2\,\text{mM}$ is constant. We then vary the monomer fraction $f_{\text{m}} := \frac{c_1}{c_{\text{tot}}}$ from zero up to 90 %. The orange curve ($f_{\text{m}} = 0$) is the dimer-only system at standard parameters, showing the long tail caused by the infinite extension cascades. For any finite monomer concentration, infinite extension cascades are suppressed, and the long tail collapses. The partitioning of complexes into various substructures is

shown in Fig. 6.15 (b) for $f_m = 70\,\%$. The tail of the distribution is now dominated by fully-hybridized duplexes. As above, duplexes with finite overlap are distinguished by the parity of their overhangs, with the addition of mixed parity duplexes, having different parity at the different sites.

Overall, the length distributions for finite monomer fractions look qualitatively similar. In the stationary state, the total mass in the system depends on the balance between in and outflux: The higher the monomer fraction, the less nucleotide mass is added by the influx, and the overall strand-length distribution shows lower concentrations.

Moreover, the lower the monomer concentration, the more of the bias towards strands with even lengths is retained. This bias is most visible for short strands, leading to the zig-zag pattern visible in Fig. 6.15 (a). For long strands, the bias vanishes.

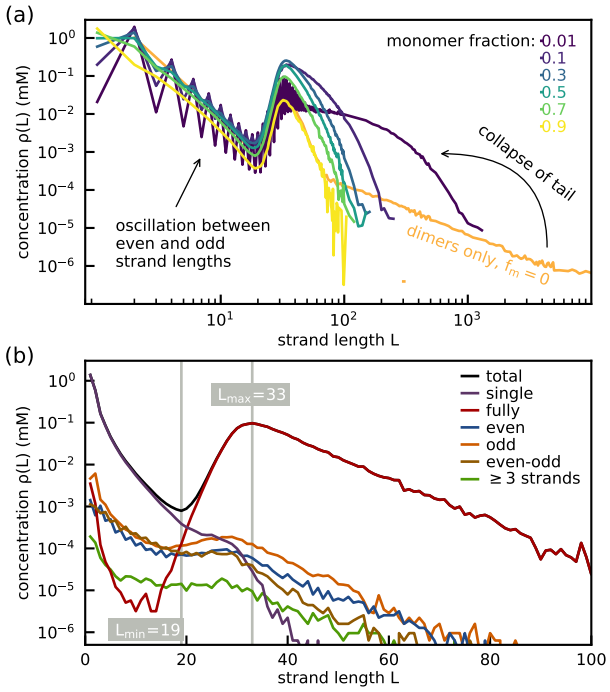


Figure 6.15.: Strand-length distributions for monomer-dimer mixtures. The monomer fraction f_m is varied between zero and 90 % at a total concentration $c_{tot} = 2\,\text{mM}$. (a) Steady state length distributions for different f_m. For low f_m the concentration between even and odd strands oscillates heavily for short strands. The long tail that is present for $f_m = 0$ (orange curve, only even strand lengths shown) collapses even for very small f_m. (b) Partitioned strand-length distribution for $f_m = 70\,\%$. In contrast to Fig. 6.7, virtually all strands with $L > L^*$ belong to a fully-hybridized duplex.

Importantly, the general understanding of the characteristic features of the strand-length distribution presented above remains valid. In accordance with Eq. (6.20), the position of the maximum is unchanged, as it does not depend on the (constant) building block concentration. The position of the minimum also remains mostly constant.

Repeating the calculations leading to Eq. (6.18) for the onset of extension cascades, with the combined extension rate for both monomers and dimers leads to the same equation, with the dimer concentration c_2 replaced by the total concentration c_{tot}:

In general, for reservoirs with mixed building blocks, the typical length L^* for the onset of extension cascades is derived analogously to the dimer-only model using the condition $1 < r_{ext}(D)/r_{off}^{dupl}(D)$, *cf.* Sec. 6.4.1.2. Instead of only considering the extension with a dimer building block, one needs to include the extension with a monomer. The extension rate thus becomes

$$\begin{aligned} r_{ext}(D) &\approx r_{ext,1} + r_{ext,2} \\ &= r_{lig} \sum_{\substack{i\in\{1,2\} \\ o_i\neq 0}} \left[c_2 e^{-\gamma(\min(|o_i|,2))} + c_1 e^{-\gamma(\min(|o_i|,1))} \right]. \end{aligned} \tag{6.24}$$

The criterion for extension cascades then reads

$$1 \leq (L_1 + L_2 - 1) r_{\text{lig}} \times \sum_{\substack{i \in \{1,2\} \\ o_i \neq 0}} \left[c_2 e^{-\gamma(l + \min(|o_i|,2))} + c_1 e^{-\gamma(l+\min(|o_i|,1))} \right]. \tag{6.25}$$

The right side of the Eq. (6.25) is maximal for the odd duplex configuration $D_{\pm 1} = (L_0, L_0, \pm 1)$, for which $l + \min(|o_i|, 2) = l + \min(|o_i|, 1) = L_0$, which leads to

$$1 \leq 2(2L_0 - 1) r_{\text{lig}} (c_2 + c_1) e^{-\gamma L_0}. \tag{6.26}$$

Consequently, L^* for monomer-dimer mixtures obeys

$$1 = 2(2L^* - 1) r_{\text{lig}} c_{\text{tot}} e^{-\gamma L^*}, \tag{6.27}$$

where $c_{\text{tot}} = c_2 + c_2$ is the total concentration of building blocks.

Like already mentioned above, Equation (6.27) is the same formula as for the dimer only-system, except that the dimer concentration c_2 is substituted by the total concentration of building blocks c_{tot}. In accordance with formula Eq. (6.27), we observe that the position of the minimum is constant $L_{\text{min}} = 19$ under variation of the monomer fraction while keeping the total concentration fixed at $c_{\text{tot}} = 2\,\text{mM}$ (see Fig. 6.16). Only the dimer-only system has a different minimum position $L_{\text{min}} = 18$.

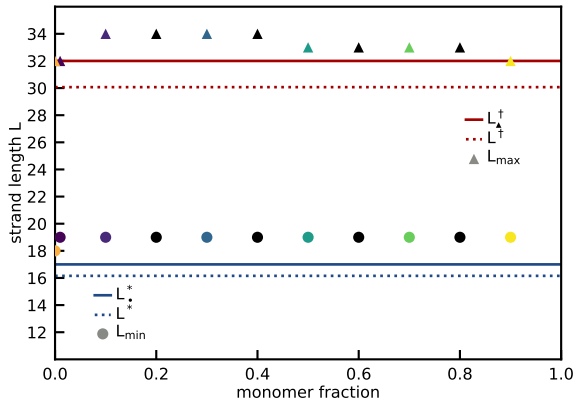


Figure 6.16.: L^* in the monomer dimer system is calculated via Eq. (6.27), which is the same as the formula for the dimer only system upon substituting the dimer concentration with the total concentration c_{tot}.

6.4.6. Dimer-trimer mixtures of initial building blocks

Fig. 6.17 (a) shows the strand-length distribution for a reservoir containing dimers and trimers, at a fixed total initial building block concentration $c_{\text{tot}} = c_2 + c_3 = 2\,\text{mM}$. Equivalent as we did for monomer mixtures, we then vary the trimer fraction $f_t := \frac{c_3}{c_{2\text{tot}}}$.

The orange curve ($f_t = 0$) is again the dimer-only system at standard parameters. The system shows the expected behavior equivalent to the monomer-dimer system: Adding a few monomers is enough to cause a collapse of the long tail as all duplexes undergoing extension cascaded become fully hybridized at some point. The position of the maximum and the minimum L_{min}, L_{max} are roughly constant. The tail of the length distribution seems to become gradually bend toward smaller lengths upon an increase of the trimer fraction from $f_t = 0$ to $f_t = 0.5$. Simultaneously the length distribution of the small strands seems to become shifted in the log-log representation, which indicates a change in the prefactor governing the power-law distribution of small strands. But in the limit $f_t \to 1$ the length

distribution does apparently not seems to follow this monotonous transformation as for $f_t = 1$, trimers only, the concentrations of short strand lengths are largest (yellow curve). Remarkably, the trimer-only length distribution expresses a plateau that spans roughly one order of magnitude $\approx 100 - 1000$.

The partitioning of complexes of the trimers only system into various substructures (see Fig. 6.17 (b)) reveals that the strands of the tail and hence the plateau are not contained in duplexes, but in complexes of higher-order $n \geq 3$.

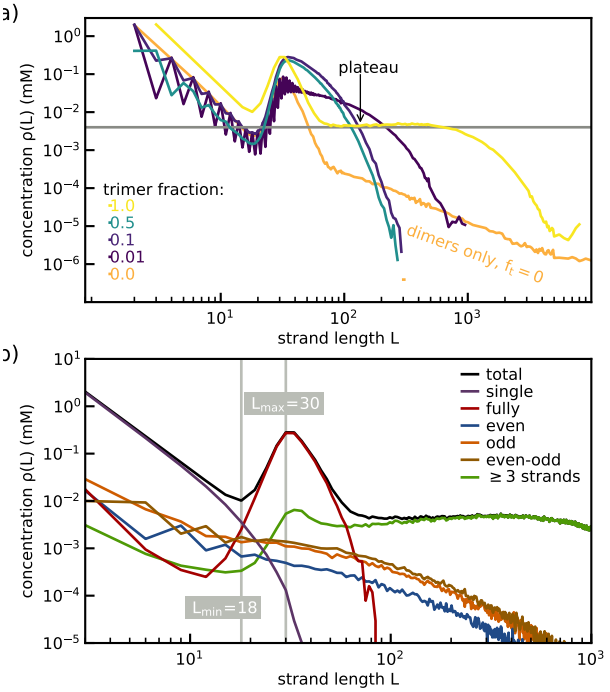


Figure 6.17.: Strand-length distributions for dimer-trimer mixtures. The trimer fraction f_t is varied between zero and 1 at a total concentration $c_{tot} = 2\,\text{mM}$. (a) Steady state length distributions for different f_t. The behavior is, except for the trimer only system $f_t = 1$, analogous to the behavior of the monomer-dimer mixture. The trimer-only system expresses a plateau where the concentration seems to be independent of the length over roughly one order of magnitude (100-1000). (b) Partitioned strand-length distribution of the trimer only system $f_t = 1$. In contrast to the other systems studied, the strands of the tails are not part of a duplex, but higher-order complexes $n \geq 3$.

6.4.6.1. Trimer only system plateau in the length distribution

In order to obtain an insight into the processes of the trimer only system that causes the plateau, we start with considering the complex length distribution $\psi(C)$, Fig. 6.18.

$\psi(C)$ expresses a pattern, for $i \in \mathbb{N}$ we have $\psi(3i) < \psi(3i+1) < \psi(3i+2)$. The pattern can be understood by considering the simplified reaction network, see Figure 6.19 (a) where we neglected the directionality of the complexes. Within this network, an even-odd complex of length $3i$ becomes by extension with a trimer either an odd complex of length $3i+1$ or an even complex of length $3i+2$. The extension $3i \to 3i+1$ involves an overlap of 2, whereas $3i \to 3i+2$ includes an overlap of 1, which favors the formation of complexes of length $3i+1$.´ For small lengths the ratio is accordingly given by $e^{-\gamma}$, see Fig. 6.18 (b).

Stable duplexes with an overhang mod $(o_i.3) = 0$ will become fully-hybridized at some point, whereas stable duplexes with mod $(o_i.3) = 1$ and mod $(o_i.3) = 2$ will continue growing thereby switching from an mod $(o_i.3) = 1$ to an mod $(o_i.3) = 2$ overhang and vice versa. These elongators cannot grow by interaction with other duplexes (duplex-duplex extension) with an overhang of length $o_i = 1$ and $o_i = 2$ and subsequent ligation as shown in Figure 6.19 (b) resulting in even longer elongators with gaps.

We can speculate that this mechanism causes the plateau in the strand length distribution: The average number of strands $\langle n \rangle$ forming the long complexes grows linearly $\langle n \rangle = 0.0023C + 2.2$, *cf.* Fig. 6.18 (c). For very long complexes we can assume that $\langle n \rangle \approx 0.0023C$. Let us further assume that the length of a complex is the number of strands times a typical

length scale L_p, $C = \langle n \rangle L_p$, which yields $L_p = 1/0.0023 \approx 435$, which coincides roughly with the center of the plateau in the strand length distribution. This mechanism is apparently weak in the dimer only system as in that case, the extension of an elongators with the most abundant building block (dimers) leads to a new overhang of length one. Two elongators with overhang one hybridizing onto each other lead to two ligation sites but no overhang, see Figure 6.19 (c). Two subsequent ligations result in a duplex, not resulting in the emergence of a plateau in the strand length scale. In a system containing monomers the plateau can not be formed neither as all duplexes become fully-hybridized at some point and the tail decays quickly.

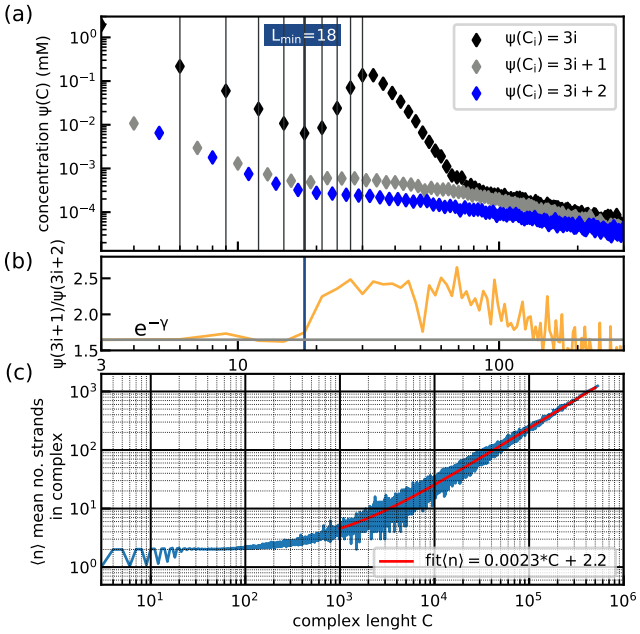


Figure 6.18.: Analysis of complexes of the trimer only system. (a) The complex length distribution $\psi(C)$: Complexes of lengths $C_i = 3i$, $i < 0$ being an integer, resemble die minimum and maximum also seen in the strand length distribution. Note that those complexes must not be in a fully hybridized configuration. The length distribution shows a pattern $\psi(3i) < \psi(3i+1) < \psi(3i+1)$ for $i \in \mathbb{N}$. (b) Ratio $\psi(3i+1)/psi(3i+2)$: For $C < L_{\min}$ the ratio is approximately equal to $e^{-\gamma}$. (c) Mean number of strands of vs length of complexes.

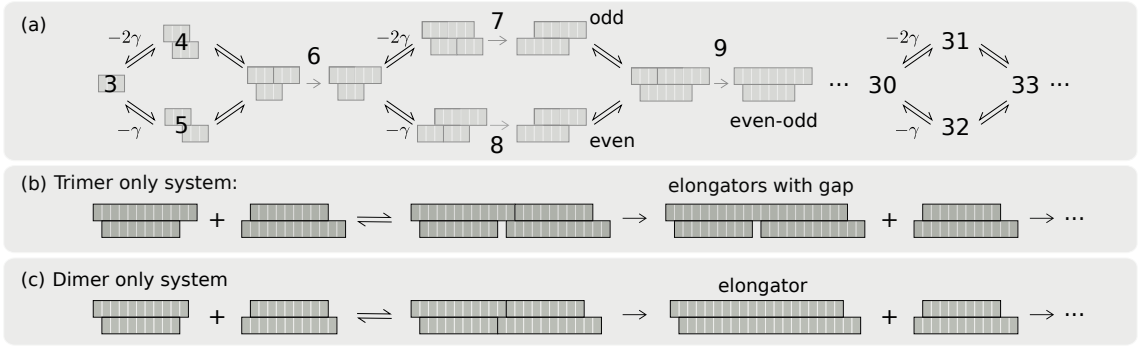


Figure 6.19.: (a) Simplified reaction scheme of trimer dynamics. The formation of a complex of length $C_o(i) = 3i + 1$ from an even-odd complex of length $C_{\text{eo}} = 3i$ is approximately proportional to $\sim e^{-2\gamma}$, whereas the formation of an even complexes of length $C_e(i) = 3i + 2$ is proportional to $\sim e^{-\gamma}$. (b) Duplex-duplex extension in a trimer only system can lead to elongators with gaps of length 1. (c) Duplex-duplex extension in a dimer only system.

6.4.7. Growth of complexes

In the previous section, we have established how the length-scales L^* and $L^\dagger$ shape the strand-length distribution. We will now consider how they determine the properties of the extension cascades in more detail.

The first length scale L^* sets the typical start of extension cascades and separates strands into two regimes. Strands that are shorter than L^* form unstable duplexes. While some ligations occur for those strands, they are relatively rare, and the dynamics is thus only weakly out of equilibrium. In contrast, strands longer than L^* form stable duplexes that usually undergo extension events. This dynamics is strongly out of equilibrium and gives rise to the nonmonotonous distribution.

Even though we know that the second length scale $L^\dagger$ relates the dehybridization time to the outflux (or transient) time scale, its role in the dynamics is not as straightforward. In the following, we will show that $L^\dagger$ is a typical scale where the self-enhancing processes that leads to the growth of strands and complexes breaks down.

6.4.7.1. Configurations of stable duplexes

The basis of our analysis are the statistics of individual trajectories of stable duplexes until they reach a fully-hybridized configuration and finally leave the system. An initial stable duplex consists of a long and a short strand of size L_{long} and $L_{\text{short}} \leq L_{\text{long}}$ with an initial overlap l_{initial}. The length of this initial duplex is $C_{\text{initial}} = L_{\text{long}} + L_{\text{short}} - l_{\text{initial}}$. Notice that different combinations of L_{long} and L_{short} correspond to the same values of C_{initial} and l_{initial}, *cf.* Fig. 6.20 (a).

These stable duplexes then grow by multiple extension events and become a fully-hybridized duplex of length $C_{\text{final}} \geq C_{\text{initial}}$. If $C_{\text{final}} = C_{\text{initial}}$, we say the trajectory grew via *pure primer extension*. In contrast, if $C_{\text{final}} \geq C_{\text{initial}}$, processes must have occurred that extended the length of the complex.

Our sampling algorithm is consistent with the actual (rate of) events that occur in a steady state and are explained in detail in Section A.5 of the Supplemental Material. In particular, the distribution $p(C_{\text{final}})$ characterizing the final complex length is proportional to their stationary concentration. For a concrete example, we sample trajectories from the stationary state of the system shown in Fig. 6.15, where monomers and dimers are kept at a total concentration $c_{\text{tot}} = 2\text{mM}$ with a monomer fraction of $f_{\text{m}} = 70\,\%$. First, we are interested in the configurations of initial stable duplexes, *i.e.*, in the statistics of l_{initial}, C_{initial}, L_{long} and L_{short}.

Fig. 6.20 (b) shows the joint distribution $p(C_{\text{initial}}, l_{\text{initial}})$. The probability is maximal for $C_{\text{initial}} \sim L^\dagger$ and $l_{\text{initial}} \sim L^*$. Since the probability decays fast away from that maximum, we say that these values define a *typical initial configuration*.

The next question regards the individual strands that form a typical initial duplex. Fig. 6.20 (c) shows the joint distribution $p(L_{\text{long}}, L_{\text{short}})$. We see that it is dominant in the lower triangle defined by $L^* \leq L_{\text{short}} \leq L_{\text{long}} \leq L^\dagger$. Within the triangle, the distribution is approximately uniform.

Knowing the properties of the initial configuration, we now turn to the final configuration. Fig. 6.20 (d) shows the joint distribution $p(C_{\text{final}}, C_{\text{initial}})$. A considerable part of the weight (around 17 %) of the distribution is on the diagonal $C_{\text{initial}} = C_{\text{final}}$, corresponding to the pure primer extension. The maximum weight (around 2.5 %) is located at $C_{\text{final}} = C_{\text{initial}} = 31 \sim L^\dagger$. We further see that off-diagonal elements are centered around this maximum. While we cannot neglect the off-diagonal elements accounting for complex extension, it is convenient to consider the pure primer-extension scenario first.

6.4.7.2. Catalytic growth processes and reassembly

Consider a stable duplex growing by pure primer extension. The generic case is shown in Fig. 6.21 (a). After a fully-hybridized duplex is reached, it will eventually dehybridize or

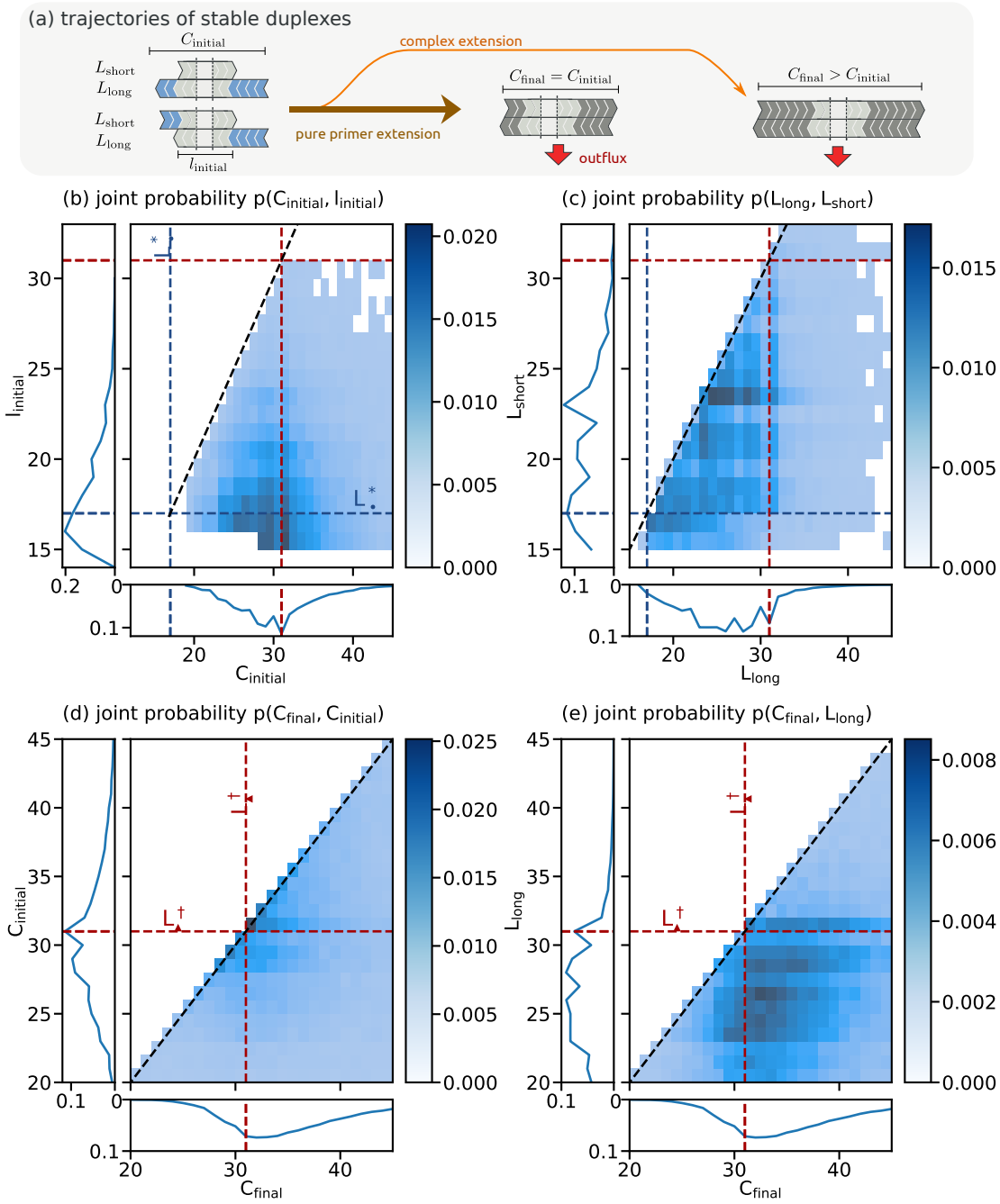


Figure 6.20.: (a) Sampled trajectories start with an initial stable duplex characterized by its strand lengths L_{long} and L_{short} together with the initial overlap l_{initial} and complex length C_{initial}. Trajectory statistics can be understood from various joint probability distributions, with the corresponding marginal histograms plotted on the axis. Horizontal and vertical dashed lines indicate the typical scales $L^*_\bullet = 17$ (blue) and $L^\dagger_\blacktriangle = 31$ (red). For the arguments made in this section, we do not distinguish between the float and ceiled values, $L^*_\bullet$ L^* and $L^\dagger_\blacktriangle$ $L^\dagger$. The black dashed line is the diagonal, where abscissa and ordinate are equal. (b) Typical initial stable configurations have $C_{\text{initial}} \sim L^\dagger$ and $l_{\text{initial}} \sim L^*$. (c) Strand combinations $(L_{\text{long}}, L_{\text{short}})$ are almost uniformly distributed in the triangle defined by $L^\dagger \geq L_{\text{long}} \geq L_{\text{short}} \geq L^*$. (d) About $\sim 17\,\%$ of trajectories grow by pure primer extension (diagonal $C_{\text{final}} = C_{\text{initial}}$) with no complex extension. (e) The joint probability $p(L_{\text{long}}, C_{\text{final}})$. The weight on the diagonal $L_{\text{long}} = C_{\text{final}}$ ($\sim 2.5\,\%$) indicates autocatalysis.

leave the system. If it dehybridizes, it may hybridize to another single-strand and thus create a stable duplex with a fresh overhang. This *reassembly* of strands is the mechanism by which long strands catalyze the formation of other long strands in the strongly non-equilibrium regime.

The reassembly probability p_{ra} is mostly determined by competition between outflux and dehybridization, resulting in a sigmoidal dependence on the length C_{final}:

$$p_{\mathrm{ra}} \sim \frac{r_{\mathrm{off}}}{r_{\mathrm{off}} + r_{\mathrm{out}}} \sim \left(1 + e^{\gamma(L^\dagger - C_{\mathrm{final}})}\right)^{-1}. \tag{6.28}$$

For complexes that have reached $C_{\mathrm{final}} \sim L^\dagger$, the reassembly probability decays exponentially to zero and the process stops. Thus the production rate of strand lengths reduces drastically above $L^\dagger$.

As long as $L^* \lesssim C_{\mathrm{final}} \lesssim L^\dagger$, longer fully-hybridized duplexes accumulate more strongly. Notice that while the concentration of single strands is strictly decaying in that regime, the concentration of fully-hybridized complexes of length C_{final} increases. This fact emphasizes the strong nonequilibrium character of the dynamics in this region, *cf.* Fig. 6.15 (b).

Enhanced accumulation and the frustration of the heterocatalytic reassembly process are thus the dynamic processes governing the emergence of the nonmonotonous strand length distribution. While this process is dominant, also autocatalysis occurs if $L_{\mathrm{long}} = C_{\mathrm{initial}} = C_{\mathrm{final}}$ and hence $L_{\mathrm{short}} = l_{\mathrm{initial}}$ (Fig. 6.21 (b)). This process is particularly strong for the typical configuration with $L_{\mathrm{long}} = L^\dagger$ and $L_{\mathrm{short}} = L^*$, see Fig. 6.20 (e): The diagonal $C_{\mathrm{final}} = L_{\mathrm{long}}$ (black dashed line) represents the autocatalytic cycles shown in Fig. 6.21(b), which is maximal around $L_{\mathrm{long}} \sim L^\dagger$. In the big picture, however, autocatalysis trajectories only represent about 2.5 % of all trajectories.

The longer strand in the initial stable duplex with lengths L_{long} are provided by the fully-hybridized double strands that are still more likely to dehybridize than to leave the system. In contrast, the shorter strand with length L_{short} are provided by the assembly processes in the weakly non-equilibrium regime. Importantly, for this typical cycle, the short strand is typically not one that is released in the dehybridization of final duplexes, making the process truly autocatalytic.

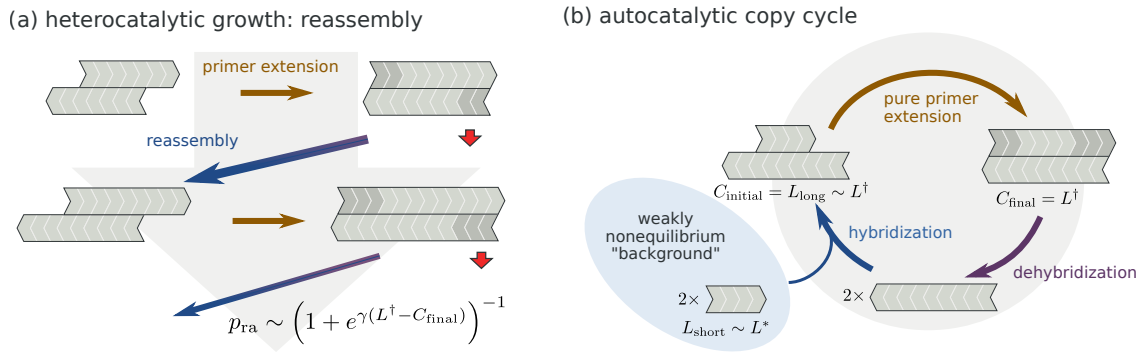


Figure 6.21.: Hetero- (a) and autocatalytic (b) processes for the growth of strands. In the strongly non-equilibrium regime, extension cascades cover the available overhang of stable duplex and form longer fully-hybridized strands. These long strands can then dehybridize and reassemble, thus creating new overhangs (copy sites) to be covered by extension cascades. The reassembly probability p_{ra} is determined by the balance between dehybridization and outflux and decays to zero fast for $L \gtrsim L^\dagger$.

6.4.7.3. Beyond pure primer extension

So far, we have pretended that only pure primer extension occurs when strands grow, *i.e.*, that no new overhang is created during the trajectory, see Fig. 6.22 (a). While pure primer extension only accounts for about $\sim 17\,\%$ of all trajectories, the discussion of the catalytic extension-reassembly dynamics does not rely on this fact. Still, in order to account for the remaining trajectories, we also discuss the processes that lead to complex extension.

First, notice that the growth happens essentially independently at each end of a duplex. It thus makes sense to take the perspective of a single end since it allows us to distinguish the two strands by their roles: We call the strand whose end is overhanging the template, whereas the other strand is called the primer. Moreover, we refer to the length of the overhang at the start of a trajectory as its initial copy site length l_{cs}.

The obvious mechanism that leads to complex extension is depicted in in Fig. 6.22 (b). It occurs when the original primer is extended with a strand that is longer than the (remaining) length of the copy site. After this extension, the roles of primer and template are reversed and a new copy site is created. We thus denote this process as *primer-template switching*. Notice that unstopped primer-template switching was responsible for the infinite extension cascades encountered in the dimers-only system.

Secondly, a complex undergoing an extension cascade is not always a simple duplex. Ligation reactions can also occur away from the stable hybridization site. We say that *template extension* occurs, if another strand facilitates the extension of the template strand, see Fig. 6.22 (c). From the perspective of the stable hybridization site, the length of its associated copy site l_{cs} has increased.

Fig. 6.22 (d) shows the number of extension events along a trajectory conditioned on the initially single-stranded length that is covered during the trajectory, $C_{\mathrm{final}} - l_{\mathrm{initial}}$. Standard primer-extension steps (p, red curve) are most common. In contrast, template extension (t, blue curve) is rare. For large $C_{\mathrm{final}} - l_{\mathrm{initial}}$, the number of events behaves strictly linear and primer-template switching (s, black curve) occurs approximately three times less than primer extension. For small values of $C_{\mathrm{final}} - l_{\mathrm{initial}}$, the relative fraction of primer-template switching increases, since a short available overhang increases the chance of primer-template switching.

6.4.7.4. Partial trajectories and copy site distribution

Each primer-template switching event along a trajectory creates a new distinct copy site, Fig. 6.22 (b). Consequently, trajectories naturally split into partial trajectories, which are defined by their copy site length l_{cs} and the sequence $(B_i)_{i\in[1,\dots,k]}$ of the building blocks used.

Fig. 6.22 (e) shows the distribution $p(l_{\mathrm{cs}})$ of copy site lengths in a double-logarithmic plot. Short copy sites are created by primer-template switching events and dominate the distribution. Longer copy sites, in contrast, are overwhelmingly created when single strands reassemble to form the initial complex at the start of trajectories. For the typical configuration $C_{\mathrm{initial}} \sim L^{\dagger}$ and $l_{\mathrm{initial}} \sim L^{*}$, the typical scale of the initial copy site is $L^{\dagger} - L^{*}$. The influence of these two regimes can be seen in the double-logarithmic representation. Different flanks of the distribution suggest a power-law decay with an exponent α_1 up to $l_{\mathrm{cs}} \approx L^{\dagger} - L^{*} = 14$ and a second exponent $\alpha_2 < \alpha_1$ beyond that. The faster decay with exponent α_2 reflects the circumstances that for a copy site of length $l_{\mathrm{cs}} > L^{\dagger} - L^{*}$, the formation of a initial stable duplex with $L_{\mathrm{long}} > L^{\dagger}$ is needed. As those duplexes are dominantly stalled in a fully hybridized configuration, the required initial stable duplexes are not formed.

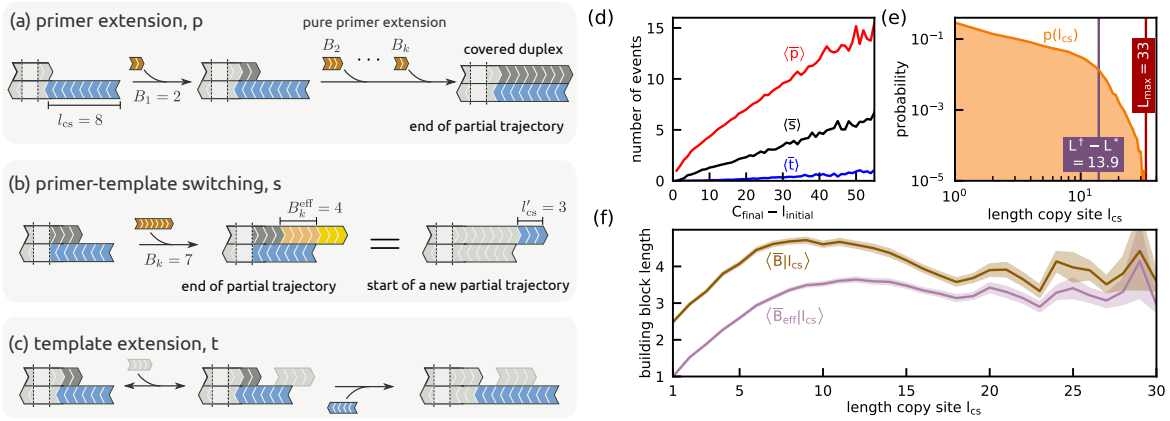


Figure 6.22.: Pure primer extension (a) and complex extension (b,c). The overhang at the beginning of a (partial) trajectory is called a copy site (blue) with length l_{cs}. (b) In primer-template switching events a building block extends the primer beyond the original copy site. The original copy site is fully covered and a new copy site is established. The roles of primer and template have changed. (c) Copy sites can grow independently of the original primer by template extension with the help of a helper strand. Right: (d) Number of extension events occurring during the covering of the total copy site $C_{\mathrm{final}} - l_{\mathrm{initial}}$. (e) Distribution of copy site lengths of partial trajectories. (f) Mean building block length conditioned on copy site length.

6.4.7.5. Building blocks and the influence of the weakly nonequilibrium regime

From the perspective of prebiotic self-assembly and the emergence of structural motifs, we also wonder about the buildings blocks used for extension. In contrast to controlled experiments, where building block sizes are prescribed, in our model the distribution of short strands arises from the weakly nonequilibrium regime. Since the weakly nonequilibrium regime is not the scope of this paper, the remainder of this section is mostly descriptive.

A first estimate of the typical length covered in an extension can be obtained from Fig. 6.22 (d). For sufficiently long copy sites, the combined slope of primer extension and template-switching has a value of about $\frac{1}{3}$. Since this slope is the average number of extension events needed to cover a single nucleotide, it implies that each extension event covers about 3 units of copy site length. In order to obtain a more precise picture, we consider the mean building block length for each copy site of a partial trajectory,

$$\bar{B} := \frac{1}{k}\sum_{i=1}^{k} B_i. \tag{6.29}$$

Fig. 6.22 (f) shows the conditional averages $\langle \bar{B}|l_{\mathrm{cs}}\rangle$ (brown curve). For completeness, we also show the mean effective building block size $\langle \bar{B}^{\mathrm{eff}}|l_{\mathrm{cs}}\rangle$ (purple curve), where the final building block is only counted according to its overlap. Both curves initially grow with l_{cs} and reach a plateau for large l_{cs}, which is close to the value determined from the slopes in Fig. 6.22 (e).

For short copy sites, covering occurs in a single step. Since the extension rate is mostly determined by the available copy site length l_{cs}, *i.e.*, the overlap, we initially have the proportionality $\langle \bar{B}|l_{\mathrm{cs}}\rangle \propto l_{\mathrm{cs}}$. For larger copy sites, extensions involve multiple extension steps. Each extension after the initial one has only a reduced overhang. Moreover, since the concentration of single strands decays with increasing length, there are also less single strands available that can make proper use of the (remaining) overhang. Consequently, the initial linear slope flattens in a way determined by the interplay between reduced concentration and increased binding energy.

The existence of a plateau for very long copy sites is also easy to understand. The longer the copy site becomes, the less influence its actual length has, and only the concentration of building blocks plays a role. Moreover, the nonmonotonous behavior for copy sites of lengths around $l_{cs} \sim 6$ corresponds to another intrinsic scale of the dynamics. At a value of $l_{lig} \sim \frac{r_{lig}}{\gamma}^{-1} \sim 6$, time scales of dehybridization and *bare ligation* become comparable. Thus, the incorporation of building blocks around that size is limited by the bare ligation rate rather than dehybridization.

6.4.8. A closer look at the relation between C_{final} and $C_{initial}$

With the understanding gained in the previous section, we consider the conditional probability distribution $p(C_{initial}|C_{final})$. It is the probability that a complex that reached a certain length C_{final} started with an initial length $C_{initial}$. Like can be seen in Fig. 6.23, complexes with a final length $C_{final} < L_{max}$ started most likely to grow from an initial complexes that had same length (diagonal) or was only slightly shorter, $C_{final} \approx C_{initial}$. This reflects the hetero- and autocatalytic cycles discussed above. In contrast final complexes larger than L_{max} start most likely with a initial complex of length $C_{initial} \approx L_{max}$, *cf.* the median $\tilde{C}_{initial}$ in Fig. 6.23. We can understand this behavior as for reaching a long final complex, one has to start to form an initial complex with the longest single strand available, which is of length $L_{long} \approx L_{max}$.

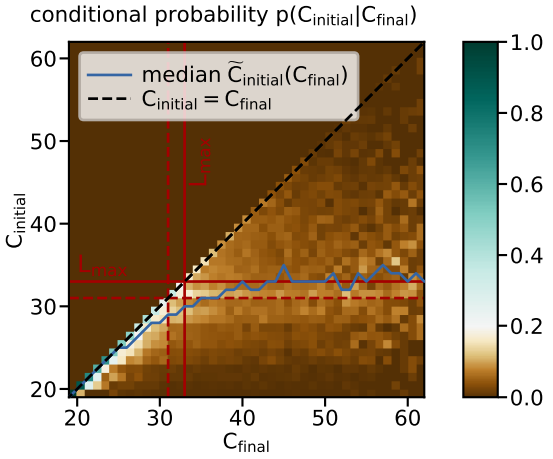


Figure 6.23.: Conditional probability distribution $p(C_{initial}|C_{final})$. It is the probability that a complex that reached a certain length C_{final} started with an initial length $C_{initial}$. Final complexes with a length $C_{final} < L_{max}$ start most likely with initial complexes of the same length $C_{final} = C_{initial}$ (diagonal), whereas longer final duplexes start most likely with initial complexes of length $C_{initial} \approx L_{max}$ (median).

6.4.9. Summary of results

In this chapter, we have analyzed self-assembly and growth processes via templated ligation. At first, we considered the transition from long-tailed to short-tailed distributions. In systems where dehybridization becomes slower with hybridization length, a competition between extension and dehybridization creates a strongly nonequilibrium regime leading to a nonmonotonous strand-length distribution. We then derived the typical scales that shape the strongly nonequilibrium regime for strands with typical lengths $L^* \lesssim L \lesssim L^\dagger$. The rapid production of long strands is self-enhancing via auto- and heterocatalysis. Catalytic enhancement stops as strands become too stable and thus inert at a typical scale $L^\dagger$, which is determined by the outflux rate. In a situation without outflux, the typical scale $L^\dagger$ is set by the balance of the observation time, and the time it takes for a reassembly, *i.e.*, dehybridization. The value of L^* as well as the building blocks used in extension cascades

are determined by the concentration and properties of the short strands in the weakly nonequilibrium regime.

The complexity of the process even in this simple, sequence-independent null model is remarkable. The universality of the arguments make it appealing to apply our insights to experiments performed on real oligonuclotides with random sequences.

6.5. Thermocycler experiments

Finally, we study an experimental system using DNA strands with random sequences from a binary alphabet. As in the simulation, we consider the strand length distribution formed by templated ligation from a reservoir of oligonucleotides of a fixed initial length acting as the starting material. Variations in temperature are achieved with a thermocycler. The ligation is performed by an evolved TAQ DNA ligase enzyme.

In order to observe the nonmonotonous length distribution, the time-scales of various processes need to be compatible. In particular, the following three requirements need to be fulfilled:

1. The effective extension rate r_{ext} determining the onset of extension cascades L^* must be larger than the rate r_{out} or τ_{obs}^{-1}, associated to the process that sets the length scale of the peak $L^\dagger$.

2. In a transient system without outflux, parameters must be such that τ_{obs} is still compatible with realistic experimental time scales.

3. To resolve the nonmonotonous nature of the strand length-distribution, the predicted length scales must be compatible with the experimental resolution, which is set by the length of the smallest building blocks.

While tuning these parameters independently of each other is easy in a simulation, this is not necessarily the case for real experiments. Usually, the generic experimental control parameters (like temperature, salt concentration, buffer composition) will affect the values of all time scales in a non-trivial manner. In particular, it may be possible that by varying a single parameter one cannot achieve an experimental situation, where all the above requirements are fulfilled. In what follows, we show that some of these difficulties can be overcome in a thermocycler, where temperature oscillation drive the extension-reassembly process.

6.5.1. Theoretical preliminaries

An important factor is the temperature dependence of the dimensionless binding energy $\gamma = \frac{\Delta G^\circ}{k_{\mathrm{B}}T}$. To zeroth order, the Gibbs standard free energy is obtained as $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$. The most drastic physical effects occur when the binding energy changes sign at the critical temperature $T_{\mathrm{c}} = \Delta H^\circ / \Delta S^\circ$, which is determined by the ratio of ΔH° and ΔG°. Around the critical temperature, a linear approximation yields

$$\gamma(T) = -\frac{\Delta H^\circ}{k_{\mathrm{B}} T_{\mathrm{c}}^2}(T - T_{\mathrm{c}}) = \frac{T - T_{\mathrm{c}}}{\sigma} \tag{6.30}$$

where $\sigma = -\frac{k_{\mathrm{B}}T_{\mathrm{c}}^2}{\Delta H^\circ} = -\frac{k_{\mathrm{B}}\Delta H^\circ}{(\Delta S^\circ)^2}$ has units of temperature and characterizes the (inverse) slope of $\gamma(T)$ around the critical temperature. In general, ΔH° and ΔS° also depend in a complicated way on the details of the hybridization site. Within our model we use an effective ΔH° and ΔS° per nucleotide.

6.5.1.1. Effective melting curves

Melting curves in oligonucleotide melting experiments can be interpreted as the probability $p(T)$ of a nucleotide to be unpaired for a given temperature. Such melting curves typically have a sigmoidal shape varying from $p = 0$ for low temperatures to $p = 1$ for high temperatures. A common fit function is the Fermi-like function

$$p(T) = \left(\exp\left(\frac{T - T_c}{\sigma}\right) + 1\right)^{-1}, \tag{6.31}$$

where T_c is the critical or "melting" temperature defined as $p(T_c) = \frac{1}{2}$. The parameter σ determines the width of the transition region.

Typical values for DNA composed of adenine (A) and thymine (T) nucleotides at common buffer conditions are $T_c \sim 55 \pm 10\,°C$, *cf.* Refs. [30, 40]. The width σ is typically on the order of 10 °C for short AT-oligomers, such that typical melting curves look like Fig. 6.24(a). In accordance with the thermodynamic considerations above, we use $\frac{T-T_c}{\sigma}$ as a proxy for an dimensionless binding energy per nucleotide γ, *cf.* Fig. 6.24(b).

While this mapping is certainly a crude approximation, we can use it as a proxy for the temperature dependence of γ within our effective model. As a consequence, the dehybridization rates $r_{off}(T) \sim r_0 \exp(\gamma L)$ should have a typical behavior as the one shown Fig. 6.24(c), where we chose a physically reasonable collision rate of $r_0 = 10^6\,s^{-1}$ [57, 43, 26]. We expect this mapping from melting curves to a parametrization for $\gamma(T)$ to only yield good results below the melting temperature, *i.e.*, for γ sufficiently negative. As soon as γ approaches zero, the kinetic and sequence-dependent details start dominating. Consequently, we do not expect a quantitative agreement between the real (experimental) dynamics and our analysis for systems above or close to the melting temperature. However, as we will see below, they can still be used for semi-quantitative prediction of experimental results.

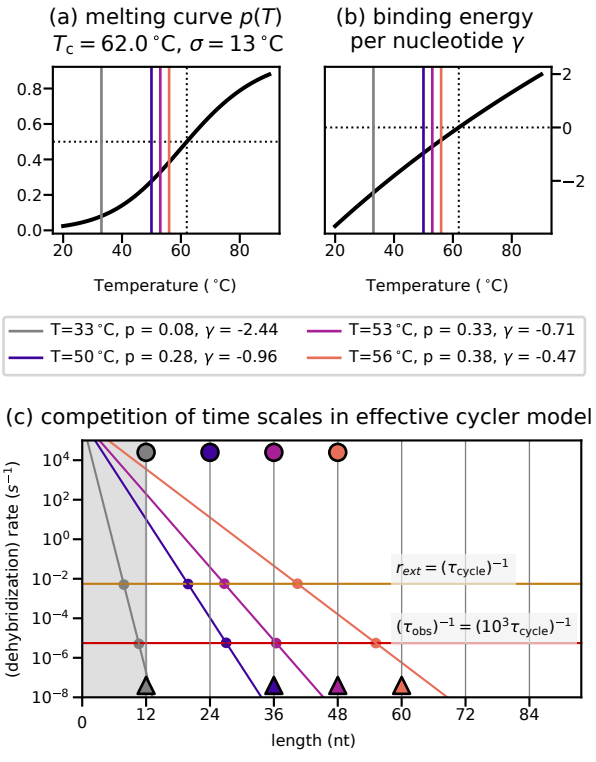


Figure 6.24.: (a) An effective melting curve with a critical temperature $T_c = 62°C$ and $\sigma = 13°C$. The gray line indicates the cold temperature $T_{cold} = 33°C$ which is optimal for extension. (b) Upon approaching the critical temperature, the binding energy in the hot phase approaches zero. (c) The effective dehybridization rate decays exponentially with a (log-)slope corresponding to the effective binding energy. Without cycling ($T_{hot} = T_{cold}$, gray curve), the system is simply too cold for anything too happen. Approaching the critical temperature, the binding energy and thus the slope become smaller in magnitude. Intersects with the horizontal lines mark the scales L^* and $L^\dagger$ (dots) and their ceiled values to the next higher multiple of $L_{bb} = 12$ (circles and triangles). Parameters: $\tau_{cycle} = 180\,s$, $r_{ext} = (\tau_{cycle})^{-1} = 5.56\times10^{-3}\,s^{-1}$, $\tau_{obs} = N_{cycle} \times \tau_{cycle} = 1.8\times10^5\,s$ and $r_0 = 10^6\,s^{-1}$.

6.5.1.2. Effective extension rates and thermocycling

Obtaining extension rates which are compatible with the above criteria for the emergence of a nonmonotonous length distribution may be challenging. In enzyme-free systems, templated ligation is a very slow process and thus not necessarily compatible with experimental time scales. Enzyme-assisted templated ligation with a ligase may speed up these extensions considerably but requires the formation of a chemical complex involving at least three strands plus the ligase itself. Generically, the probability of finding such a complex decreases with increasing temperature. Further, the ligase activity itself is temperature dependent. These effects generally lead to an experimental situation where the effective extension rate has a non-trivial temperature dependence.

In particular, one may encounter a "stalemate" situation in an isothermal system: On the one hand, for high temperatures, the extension rate is small because the formation of the complexes necessary for ligation is thermally suppressed. On the other hand, for low temperatures, the dehybridization rate is so small that the system is essentially frozen and the relevant dynamics come to a halt.

Fortunately, this stalemate can be resolved with the help of a thermocyler, which periodically cools the system to a temperature that is optimal for ligation *cf.* Fig. 6.25(a): During the cool phase, the rate of ligation is initially very high, until all ligation sites in existing complexes have been used up. However, the formation of any *new* complexes allowing for ligations is drastically slowed down. Hence, the hot phase is required in order to create the new ligatable complexes.

Recall that the extension rate r_{ext} is defined as the effective rate with which any given duplex binds a third strand which then subsequently ligates. In the scenario described above this would correspond one extension per cycle and thus $r_{\text{ext}} \sim \tau_{\text{cycle}}^{-1}$. In a transient experiment without outflux, the inverse observation time τ_{obs}^{-1} which replaces r_{out} in determining $L^\dagger$ (and thus $L_{\max}$) is given as $\tau_{\text{obs}} = N_{\text{cycles}}\tau_{\text{cycle}}$, where N_{cycles} is the number of cycles performed in the experiment. For cycles with a duration $\tau_{\text{cycle}} = 180\,\text{s} = 3\,\text{min}$ and $N_{\text{cycles}} = 1000$, we obtain the two horizontal lines shown in Fig. 6.24(c). The intersection of these lines with the dehybridization rate determines the scales L^* and $L^\dagger$. For a system whose smallest building blocks are of length $L_{\text{bb}} = 12$, the big symbols denote the values $L^*_\bullet$ and $L^\dagger_\blacktriangle$ obtained by ceiling to the next integer multiple of L_{bb}.

6.5.2. Experimental method and results

As described above, the experimental system consists of A (Adenine) and T (Thymine) only DNA strands with a building block length $L_{bb} = 12\,\text{nt}$ and random sequences, thus including all possible $2^{12} = 4096$ sequences. The ligation is performed by an evolved enzymatic molecule, the TAQ DNA ligase from *NEB*. This allows for very high ligation rates in comparison to chemical ligation [64]. In the experiment two strands are ligated in the same way as in the theoretical model: two (substrate) strands hybridize on a third (template) strand, that overlaps both substrate strands, by Watson–Crick base pairing. The ligase then connects the sugar backbones of the 3′ end of the first strand to the 5′ end of the second strand. The rather high building block length $L_{\text{bb}} = 12$ was enforced due to the properties of the ligase. This experimental setup is similar to the setup used in [103]. The resulting coarse resolution for the strand-length corresponds to the discretization shown in the schematic plot for effective thermocyler experiments shown in Fig. 6.25(b).

Unlike the theory described above, the experimental setup is a closed system without in- or outflux and thus does not reach a comparable steady-state. While transient observation time τ_{obs} is connected to the outflux rate r_{out} (see Sec. 6.4.4), depletion effects are present.

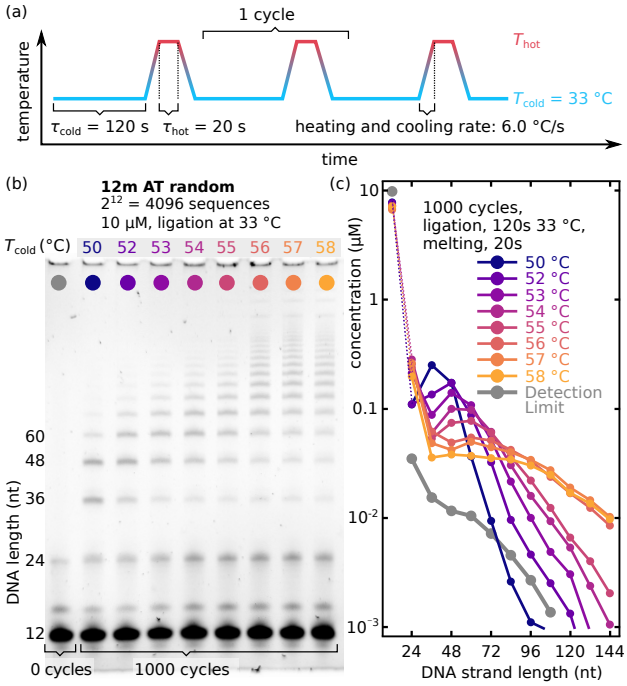


Figure 6.25.: Product concentration analysis for a 12nt random sequence AT-only pool. (a) Experimental temperature profile. Ligation occurs for 120 s at 33 °C after which the sample is heated to the variable hot reassembly temperature T_{hot} for 20 s. (b) Image of a PAA gel with SYBR gold post stained DNA. The first lane on the left shows the "baseline" sample, which is similar to the other lanes but was not subjected to temperature cycling. The other lanes have the same ligation conditions but different temperatures for dissociation. (c) Quantitative results for the strand-length distribution obtained via our custom software. From 50 °C to 58 °C the transition of a quickly exponentially falling product length distribution to a shallowly decreasing exponential distribution is notable. The transition shows the feature simulated before, with a clear peak.

Analysis of the length distribution are done by running the samples in polyacrylamide gel electrophoresis (PAGE), post staining the DNA with intercalating SYBR gold dye and taking fluorescent images of the gel in a *BioRad* ChemiDoc MP. Concentration quantification for the experimental data is then done with a custom software that extracts lane intensity from gel photos. For the baseline correction and normalization the software needs a reference sample in one lane per gel that is the same as all the other samples, but was stored in the fridge and not subjected to temperature cycling, as described in Sec. A.6.3 in the Supplemental Material. The bands visible at lengths of 16 and 24 nt are artifacts from buffer and DNA synthesis and visible for all lanes.

The temperature cycling necessary to prevent the stalemate conditions discussed above are done with a *ThermoFisher* ProFlex PCR system thermocycler. We analyzed the strand length distribution for various observation times τ_{obs} at different isothermal conditions, whereas the temperature alternates between $T_{cold} = 33\,°C$ for $\tau_{cold} = 120\,s$, and $\tau_{hot} = 20\,s$ at variable temperature T_{hot} for cycling conditions. Linear temperature ramps of 20 s connect the hot and cold phases, as shown in Fig. 6.25(a).

Isothermal experiments resulted in no product formation within 60 and 116.5 hours for none of the considered temperatures. We could not observe strands longer than the initial 12 nt (see Fig. S11), as already expected for the stalemate condition.

For the temperature-cycled experimental conditions, the variation in T_{hot} yields very different product distributions, as shown in Fig. 6.25(b). The strand-length distribution decays quickly for a dissociation temperature of $T_{hot} = 50\,°C$, with much slower decay at $T_{hot} = 58\,°C$. All strand-length distibutions show a non-monotonous behaviour exhibiting a local minimum with L_{min} between 36 and 48 nt and a maximum at L_{max} ranging from 36 and 72 nt. Note that these lengths are consistent with the semi-quantitative predictions of the effective cycler model shown in Fig. 6.24(c). For higher dissociation temperatures, the peak becomes flatter and wider. Importantly, the overall shape of the distribution changes significantly in a very limited temperature range for T_{hot} in the vicinity of typical melting curves for DNA composed of A and T only.

We also recorded the transient behavior of the strand length distribution at various temperatures. Fig. 6.26 shows how the multimer composition changes with T_{hot} of 52, 54

and 56 °C over the curse of 500, 1000, 1500 and 2000 temperature cycles. For $T_{\text{hot}} = 52\,°\text{C}$ at 500 cycles, the length distribution is quickly monotonously decaying and the maximum appears not before 1000 cycles. For $T_{\text{hot}} = 54\,°\text{C}$ we already see a saddle point at 500 cycles at 48 nt (corresponding to the position of the developed minimum L_{min}) and for $T_{\text{hot}} = 56\,°\text{C}$ the maximum has already fully formed at 500 cycles.

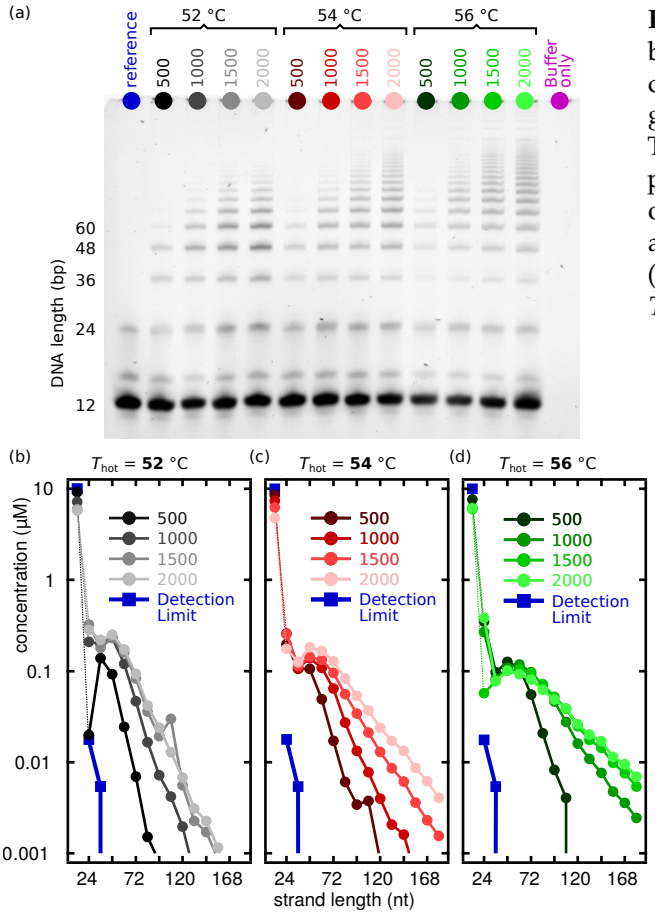


Figure 6.26.: Transient strand-length distributions after 500, 1000, 1500 and 2000 cycles. (a) Gel electrophoresis image of SYBR gold stained DNA with marked sample lanes. The reference lanes is the same for all samples. The rightmost lane is the ligation buffer only and shows no bands. Quantitative analysis of the strand-length distribution for (b) $T_{\text{hot}} = 52\,°\text{C}$, (c) $T_{\text{hot}} = 54\,°\text{C}$ and (d) $T_{\text{hot}} = 56\,°\text{C}$

6.5.3. Comparison with theory

Our experimental results are in good qualitative agreement with our theory. The isothermal system does not show any product for reasonable experimental time scales. We believe the system to be in the stalemate situation described above: For low temperatures, even the shortest duplexes with strands of length L_{bb} cannot separate efficiently. For high temperatures, the effective extension rate is suppressed because virtually no stable ligatable complexes are formed.

For the cycled systems, the results are consistent with the effective theory described above: We obtain long-tailed length-distributions with a pronounced ladder of long strands for all probed values of T_{hot}. This strongly suggests that during the cold phase at $T_{\text{cold}} = 33\,°\text{C}$, the system allows for an effective ligation of available ligatable complexes.

We were able to observe non-monotonous length distributions for all dissociation temperatures between $50\,°\text{C} \leq T_{\text{hot}} \leq 58\,°\text{C}$. The experimentally observed values of L_{min} and L_{max} and their change with temperature agree well with the effective theory illustrated in Fig. 6.24(c). Similar as in standard polymerase chain reactions, the main role of temperature cycling in the experiment is to drive the reassembly process by separating long fully-hybridized strands.

We conclude that the basic principles shaping the self-assembly process are sufficient to understand experimental results. This is a remarkable, since there are various factors that make the experiment different from the idealized simulation conditions: For instance, depletion of building blocks and the degeneration of chemicals play a role in the experiment. Self-folding of longer strands may also be an important experimental mechanism that is absent from the simple theory, as suggested in [103]. Moreover, the quantitative evaluation of the gel-electrophoresis plots becomes more difficult for long strands, due to the resolution limit.

Future theoretical work could build on the general understanding established in this work, guided by further experiments. While such extended theory and experiments are certainly important, a more exhaustive experimental study is beyond the scope of this work.

6.6. Summary and discussion

Self-assembly and growth of oligonucleotides by template-directed ligation play a key role in prebiotic chemistry and the emergence of self-replicating systems on early Earth. The microscopic dynamics of this process exhibits a large complexity due to the vast amount of possible chemical structures involved. As a consequence, the majority of previous studies has either focused on pure primer extension scenarios [99, 90, 53, 72, 45, 49, 79, 86] or considered simplified models [78, 74, 89, 47, 82, 70, 92, 98, 95, 88, 8].

In this work, we presented and analyzed a model for this self-assembly process. Ignoring any dependence on sequences, the binding energy of two oligonucleotide strands only depends on the length of their hybridization site and ligation occurs with a constant rate. Crucially, we studied the self-assembly and growth of short building blocks via templated ligation in an *ab initio* scenario without the requirement for further *a priori* assumptions.

We showed that the strand-length distribution arising in this situation is determined from the competition of three natural time scales, or equivalently, their corresponding rates:

1. The dehybridization rate r_{off} which decreases exponentially with strand length L with a rate determined by the binding energy per nucleotide γ.

2. An effective extension rate r_{ext} of strands in hybridization complexes, which is determined by the ligation rate r_{lig}, γ and system properties like initial conditions and/or coupling to an environment.

3. A global time-scale either determined by the outflux rate r_{out} (which is the inverse average life-time of any complex) or a global observational time scale τ_{obs} (which is the maximal life time of any complex in a transient system).

The competition between r_{ext} and r_{out} (2 and 3) determines whether we see a long-tailed distribution at all: If r_{out} is larger than r_{ext}, there is not enough time on average to have any ligations. The competition between r_{off} and r_{ext} (1 and 2) leads to the emergence of extension cascades at a typical length scale L^*: As soon as strands in a hybridization complex have a length such that $r_{\text{ext}} > r_{\text{off}}$, they will undergo extension cascades that lead to persistent configurations, which can not extend any further. The fate of such a persistent configuration is determined by a competition between r_{off} and r_{out} (1 and 3): Fully-hybridized duplexes that are shorter than $L^\dagger$ dehybridize before they leave the system. The single strands created in this way act as templates in other extension cascades, enabling further strand growth.

We showed that these simple arguments allow us to explain the assembly and growth dynamics of oligonucleotides via templated ligation in both *in vitro* and *in silico* experiments. In what follows, we discuss our results in the context of the broader research field.

6.6.1. The simulation framework

A major part of this work focused on the emergent length-scales defined via the competition of time scales in the self-assembly process of long strands. While the results may be obvious in hindsight, having access to a microscopic stochastic simulation was key in their formulation. Moreover, the detailed analysis of self-assembly and primer extension would not have been possible without access to simulated trajectories of individual chemical structures.

Even though the concept of such a simulation seems natural, no simulation framework capable of handling the full complexity of the process was available. The biggest challenge in any implementation is the immense size of the stochastic state space. At any time, the system state is defined by an occupation number for each of the infinite number of species defined by different hybridization complexes. Enumerating or counting the different species is already a non-trivial combinatorial problem. As a consequence, there is no practical way to specify the states and the stochastic transitions between them *a priori*. Instead, our algorithm dynamically generates the relevant part of the network of states just in time. Even then, due to the vast range of relevant time scales in this problem, simulations were executed on high-performance computing hardware and optimized for speed.

6.6.2. Joint experimental, computational and theoretical efforts

Experimental studies of templated ligation and other prebiotically relevant dynamics of informational polymers have attracted considerable attention in recent years [100, 58, 14, 11, 13, 105, 99, 90, 53, 72, 45, 49, 79, 86, 56]. A huge experimental challenge for probing the origins of life *in vitro* are the potentially long time scales. While interesting prebiotic phenomena can occur fast in terms of geological time scales, the involved time scales can be prohibitively long from the perspective of the relevant experiments.

We have seen how this problem occurs in isothermal experiments and how it can be overcome with the use of thermal cycling. Finding the correct points in a large experimental parameter space is a challenge and usually requires good experimental intuition. Analytical arguments and the simulation tool brought forward in this work can help to guide this intuition. In particular, the theory presented here allows us to experimentally *control* the characteristic scales of oligunucleotide structures emerging in the self-assembly process.

6.6.3. Step by step towards the RNA world

Accepting the RNA world hypothesis, one of the central question regarding the origin of life is the path from prebiotic chemistry to the first self-replicating systems of informational molecules [61, 71, 100].

Since major transitions in evolution appear to have occurred due to smaller entities coming together to form larger entities [19], a multi-steps scenario also seems natural in a prebiotic context, *cf.* Fig.6.27. While the importance of templated ligation in this scenario is generally accepted [18], the mechanisms emerging from the combined action of short building blocks are not yet fully understood. While most other work focuses on scenarios further down the evolutionary road, our study shows the emergence of oligomer structure in a kinetically and thermodynamically consistent *ab initio* scenario.

Since both characteristic scales are to leading order proportional to the reciprocal binding energy per nucleotide in units of $k_B T$, their values will differ for different nucleotide pairs with different thermodynamic properties [83, 52, 32, 23].

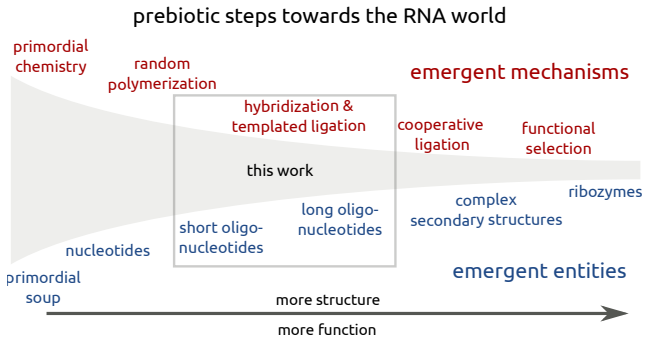


Figure 6.27.: Evolution is a multi-step process that creates new emergent entities which exhibit emergent mechanisms of interaction. As a process far from equilibrium, evolutionary dynamics is able to funnel the phase space of all possibilities into distinct regions exhibiting ever more complex structural entities. Our work outlines the emergence of structured oligonucleotides from the smallest building blocks in a thermodynamically and kinetically consistent model.

In addition, the scale of the start of extension cascades, L^*, crucially depends on the kinetic aspects of the ligation reaction [76, 58].

The typical strand-length scale, $L^\dagger$, on the other hand emerges from a competition of transient or degradation time-scales and dehybridization. It is thus independent of the ligation kinetics. Coupling to other non-equilibrium driving forces like temperature or concentration cycles which naturally occur in prebiotic environments, shape the relevant time and length-scales [48, 97, 85]. Our experiments show that the value of the emergent characteristic length scales can be tuned without changing the underlying chemistry of the oligonucleotides.

As such, the strands of a characteristic scale $L^\dagger$ produced by the basic self-assembly mechanisms presented here, can act as the basic building blocks of a higher-level form of organization. For instance, the work of Eigen and Schuster on hypercycles and quasi-species [6] as well as Kauffman's notion of an autocatalytic set [17] provided valuable conceptual frameworks. However, recent theoretical advances along this lines have relied on premises arising from coarse-grained assumptions [8, 55, 74, 88, 82, 100, 47]. Our results and the use of an elementary stochastic simulation framework will help to provide a solid base for further research.

6.6.4. Connection to evolutionary dynamics

While various definitions of the necessary properties of evolutionary dynamics exist, most researchers agree that they include:

- **Mutation**, *i.e.*, stochasticity, possibility of rare events.
- **Selection**, which arises from replication (*i.e.*, autocatalysis) coupled to global constraints (*i.e.*, global degradation, carrying capacities *etc.*).
- **Cooperation**, for instance in the form of mutually catalytic networks, see Refs. [100, 95].
- **Structural entities** that persist and allow for information storage (strands, sequence motifs, genes, cells, individuals).

The results and the simulation method used in this chapter lend themselves naturally to further studies. While we did not treat sequence information explicitly, the emergent self-assembly and growth processes in our model exhibit all of these properties: Mutation arises

directly from the stochastic simulation framework. Selection results from a competition between dehybridization and degradation (or transient) time scales and thus leads to structure on the level of strand lengths. The combined hetero- and autocatalytic nature of the assembly process emphasize the importance of cooperation, see Sec. 6.4.7.

Since extension cascades enable the classical primer-extension scenario, additional selection and cooperative behavior is a natural consequence, *cf.* Ref. [100]. Under this premise, the results of Refs. [100, 95] as well as the more conceptual ideas of Eigen and Kaufmann imply structure formation in sequence space.

Finally, our work shows the emergence of distinct structural entities, *i.e.*, strands of a typical length, in a prebiotically plausible self-assembly scenario. In a linguistic analogy, where letters amount to nucleotide identities, our work allows us to understand the typical length of words and sentences: Letters and words occur as the small, single strands of length $L \lesssim L^*$. In contrast, the length of typical sentences is determined by the scale $L^\dagger$. The catalytic self-assembly dynamics analyzed in this context then allow us to understand a part of the underlying syntax of the prebiotic language. However, the semantics of this language only emerges in the context of an evolutionary process. We hope that our work contributes towards understanding a crucial step within the larger story.

6.6.5. Switching mode of operation of a hypothetical protocell genome

In Section 6.4.3 we saw that the average strand length as a function of the cutoff rate $r_{\text{cut}} = e^{\gamma l_{\text{cut}}}$ exhibits a maximum at $l_{\text{cut}} = 16$. The average length at the maximum is eight times larger than the average length of a system without a cutoff, $l_{\text{cut}} \geq 50$. Like discussed, this corresponds to a growth mode where duplexes dehybridize after a single extension leading to a rapid increase of strand length. However, this mode of operation does not allow for copying of longer sequences in a templated fashion. We consequently term this regime the *fast incorporation regime* and the regime without cutoff the *copy regime*.

We can transition from the copy regime towards the fast incorporation regime by *e.g.*, increasing the ambient temperature or lowering the pH which reduces the melting temperature [91].

While the change in temperature can only be induced through a change in the environment, there are indications that the latter could also be self induced by the membrane growth of a hypothetical protocell: In [62] they show that the pH of a model protocell with a fatty acid bilayer can be reduced by constant incorporation of fatty acids into the membrane (growth).

Such a coupling between the growth of the membrane and the mode of operation of the genome (copy mode, fast incorporation mode) could enable a synchronization of inner protocell processes to its growth-division cycle.

6.6.6. Outlook

The null model considered in this chapter serves as the most simple model allowing the study of structure formation through templated ligation. Never the less the secondary structures so far included in the model are limited, most evident in the fact that we excluded self folding of strands. Self folding could lead to the emergence of yet an other length scale and it would be interesting to see its impact on the length distribution. Though desirable, extending the algorithms to include self folding is a possible but daunting task. Further, coupling the system to an influx and outflux that mimic diffusion of informational polymers through membranes could be considered. In particular, this would require fluxes which are length dependent and proportional to the difference in the concentration between the

environment and the reactor. One could for example consider a scenario where complexes of length larger than $C \geq 3$ can not leave the system. Cell division can be mimicked relatively simple by stochastically removing half of the complexes or by implementation of serial dilution. Further, the possibility of driving time resolved temperature cycles in the reactor would be beneficial as it would allow for a more direct comparison with the experimental setup.

7. Preliminary results for a sequence-dependent model

The simulation by design can handle sequence-dependent energy models. Here, we show preliminary results for a system of oligunucleotides with a binary alphabet. The energy model used here is a nearest-neighbor model. Non-complementary base pairs introduce a thermodynamic penalty and optionally a stalling effect on the rate of templated ligation. The model extensions got designed in cooperation between Tobias Göppel and me. The simulation code got extended by Tobias Göppel to include the breaking of oligonucleotides (hydrolysis) for both single and double strands. He also wrote the code for the analysis of the error fractions, discussed below.

7.1. Overview of the model

The model presented here uses a binary alphabet $\mathcal{A} = A, U$, where $A.U$ and $U.A$ are considered matches, whereas $A.A$ and $U.U$ are mismatches. This constitutes a simplification with respect to the four nucleotide alphabet used today by biology; however, it is considered a possible prebiotic scenario [2, 15, 58]. Further, there are examples for ribozymes composed of only two bases [29, 42] and many theoretical models use a two-letter alphabet for simplicity, *cf.* Refs. [8, 39, 58, 82, 70, 41, 104, 37].

Energy model for a binary alphabet

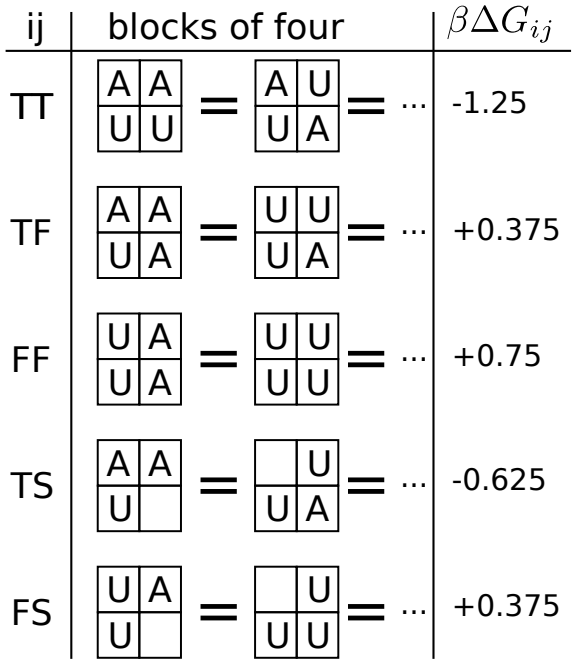


Figure 7.1.: Dimensionless binding energies $\gamma = \frac{\Delta G}{k_B T}$ for all possible block configurations. Energy values only depend on the number of matching base pairs (T) and unpaired nucleotides (S). The configurations TS and FS consisting of 3 nucleotides are also known as "dangling end" contributions in common nearest neighbor models [52, 32].

We choose an energy model that assigns a binding energy to each possible 4-block of

neighboring nucleotides, *cf.* Fig. 7.1. As such, it is similar to common nearest neighbor models [52, 32]. Here, for the sake of simplicity, we only distinguish between complementary (T) and non-complementary (F) nucleotide pairs and single (S) nucleotides, *cf.* Fig 7.1.

Sequence-dependent ligation rates

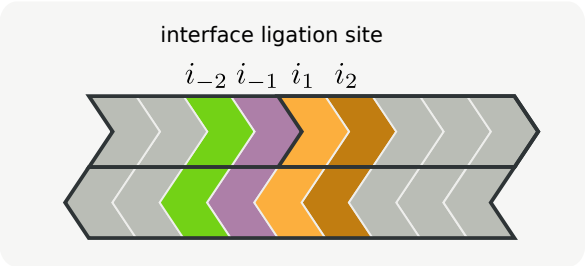


Figure 7.2.: Mismatches in the vicinity of the ligation site decrease the ligation rate. We include the influence up to the next nearest neighbors of the ligation site.

From experimental observations [49], it is well known that enzyme-free extension of a primer is considerably slowed down after a mismatch. This effect is usually known as "stalling" and has been shown to depend on the presence of mismatches around the ligation site. Stalling due to non-complementary base pairs is also observed for ligation-reaction catalyzed by a ligase [46, 77].

In our extended model we can include stalling by making the bare rate of templated ligation, r_{lig}, explicitly dependent of the sequence context. In particular, the rate of templated ligation decreases for any mismatch at a distance up to two nucleotides away from the ligation site, see Fig. 7.2.

To be close to experimental observation, we parametrize r_{lig} in the following way:

$$r_{\text{lig}}(i_{-2}, i_{-1}, i_1, i_2) = \frac{r^0_{\text{lig}}}{s(i_{-2}, i_{-1}, i_1, i_2)}. \tag{7.1}$$

In this formula, r^0_{lig} is a neutral (maximal) ligation rate. The neutral ligation rate is modified by a stalling factor $s(i_{-2}, i_{-1}, i_1, i_2) \geq 1$ that depends on the sequence context of nearest (index ± 1) and next-nearest neighbours (index ± 2) of the ligation site. Using the boolean function $b(i)$ for the basepair bp_i at position i,

$$b(i) = \begin{cases} 0, \text{ if } \text{bp}_i = T, S \\ 1, \text{ if } \text{bp}_i = F \end{cases}, \tag{7.2}$$

a convenient multiplicative form of the stalling factor is given by

$$s(i_{-2}, i_{-1}, i_1, i_2) := \sigma_1^{\sum_{j\in\{-1,1\}} b(i_j)} \sigma_2^{\sum_{j\in\{-1,1\}} b(i_j) b(i_{2j})} \tag{7.3}$$

where σ_i is an elementary stalling factor attributed to mismatches that are i base pairs away from the ligation site. Values of $\sigma_1 = 20$ and $\sigma_2 = 10$ are consistent with experimtal data in [64, 49].

7.1.1. Degradation by hydrolysis and double-strand breakage

In addition, we implemented cleavage of single strands with elementary rate r_{cs} and the breaking of double strands with rate r_{db}, *cf.* Fig. 7.3. A process leading to the spontaneous cleavage of single-stranded segments is auto-hydrolysis, while double-strand breakage can

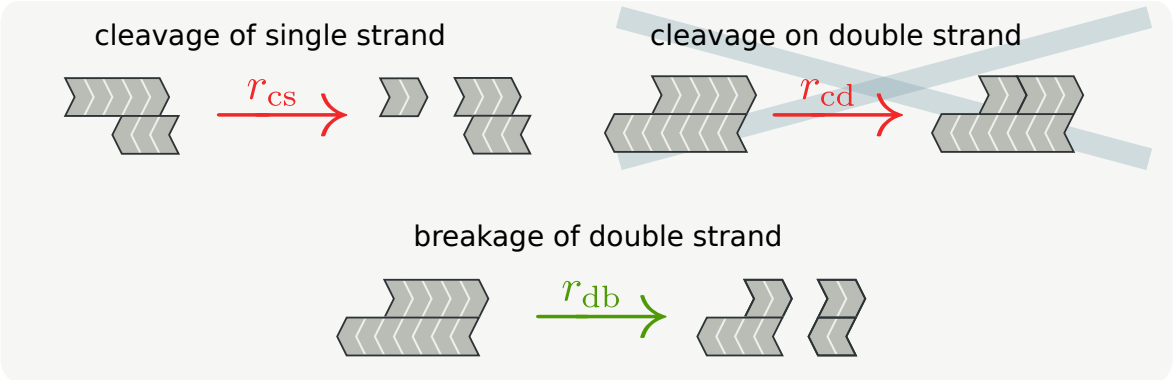


Figure 7.3.: (top left) Cleavage of single strand with elementary rate r_{cs}. (top right) Cleavage of a hybridized strand which is not included in our model. (bottom) Breakage of a double strand with elementary rate r_{db}.

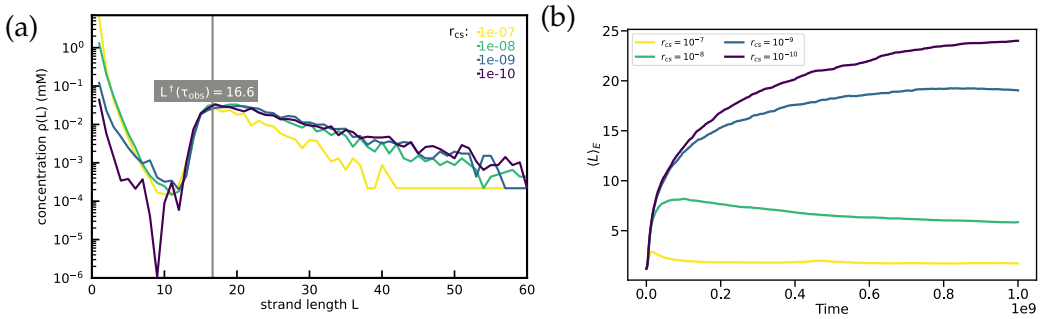


Figure 7.4.: Closed binary alphabet system with cleavage of single strands as the only degradation mechanism. (a) Strand-length distribution for various values of the single-strand cleavage rate r_{cs}. In all cases, the position of the maximum is predicted by the competition of the length-dependent dehybridization time, r_{off}^{-1} and observation time τ_{obs} via Eq. (7.4). (b) The mean strand length as a function of time.

be caused by ionizing radiation [22, 68, 75]. For the purpose of this simple model extension, we neglected auto-hydrolysis of strand segments that are hybridized, see Fig. 7.3 (top right). Experimental evidence suggests that this phenomenon is negligible since the double-stranded configuration inhibits the attacking 2'-OH from attacking the phosphodiester bond [21]. Further, the rates of auto-hydrolysis are different between DNA and RNA, with experiments suggesting that under similar conditions, the rate constant for DNA auto-hydrolysis is by a factor of 1×10^5 smaller than that of RNA [22]. For r_{lig}^0 we used the standard value of Chapter 6, $r_{lig}^0 = e^{-6}$.

7.2. Preliminary results

We report preliminary results on the strand-length distribution obtained in this model and its dependence on the cleavage processes. In particular, we demonstrate that a time-scale (or rate) analysis that is similar to the one presented in Chapter 6 can be used to predict the relevant features of the distribution.

7.2.1. Cleavage of single strands

We first consider a closed system where we only allow for cleavage of single-stranded segments. The system is initialized with a mixture of all possible n-mers at at initial

concentrations c_n with values $c_1 = 8.7\,\mathrm{mM}$, $c_2 = 0.696\,\mathrm{mM}$, $c_3 = 0.348\,\mathrm{mM}$ and $c_4 = 0.174\,\mathrm{mM}$. Sequence-dependent parameters are kept constant for all simulations. The single strand cleavage rate r_{cs} (as always, expressed in units of r_0) is then varied between between 1×10^{-10} and 1×10^{-7}.

Fig. 7.5 (a) shows the resulting length distributions at an observation time $\tau_{\mathrm{obs}} = 1\times10^{9}$. Depending on whether the inverse cleavage rate is comparable to the system time, systems have (approximately) reached a semi-stationary state. This behavior is visualized by the time-dependent average length shown in Fig. 7.5 (b).

We see the emergence of the typical shape of the non-monotonous strand-length distribution in all cases. Importantly, the position of the maximum is independent of the cleavage rate.

This observation is consistent with our general understanding of the competition of time scales involving the dehybridization rate of duplexes. Since duplexes are not affected by cleavage at all, the single-strand cleavage rate does not influence the position of the maximum. Notice, however, that it changes the distribution of the short strands, which are predominantly single-stranded. Moreover, since we consider a closed system here, the overall mass of the system is constant, and thus the exact form of the distribution will depend on this semi-stationary balance.

For $r_{\mathrm{cs}} \geq 10^{-8}$ a steady state was reached and L_{max} can be estimated by calculating the strand length for which the dehybridization rate of a fully hybridized configuration with zero mismatches becomes equal to the total double-strand breakage rate $(L-1)r_{\mathrm{db}}$ of the duplex:

$$L^{\dagger}(\tau_{\mathrm{obs}}) \text{ via } \frac{e^{-1.25(L-1)}}{2L-1} = \frac{1}{\tau_{\mathrm{obs}}}. \tag{7.4}$$

Solving this equation yields the prediction $L^{\dagger}(t_{\mathrm{sys}}) = 16.6$, which is represented by a vertical bar in Fig. 7.5 (a), which agrees well with the position of the maximum.

For completeness, we also note that increasing the value of the degradation rate triggers another transition, akin to the transition from long- to short-tailed distributions for high outflux rates. If degradation is simply too fast with respect to extension, no long-tailed distribution can form. In the present situation, this competition can even lead to the "extinction" of all long strands: Starting at a cleavage rate of about $r_{\mathrm{cs}} = 10^{-6}$, the system quickly reaches a state where only monomers are present. Then, any strand growth becomes impossible since templated ligation requires at least dimers as template.

7.2.2. Double-strand breakage

Next, we consider the same system as in the previous section but also allow for double-strand breakage at a rate which we assume to be 100 times slower than single-strand cleavage, i.e., $r_{\mathrm{db}} = r_{\mathrm{cs}}/100$. In contrast to the single-strand cleavage rate, double strand-cleavage is a process that affects the extension-reassembly process and thus, potentially the position of the maximum $L^{\dagger}$.

However, we will only see that effect when the lifetime of the most stable duplexes becomes shorter than or comparable to the transient system time. For the given parameters, we would thus assume such a stationary balance only for the largest double-strand breakage rates. Then, $L^{\dagger}$ can be calculated by equating the rate of dehybridization of a fully-hybridized

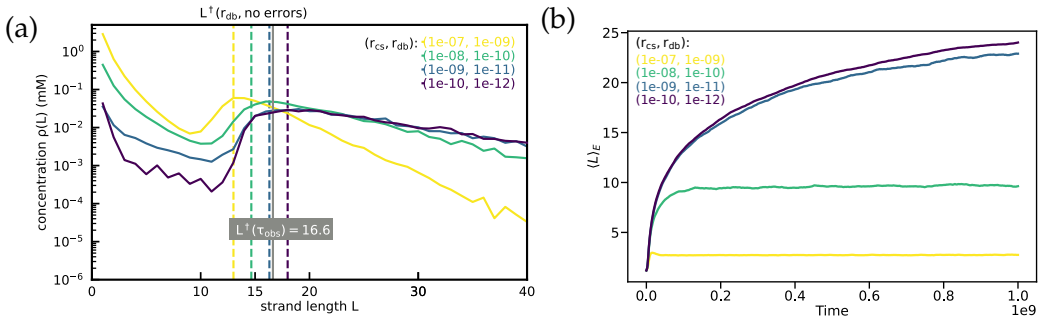


Figure 7.6.: Simulation results with single strand cleavage and double strand breakage r_{cs} and $r_{db} = r_{cs}/100$: (a) Strand length distributions. Vertical dashed lines indicate the corresponding $L^{\dagger}$ calculated from Eq. (7.5). The grey solid line is the length-scale determined by the system time via Eq. (7.4) and independent of breakage. (b) Mean strand length as a function time.

duplex with its (total) breaking rate $(L-1)r_{db}$:

$$L^{\dagger}(r_{db}) \text{ via } \frac{e^{-1.25(L-1)}}{2L-1} = (L-1)r_{db}. \tag{7.5}$$

Again, for the energy parameter $\gamma = -1.25$ we use the value corresponding to matching blocks. For the fastest double-strand breakage rates, $r_{db} = 1\times10^{-9}$ and $r_{db} = 1\times10^{-10}$, the formula (7.5) yields $L^{\dagger} = 13.0$ and $L^{\dagger} = 14.65$, respectively. Fig. 7.7 (a) shows the simulated strand-length distributions.

For the largest rates of breakage, the scales determined by Eq. (7.5) coincide well with the observed maxima at $L_{max} = 13$ and $L_{max} = 16$. For slower rates of double-strand breakage, the system time τ_{obs} determines the limiting rate, and Eq. (7.4) needs to be used. The variation of the mean strand length as a function of time confirms this transient behavior, *cf.* Fig. 7.7 (b).

7.2.3. Error fraction of fully hybridized strands

In the system considered in the last section we saw that the estimate for L_{max} obtained via considering only fully hybridized duplexes (Eq. (7.5)) underestimated the actual value, $L^{\dagger}(10^{-10}) = 14.65 < L_{max}(10^{-10}) = 16$. This can be interpreted as an indication that for larger L_{max} (smaller r_{db}) mismatches need to be considered in formula Eq. (7.5). But in this case, Eq. (7.5) would loos its predictive power as we do not have a formula for the average numbers of mismatches contained in a strand. Let us therefore consider the error fraction of fully hybridized duplexes, see Fig. 7.8.

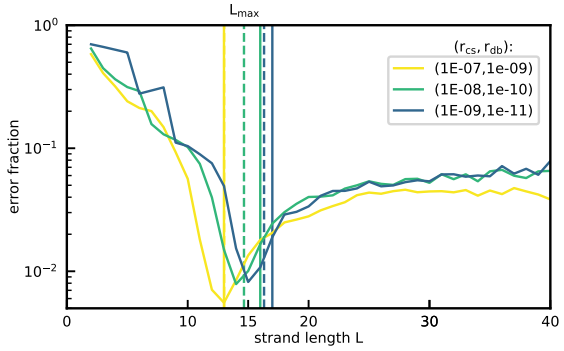


Figure 7.8.: Error fraction of fully hybridized duplexes vs strand length. Only for the largest cleavage rate the position of the maximum (vertical lines) coincides exactly with the position of the minimum.

The error rate seems to decay exponentially for small lengths until $L \approx L_{min}$ where we

conjecture the onset of extension cascades, L^* (Fig. 7.8). At $L \approx L_{\min}$ the error fractions drops rapidly, forming a minimum at around $L_{\max}$. Hence, the appearance of primer extension cascades leads to a strong decrease of the error fraction. For $r_{\mathrm{db}} = 1\times10^{-7}$, $L_{\max}$ is exactly at the position of the minimum in the error rate, but for $r_{\mathrm{db}} < 1\times10^{-7}$, there is an offset between both quantities. Like expected from the fact that $L^\dagger(1\times10^{-10}) < L_{\max}(1\times10^{-10})$, the error rate at $L_{\max}(1\times10^{-10})$ is larger than the error rate for $L_{\max}(1\times10^{-9})$.

Correcting formula Eq. (7.5) by neglecting the possibility of multiple errors, we can estimate $L^\dagger(1\times10^{-10})$ by:

$$
\begin{aligned}
p_e(0) + p_e(1) &\approx 1 \qquad (7.6)\\
p_e(1) &= 0.272\\
\langle \Delta G \rangle (L) &= (L-1) * (-1.25) * (1-0.272) + ((L-3) * (-1.25) + 2 * 0.375) * 0.272\\
L^\dagger(r_{\mathrm{db}}) \text{ via } \frac{e^{\langle \Delta G \rangle (L)}}{2L-1} &= (L-1) r_{\mathrm{db}},
\end{aligned}
$$

where $p_e(n)$ is the probability that the duplex contains n mismatches. Applying this formula yields $L^\dagger(10^{-10}) = 15.29$, which is closer to the actual value $L_{\max} = 16$.

7.2.4. Double strand breakage in a system without stalling

In order to demonstrate the principles discussed above in a system closer to the null model discussed in Chapter 6, we again run the simulation of the double strand breakage system using identical conditions as above, but without a context dependent stalling factor, hence $r_{\mathrm{lig}}(i_{-2}, i_{-1}, i_1, i_2) = r^0_{\mathrm{lig}} = e^{-6}$. The results are shown in Figure 7.9. It is to no surprise that the results are qualitatively identical, with the exception that the error rate seems to be larger in the system without stalling, compare Fig. 7.8 with Fig. 7.10 (c). Hence, as one would expect, the context-dependent stalling factor reduced the error rate, and the positions of the maximum could be better approximated with the formula only considering matches, Eq. (7.5).

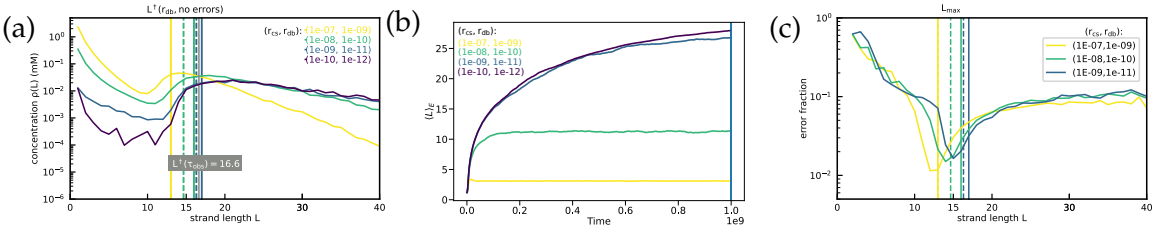


Figure 7.9.: Simulation results with single strand cleavage and double strand breakage r_{cs} and $r_{\mathrm{db}} = r_{\mathrm{cs}}/100$: (a) Strand length distributions. Vertical dashed lines indicate the corresponding $L^\dagger$ calculated from Eq. (7.5). The grey solid line is the length-scale determined by the system time via Eq. (7.4) and independent of breakage. (b) Mean strand length as a function time.

7.3. Conclusion

In this chapter, we have shown that the basic arguments presented in the simple, sequence-independent null model can also be applied to systems with a sequence-dependent energy model and a different degradation mechanism (breakage of strands). By using a sequence-dependent ligation rate that includes stalling after mismatches, we further showed that the important features of the assembly process do not depend on these microscopic details.

Further, we demonstrated that the assembly processes far away from equilibrium will have an impact on the error rate of the fully hybridized complexes and hence on the sequence space. It is an intriguing open question to study how information is copied in such a system. Further, this study should be repeated with an alphabet of four nucleotides in the future.

8. Summary and outlook

A main goal in the origins of live research is finding a system of interacting informational polymers that exhibits replication of information. As there was no appropriate simulation software available to approach this question, we developed the simulation framework presented in this thesis. In the last chapter, we demonstrated the capability of the simulation to handle more evolved models, including cleavage and sequence dependence. One of the future challenges will be to not become lost in the forest of endless potential models that can be simulated. The null-model studied in detail in [106] and this thesis will be of great help in selecting promising models in order to find a self-replicating system. Further, quantities that allow measuring the degree of replication must be determined.

8.1. Toehold/branch migration

A possible extension of our simulation could be the toehold/branch migration mechanism, as experimentally studied in [101]. The authors claim that it could circumvent the strand separation/re-annealing problem. Nevertheless, the article does not explain how the toehold configuration does form from a fully-hybridized configuration in the first place.

The separation/re-annealing problem describes the following process: After a full extension of a primer-template duplex, the fully-hybridized duplex is separated by a rise in temperature (or change in pH, or salt concentration *etc.*). Thereafter the temperature is lowered again, but the formation of new primer-template duplexes is suppressed due to the re-hybridization of the template strand and its full complement, which results again in a non-extendable fully-hybridized duplex. However, it is not clear to me if this problem is primarily due to the consideration of simplified scenarios where there are, e.g., only primers, templates, and nucleotides and could be circumvented by using mixtures of strands of different lengths.

8.2. Coupling to ribozyme activity

An intriguing open question is how a system of informational polymers undergoing the basic reactions described in this thesis can couple to ribozymes that emerge spontaneously. One possibility discussed in the introduction was that the genome of a hypothetical protocell could favor the appearance of a ribozyme that catalysis membrane growth.

Alternatively, the spontaneous emergence of a cleaving ribozyme could be an interesting scenario to study. The so-called hammer head ribozyme I/III, see [20, 59, 60], is a catalyst that accelerates the cleavage of single strands with a minimum size of about 22 nucleotides. Its cleavage rate is sequence-dependent, but as it is one of the most studied ribozymes, sufficient data should be available such that its specific impact on the sequence space can be modeled at least quantitatively. The concentration of the ribozyme could be assumed to be proportional to the concentration of the sequences constituting its reactive core of 13 nucleotides[1]. The introduced sequence-dependent cleavage of single strands will likely have

[1] not counting nucleotides that must be present in the substrate strand of the I/III configuration

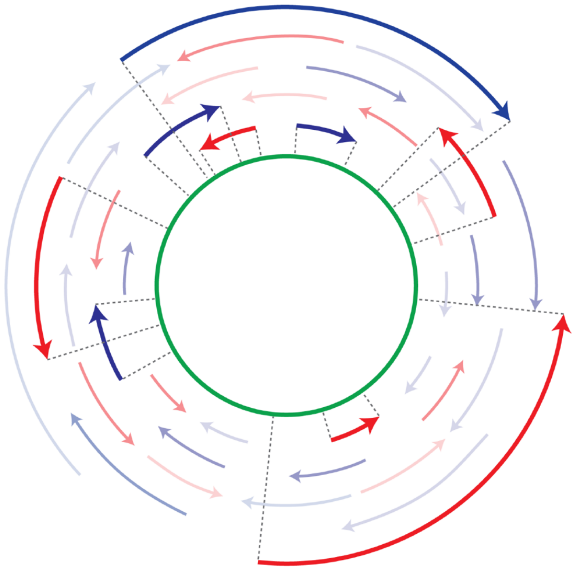


Figure 8.1.: The virtual circular genome G, green cycle in the middle, is not assumed to be present in the system. Instead its sub-sequences are assumed to be present. Hence a replication of the cyclic genome can be achieved by the growth in concentration of each of its sub-sequences. The image was taken from [107].

an impact on the hetero-catalytic cycles.

8.3. Virtual circular genome

Recently Zhou,Ding, and Szostak [107] suggested the concept of a virtual circular genome. In this section, I will briefly outline how one could study the suggested mechanism with the help of the simulation framework presented in this thesis. For this purpose, we first need to define the required quantities: We denote the sequence of the cyclic genome by G, see green circle in Fig. 8.1, and denote the set of all sub-sequences by $\mathcal{S}_G$. The number of different sub-sequences is then given by $M = |\mathcal{S}_G|$. We further denote a sub-sequence species as g_i, $i \in \{1, 2, .., M\}$ and the frequency of g_i in G by x_i. A sub-sequence has a length L_i which is smaller or equal to the length of the genome L_G. We further define the number of sub-sequence species of length L as $Q_L := \sum_i = 1^M \delta_{L_i,L}$. We denote the set of species being present in the system as C by $\mathcal{S}_C$ (c = cell) and the copy number of species s_i in C by n_i.

The actual circular genome G is not present in the system C, instead (almost) all of its sub-sequences g_i are assumed to be contained. The contained sub-sequences must constitute a full coverage of G, such that $\bigcup_{g_i \in C} g_i = G$.

Further, we require a decaying strand length distribution in the interval $[1, L_G]$ such that $N_{L+1} = \alpha N_L$, $\alpha < 1$, where N_L is the frequency of strands of length L. The authors conclude that: "A surprising consequence of such a concentration versus length gradient is that average oligonucleotide growth by as little as one nucleotide could result in replication of the entire genomic ensemble"[107].

Our simulation offers a suitable tool to study the suggested mechanism theoretically.

I suggest starting with a randomly generated circular genome of length $L_G = 30$ with an alphabet of 4 nucleotides. To provide presumably optimal conditions for the replication of the genome, we initialize the system only with its (single stranded) sub-sequences g_i weighted by their occurrence in the genome (this constraint can be relaxed in further

experiments). We thereby restrict the system to initial sub-sequences of a maximal length of $L_s = 10$.

We set $N_1 = 1.5 \times 10^4$ monomers corresponding to a concentration of $c_1 = 1\,\mathrm{M}$ and choose $\alpha = 1/2$, which yields an initial gradient of $N_L = e^{-\ln(2)(L-1)} N_1$, for $L \leq L_s$ and $N_L = 0$ for $L > L_s$. The initial total number of strands for each length is consequently given by

L	1	2	3	4	5	6	7	8	9	10
N_L	15000	7500	3750	1876	938	469	235	118	59	30

Note that the quantities were chosen such that the smallest copy number $N_{10} = 30$ does not fall below the maximal number of sub-sequences n_{10}, which is for all lengths limited by $L_G = 30$. Therefore all sub-sequence species are included initially in the system and the copy numbers of a sub-sequence can simply be calculated by[2]

$$n_i = \frac{x_i}{Q_{L_i}} N_{L_i}. \tag{8.1}$$

We further chemostat the monomer, dimer and trimer concentration. and apply a constant outflux rate, see Chapter 6. The outflux rate must be tuned s.t. a steady state can be reached in a reasonable amount of time.

For the first conceptual study, I would suggest an energy model similar to the one presented in the last chapter that assigns energies to blocks of four nucleotides based on the number of matches and mismatches. Also, a further simplified model could be considered, where hybridization is only allowed for perfectly matching sequences.

Regarding a cutoff in the dehybridization rate, both the bound and the unbound model are interesting scenarios to consider for this experiment.

Now the remaining question is, what are valuable observables for quantifying the strength of the replication reaction network associated with the cyclic genome? A simple observable could be the fraction of strands which are part of the cyclic genome in dependence of the strand length,

$$f_G(L) = \frac{\rho(L \,|\, \text{strand in } G)}{\rho(L)}. \tag{8.2}$$

Initially we have $f_G = 1$, for $L < L_s$ and $f_G = 0$ for $L > L_S$, as all strands are initially part of the genome. But in the course of time, new strand sequences will be created. If the system produces completely random strands, then $f_G(L)$ should decay towards $p = 1/4^L$.

Another interesting variable would be the number of cyclic genomes that can be assembled by alignment of the strands contained in the system at a certain time point, N_G.

We therefore define the entirety of strands present in the system at time t that are sub-sequences of the genome G as $\mathcal{Z}(t)$.

We would then uniformly draw a strand from $\mathcal{Z}(t)$ and subsequently sample all alignments where it matches the genome and covers a nucleotide that was not covered through a previous attempt. When the whole genome is covered, the copy number of the genome N_G is incremented. This repeats until there are no more strands left to draw from.

[2] A small python script that can be used for this calculation can be found on my github account [link].

Appendices

A. Supplemental material:

A.1. Hybridization

The number of possible hybridizations two complexes can undergo is denoted as Θ. We do not allow for rejection if a hybridization between the two colliding complexes is possible ($\Theta > 0$), and therefore set

$$p_{\mathrm{a}} = \begin{cases} 0, & \Theta = 0 \\ 1, & \Theta > 0. \end{cases} \tag{A.1}$$

This corresponds to setting the penalty for the formation of the first base pairing to zero, $\nu = -\ln(p_a) = 0$. We equally weight the different hybridizations upon collision

$$p_c = \frac{1}{\Theta}. \tag{A.2}$$

We call $1/\Theta$ the channel factor. In the main text p_a and p_c got combined into the hybridization probability $p_{\mathrm{hyb}} = p_a p_c$. Consequently the hybridization rate and the corresponding rate constant are given by

$$r_{\mathrm{on}} = \frac{1}{\Theta}\frac{1}{VN_A c^\circ}\frac{1}{t_0} \text{ and } k_{\mathrm{on}} = 2^{m_2-\delta_{ij}}\frac{1}{\Theta}\frac{1}{c^\circ t_0}. \tag{A.3}$$

Note that in the simulation the concentration of a reference species c_{ref} in c° and its corresponding copy number N_{ref} are input parameters.

The dehybridization rate and rate constant are given by

$$r_{\mathrm{off}} = \frac{1}{\Theta}e^{\gamma l}\frac{1}{t_0} \text{ and } k_{\mathrm{off}} = \frac{1}{\Theta}2^{m_1(1-\delta_{ij})}e^{\gamma l}\frac{1}{t_0}. \tag{A.4}$$

A.2. Scaling of the kinetic parameters of a stationary system

Consider the reaction fluxes ϕ in terms of the concentration vector $\vec{c}$ and the rate constants:

$$\begin{aligned}
\phi_{\mathrm{on}}(\vec{c}, k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{lig}}, k_{\mathrm{out}}) &\propto c_i c_j k_{\mathrm{on}}, \\
\phi_{\mathrm{off}}(\vec{c}, k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{lig}}, k_{\mathrm{out}}) &\propto c_k k_{\mathrm{off}}, \\
\phi_{\mathrm{lig}}(\vec{c}, k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{lig}}, k_{\mathrm{out}}) &\propto c_k k_{\mathrm{lig}}, \\
\phi_{\mathrm{out}}(\vec{c}, k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{lig}}, k_{\mathrm{out}}) &\propto c_k k_{\mathrm{out}}.
\end{aligned}$$

A transformation of the form

$$\vec{c} \to \vec{c}' = \alpha_c \vec{c}, \quad k_{\mathrm{on}} \to k'_{\mathrm{on}} = \alpha_{\mathrm{on}} k_{\mathrm{on}}, \quad k_{\mathrm{off}} \to k'_{\mathrm{off}} = \alpha_{\mathrm{off}} k_{\mathrm{off}}, \quad k_{\mathrm{lig}} \to k'_{\mathrm{lig}} = \alpha_{\mathrm{lig}} k_{\mathrm{lig}}, \quad k_{\mathrm{out}} \to k'_{\mathrm{out}} = \alpha_{\mathrm{out}} k_{\mathrm{out}}$$

that scales all reaction fluxes by the same amount changes only the intrinsic time-scale of the dynamics. In particular, it leaves the stationary distributions invariant.

By forming the three independent ratios of the reaction fluxes, ond finds that all transfor-

mations that lead to the same ratios

$$\alpha_1 := \frac{\alpha_{\text{off}}}{\alpha_{\text{lig}}}, \qquad \alpha_2 := \frac{\alpha_{\text{off}}}{\alpha_{\text{out}}}, \qquad \alpha_3 := \frac{\alpha_{\text{off}}}{\alpha_{\text{on}}\alpha_{\text{c}}},$$

are equivalent with respect to the stationary distribution. In particular, it shows that scaling the hybridization rate has the same effect as scaling the concentration.

A.3. Initiation penalty

Throughout this article we set the acceptance probability constant to $p_a = 1$, and thus the initiation penalty to $\nu = -\ln(p_a) = 0$. We did not vary this parameter explicitly. In this section, we apply the scaling relations derived in the previous section, in order to show that a variation of the initiation penalty is equivalent to a variation in concentration. In order to compare the thereby obtained values of the initiation penalties to the values given in the literature, we provide formulas for the conversion between different energy units in the next section.

A.3.1. Conversion between units of free energy

For clarity, in this section we denote the value of a physical quantity x by curly brackets and its unit by squared brackets, $x = \{x\}\,[x]$. For example, the Avogadro constant is written as $N_A = \{N_A\}[N_A]$, where $\{N_A\} = 6.022 \times 10^{23}$ and $[N_A] = \text{mol}^{-1}$. This notation will be useful in this section to give simple expressions for converting different energy units.

Free energies measured in experiments are usually given in units of kcal/mol, $\Delta G^m = \{\Delta G^m\}\,\text{kcal/mol}$, which can be converted to the free energy of a single molecule in Joule via $\Delta G^s = \frac{4184\{\Delta G^m\}}{\{N_A\}}\text{J}$. Dividing by the Boltzmann constant leads to

$$\frac{\Delta G^s}{k_{\text{B}}} = \frac{4184\,\{\Delta G^m\}}{\{R\}}\text{K} = 503.22\,\{\Delta G^m\}\,\text{K}, \tag{A.5}$$

where $R = 8.31\,\text{J}\,\text{K}^{-1}\,\text{mol}^{-1}$ is the gas constant. Thus, standard free energies measured at $T^\circ = (273.15 + 37)\,\text{K}$ can be converted to their corresponding values in $k_{\text{B}}t_0$ via

$$\frac{\Delta G^s}{k_{\text{B}}t_0} = 1.62\,\{\Delta G^m\}\,. \tag{A.6}$$

For example the standard binding energy per nucleotide $\gamma = -0.5$ would correspond to $-0.3\,\text{kcal/mol}$.

A.3.2. Scaling of concentrations is equal to a variation of the initiation penalty

In Sec. 4.8.1 we set the acceptance probability to $p_a = 1$ and consequently the initiation penalty to $\nu = -\ln(p_a) = 0$. Introducing an acceptance probability $p_a < 1$ into our model corresponds to a scaling of the hybridization rate with the factor $\alpha_{\text{on}} = p_a$. As shown in Sec. A.2, this is in turn the same as scaling the concentration with $\alpha_{\text{c}} = p_a$. We can thus map the concentration sweep (*cf.* Fig. 6.9(d)) onto an initiation penalty sweep. For this, we use the system with the highest simulated dimer concentration $c_2^* = 5.5\times10^{-3}\text{Mm}$ as a reference system. A concentration can then be mapped onto the corresponding initiation penalty via $\nu = -\ln(\alpha_c)$, where $\alpha_c = c_2/c_2^*$. For $c_2^* = 5.5\text{Mm}$ being the reference concentration we have

c_2 (mM)	5.5	3.5	2.5	1.5	1	0.5	0.2
α_c	1	0.64	0.45	0.27	0.18	0.09	0.04
ν	0	0.45	0.79	1.30	1.70	2.40	3.31

Measured values for the free energies of the initiation penalty for RNA under standard conditions at $T = 37\,°\mathrm{C}$ are $\Delta G_{\nu,\mathrm{RNA}}(37\,°\mathrm{C}) = 4.09\,\mathrm{kcal/mol}$, $\Delta H_{\nu,\mathrm{RNA}} = 3.61\,\mathrm{kcal/mol}$ [52]. The associated entropy is therefore $\Delta S_{\nu,\mathrm{RNA}} = -1.5\times10^{-3}\,\mathrm{kcal/mol}$.

For DNA in a magnesium (Mg^{2+}) rich buffer at $T = 25°\mathrm{C}$ measured values for the free energies and entropy are $\Delta G_{\nu,\mathrm{DNA}}(25\,°\mathrm{C}) = 0.91\,\mathrm{kcal/mol}$, $\Delta H_{\nu,\mathrm{DNA}} = 2.77\,\mathrm{kcal/mol}$ and $\Delta S_{\nu,\mathrm{DNA}} = 6.2\times10^{-3}\,\mathrm{kcal/mol}$ [83]. Hence at $T = 37\,°\mathrm{C}$ this corresponds to a Gibbs free energy of $\Delta G_{\nu,\mathrm{DNA}}(37\,°\mathrm{C}) = 0.847\,07\,\mathrm{kcal/mol}$.

Hence in $k_\mathrm{B}T^\circ$, with $T^\circ = 310.15\,°\mathrm{C}$, the Gibbs free energies of the initiation penalty for a single molecule are (see Sec. A.3.1):

$$\nu_\mathrm{RNA} = \frac{\Delta G^s_{\nu,\mathrm{RNA}}(37)}{k_\mathrm{B}t_0} = 6.64,$$
$$\nu_\mathrm{DNA} = \frac{\Delta G^s_{\nu,\mathrm{DNA}}(37)}{k_\mathrm{B}t_0} = 1.37.$$

Assuming that the standard binding energy per nucleotide $\gamma = -0.5$ corresponds to $T = T^\circ$, the measured value of the initiation penalty for DNA, $\nu_\mathrm{DNA} = 1.37$, lies within the simulated values, cf. Table **??**.

A.4. Smoothing filter

In this section, we describe the smoothing filter that we used in our data analysis. We applied it onto length distributions $\rho(L)$ in steady state to smooth the long noisy tail of the distribution.

The filter is a moving average that takes data points (x_i, y_i), $i \in \{1, ..., N\}$, as an input and generates a smoothed set of data points (a_i, b_i), $i \in \{1, ..., M\}$, $M \leq N$. The principle of the algorithm is illustrated in Fig. A.1(a). The size of the window for the moving average is adaptively increasing with the value of x_i. We define a vector that sets the boundaries at which the window size changes $v = (100, 500, 1000, 5000, 10000, ...)$. If $x_i < v(0)$ no smoothing is applied and $(a_i, b_i) = (x_i, y_i)$. Each interval given by $[v(j), v(j+1)]$ is divided into $m_j = 10(v(j+1)/v(j) - 1)$ sub-intervals s, $s \in \{1, ..., m_j\}$, of length $\Delta L_j = v(j)/10$. For each of these sub-intervals a value a_s is calculated, by assigning it the lower limit of each sub-interval (*cf.* dashed orange line in Fig. A.1(a))

$$a_s = v(j) + (s-1)\Delta L_j. \tag{A.7}$$

The corresponding value b_s is calculated by taking the average over all y_i, for which the corresponding x_i is not further away than $\Delta L_j/2$ from a_s:

$$b_s = \frac{1}{|I_s|}\sum_{i\in I_s} y_i, \text{ where } I_s = \left\{i : x_i \in \left[a_s - \Delta L_j/2, a_s + \Delta L_j/2\right]\right\}. \tag{A.8}$$

Let us consider two examples: (1) Consider $j = 0$, hence $v(0) = 100$ and $v(1) = 500$ and thus $\Delta L_0 = 10$. We calculate $a_2 = 100 + 10 = 110$. Thus, b_2 is the average over all y_i, where $105 \leq x_i < 115$. In a system that is coupled to a reservoir containing only dimers, only

strands of even length are present, i.e. $x_i \in x_i \in \{106, 108, 110, 112, 114\}$. Hence b_2 is an average over 5 data points.

(2) Consider $j = 2$, hence $v(2) = 1000$ and $v(3) = 5000$, thus $\Delta L_2 = 100$. We calculate $a_2 = 1000 + 100 = 1100$. Thus, b_2 is the average over all y_i, where $1050 < x_i < 1150$.

The filter's effect on the length distribution is shown in Fig. A.1(b). Only the noise in the tail is smoothed, the shape of the length distribution is preserved.

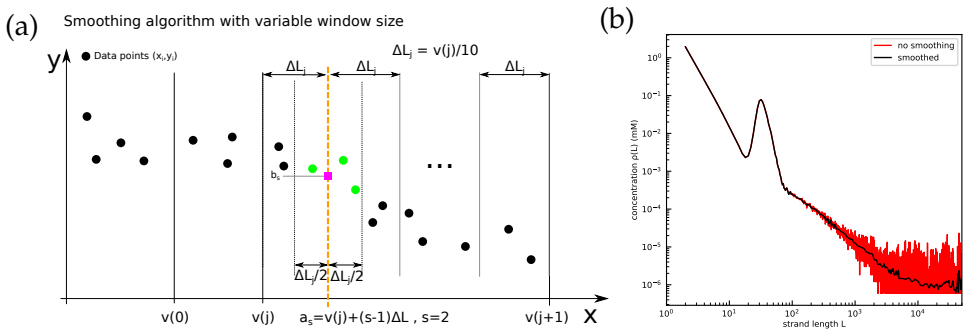


Figure A.1.: Illustration of the smoothing algorithm: (a) The pink square labeled b_s is the average values over all y_j where the corresponding $x_j \in [a_s - \Delta L_j/2, a_s + \Delta L_j/2]$. The case illustrated corresponds to $s = 2$. (b) Application of smoothing filter on the length distribution (dimer only reservoir, standard parameters). The red curve is the original curve, and the black curve the resulting smoothed length distribution. The effect of smoothing is only visible in the tail and makes it easier to visualize the trend.

A.5. Trajectories of extension cascades

For further investigation of the assembly and growth processes in our model, we analyzed the trajectories of duplexes that undergo extension cascades resulting in a fully-hybridized duplex.

A.5.1. Sampling of trajectories

We start with a background obtained form a simulation that reached steady state. We set the complexes as background species, hence do not allow for reactions within this background, as described in Sec. 5.1.4. As mentioned in Sec. 5.1.1, internally, the simulation implements complexes via segments, all of which have a poly(A) sequence. We insert a dimer with a specific sequence ("TB") into the system, which serves as a label for an individual tracer complex. The dimer can undergo hybridizations with all background species. Thereby the label becomes integrated into other complexes. The complex with the specific label is always kept as the only non-background species in the system (it undergoes reactions with the background). We call this specific complex the tracked complex. After each reaction we check if the complex contains a stable sub-duplex (see Fig. A.2(a)).

If one of the sub-duplexes starts to undergo an extension cascade (the dehybridization rate of its hybridization site is smaller than its extension rate), we start to track the sub-duplex within the tracked complex. We store the initial stable sub-duplex in a buffer. Whenever the stable sub-duplex is extended via templated ligation, the newly formed duplex is appended to the buffer. The stored sequential snapshots of the extensions of the stable sub-duplex is what we call a trajectory. If the duplex dehybridizes, the trajectory is deleted, and the recording restarts as soon as a new stable duplex that undergoes extension cascades is

formed. When the tracked complex leaves the system via outflux, we save the trajectory to disk and the assembly process restarts with a labeled dimer. The schema is illustrated in Fig. A.3. As mentioned above, the tracked complex can undergo hybridizations with the background. There is, however, one caveat: The tracked complex is only allowed to undergo hybridizations with background complexes that do not itself contain sub-duplex that undergo extension cascades. We further reject trajectories where two duplexes that can undergo extension cascades are formed within a complex. These restrictions guarantee the sampling of trajectories that start with the onset of a extension cascade and finish in a fully-hybridized configuration. For the rejected trajectories it would not be possible to identify a unique starting point. It can be expected that the thereby obtained strand length distribution tends to underestimate the length distribution obtained via the full simulation.

A comparison between the strand length distribution obtained via the full simulation and the sampled trajectories reveals that the latter resembles the first one, *cf.* Fig. A.2(b). Hence the sampling is consistent with the dynamics. But indeed, the length distribution obtained via sampling trajectories underestimates the concentration of strands of length $L \geq 50 \geq L_{\max}$. For $L = 40$, the relative deviation is only -3%, whereas for $L = 50$ the deviation is -19%.

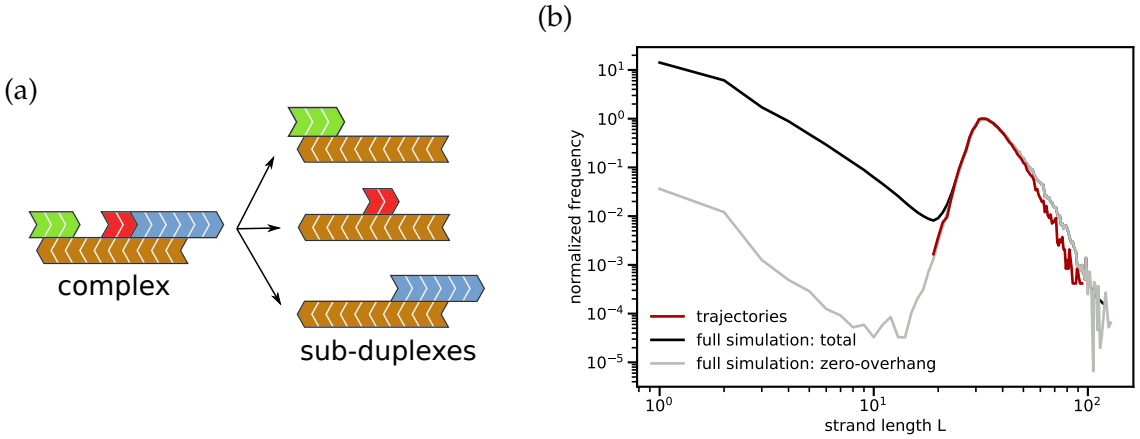


Figure A.2.: (a) The tracked complex of order n ($n-1$ hybridization-sites) can be decomposed into $n-1$ sub-duplexes. (b) The length distribution of fully-hybridized strands obtained from the trajectories (red curve) resembles the length distribution of the fully-hybridized strands obtained from the full simulation (gray curve). The length distribution for all strand lengths (black line) is plotted for orientation. The system shown here is the monomer-dimer system with a monomer fraction $f_{\mathrm{m}} = 0.7$. The length distributions are normalized on the concentration at the maximum $L_{\max} = 33$. Only at the tail the length distribution obtained via trajectory sampling underestimates the concentration of long strands. This behavior is expected since certain trajectories leading to long fully hybridized complexes are rejected.

A.5.2. Analysis of trajectories

We denote the total number of sampled trajectories as Ω and denote a single trajectory by ω. A trajectory contains the assembly information starting from a stable sub-duplex (as explained in the last section) of length C_{initial} until it reaches a fully-hybridized duplex of length C_{final} (*cf.* Fig. A.4(a)). A trajectory ω can be divided into the covering of $P(\omega)$ copy sites j of length $l_{\mathrm{cs}}(\omega, j)$ (*cf.* Fig. A.4(b)), which we call a partial trajectory. In the following we will use the notion of *template* and *primer* introduced in Sec. 6.4.7 of the main text. Each copy site can again be resolved into building blocks i of length $B(\omega, j, i)$ used to built the

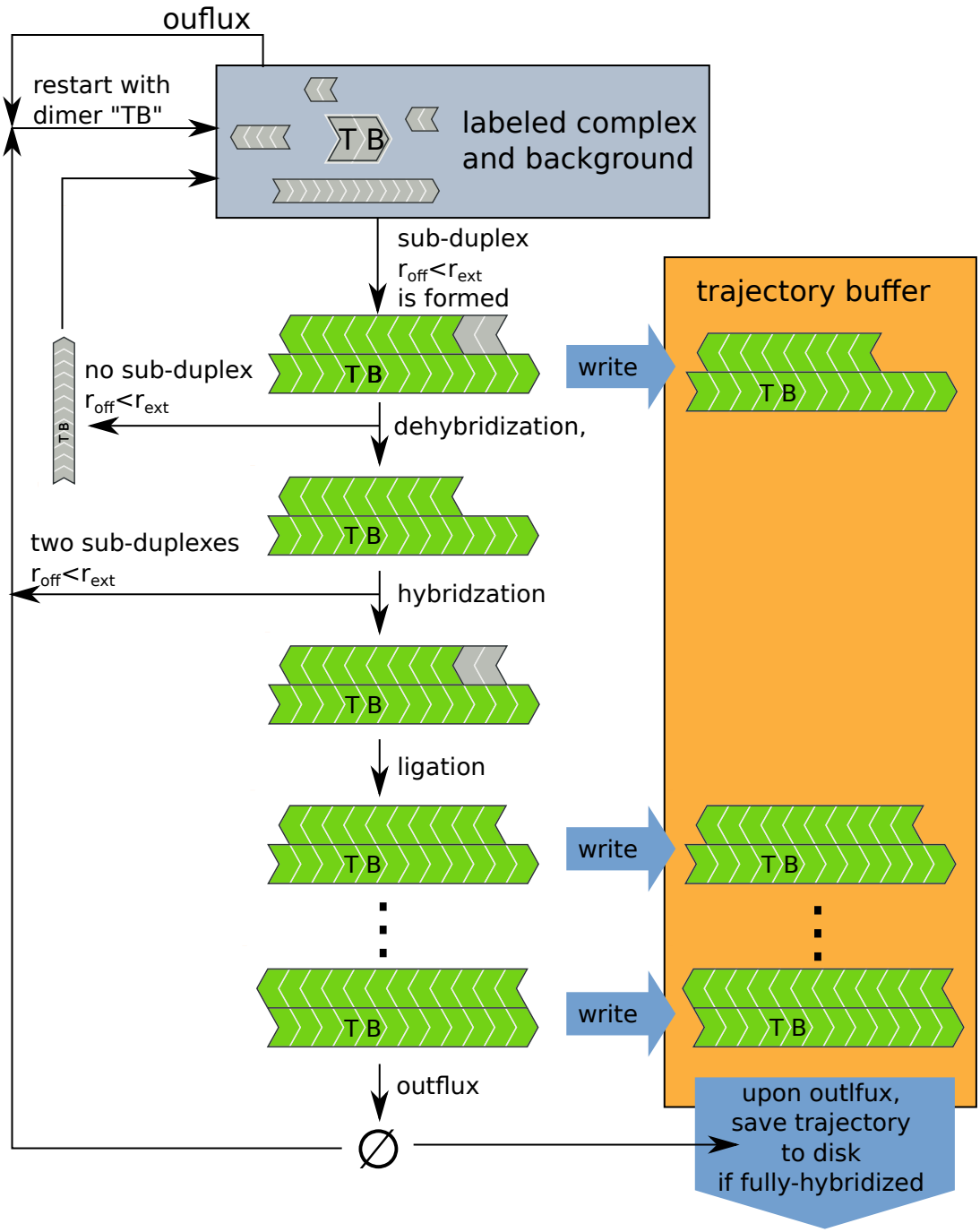


Figure A.3.: The simulation loads the complexes from a simulation that reached a steady state and utilizes them as background species. A labeled dimer ("TB") is inserted into the reaction vessel (simulation). The complex containing the ("TB") motif is called the tracked complex. It can undergo dehybridizations and ligations, and collisions with the background. As soon as the tracked complex contains a stable sub-duplex, the trajectory of the sub-duplex is recorded. Whenever a ligation happens, the sub-duplex structure is written to a buffer. Hybridizations and dehybridizations are not explicitly tracked, i.e. the resulting new complex structure is not written to the buffer. However, if the tracked stable duplex disassembles via dehybridization, the trajectory is rejected and the buffer is cleared. We again start saving a trajectory as soon as the complex containing the "TB" motif contains a stable sub-duplex. We also neglect trajectories that include two stable duplexes at some point, and in this case, restart with a "TB" dimer. If the complex leaves the reaction vessel via outflux and if it has reached a fully hybridized configuration, we save the trajectory to disk. The buffer gets cleared and we restart with the dimer motif.

partial trajectory (*cf.* Fig. A.4(c)). A building block is incorporated into the primer template duplex via one of the following growth processes (*cf.* Fig. A.4(b)):

(p) primer extension: The building block extends the primer and is shorter than the remaining overhang.

(s) switching: The building block extends the primer but is longer than the remaining overhang resulting in a new copy site.

(t) template extension: The building block extends the template.

The length that is covered by the incorporation of a building block is called $B^{\text{eff}}(\omega, j, i)$. In the main text, we referred to this quantity as the effective building block length. If the building block is incorporated via primer extension its effective building block length is equal to its regular length, $B^{\text{eff}}(\omega, j, i) = B(\omega, j, i)$. In contrast, if it is incorporated via template extension the effective building block length is $B^{\text{eff}}(\omega, j, i) = 0$. In the case of switching $B^{\text{eff}}(\omega, j, i)$ is the length of the last overhang of the copy site, or equivalently, the difference between the building block length and the length of the new copy site. The assembly over the course of time of a copy site is illustrated in Fig. A.4(d).

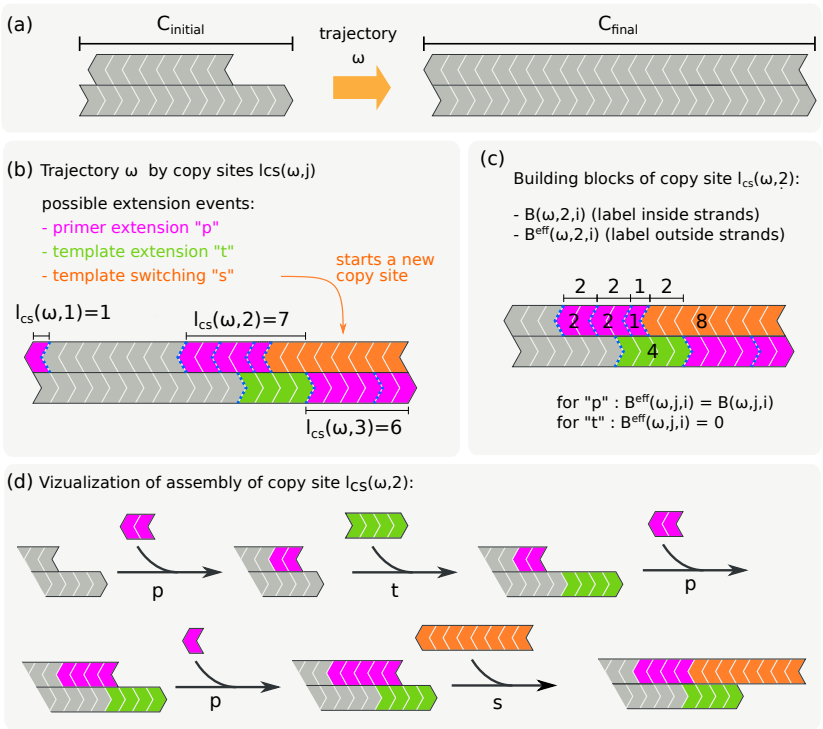


Figure A.4.: (a) A trajectory ω contains the assembly information of a fully-hybridized duplex of length C_{final}, starting from a stable sub-duplex of length C_{initial}. (b) A trajectory can be split into several copy sites of length $l_{\text{cs}}(\omega, j)$. The assembly of a copy site is called a partial trajectory. A switching event creates a new copy site and reverses the role of primer and template. If the copy site becomes fully covered and no switching happens, it results in a blunt end. A duplex with two copy sites reaching a blunt end is a fully-hybridized duplex. (c) The length of the building blocks $B(\omega, j, i)$ and effective building blocks $B^{\text{eff}}(\omega, j, i)$ that assemble copy site $j = 2$. (d) Assembly of copy site $j = 2$. In the last step, the switching events creates a new copy site.

A.6. Laboratory experiment 12 nt random A-T-DNA strands

A.6.1. Initial sequence space of 12 nt A-T strand

The initial 12mer AT random DNA pool was ordered as 5′-WWWWWWWWWWWW-3′ with 5′- phosphate modification from biomers.net. The randomness of this initial pool can only be accessed after next generation sequencing, which might introduce a bias due to several enzyme-driven PCR- and ligaseextension reaction necessary for the attachment of the primers and barcodes to the strands. Sequencing results show an overall bias towards A-rich sequences and a lack of poly-T sequences. Fig. A.5(a) shows all 4096 (=212) possible 12mer binary sequences. Fig. A.5(c) shows the abundance of 6 nt long subsequences (26=64 possible sequences). Motives with poly-T are rare in comparison to other strands. Motives with poly-A are overrepresented. Overall, there are 4067 of the 4096 possible submotives.

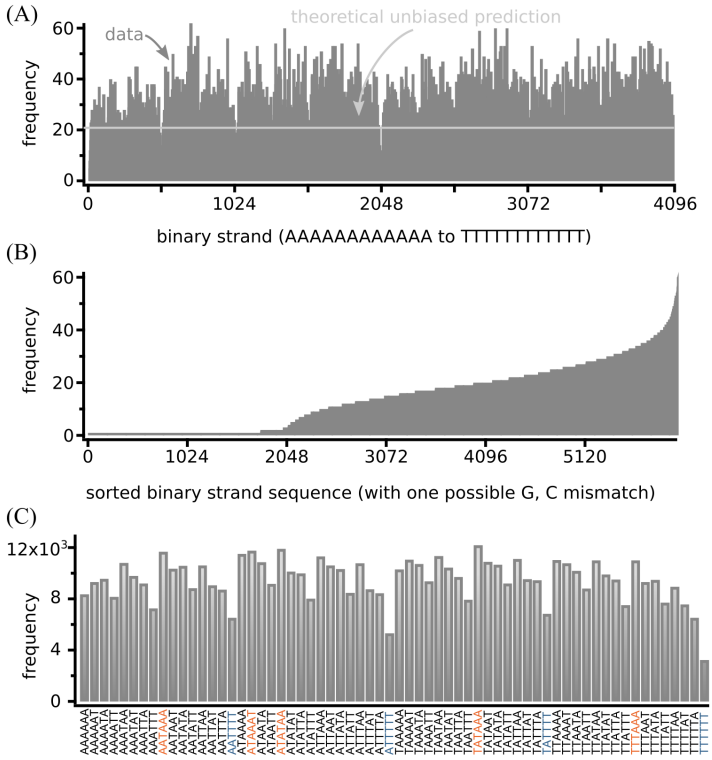


Figure A.5.: AT-only random sequence 12mer DNA pool: (a) Frequency of all possible 4096 sequence motives sorted in a binary way (0:A, 1:T). (b) Abundance of sequences, including single misreads (G, C in AT-only DNA). (c) Frequency of 6 nt (26=64) submotives in 12mer "monomer" strands. Poly-T motives are underrepresented while poly-A motives are overrepresented.

A.6.2. Resulting PAGE gels

Fig. A.5 shows the frequency of sampled 12mer strands with maximal one mismatched base. In AT-only DNA reads for G and C are, by definition, false reads. The long tail between binary strand sequence 0 and about 1800 is made almost entirely of single mismatch reads. The assumption of a random sequence pool with a slight bias towards poly-A sequences is a valid starting point for the experiment. In direct sequence analysis, the bias might influence the result if the system selects for specific sequences.

The numerical model assumes a constant ligation temperature T_{cold} for the entire reaction time until a steady-state is reached. In the experiment, a constant T_{cold} does not produce

multimer products over 60 hours, see Fig. A.6(a) and (b). Therefore, we assume that the binding energy is too large to repeatedly build complexes, ligate, and dissociate in the experimental duration of 60 h. The dynamics are essentially frozen. Temperature cycling is an easy way to "reset" dsDNA to their ssDNA state to promote further hybridization and ligation reactions.

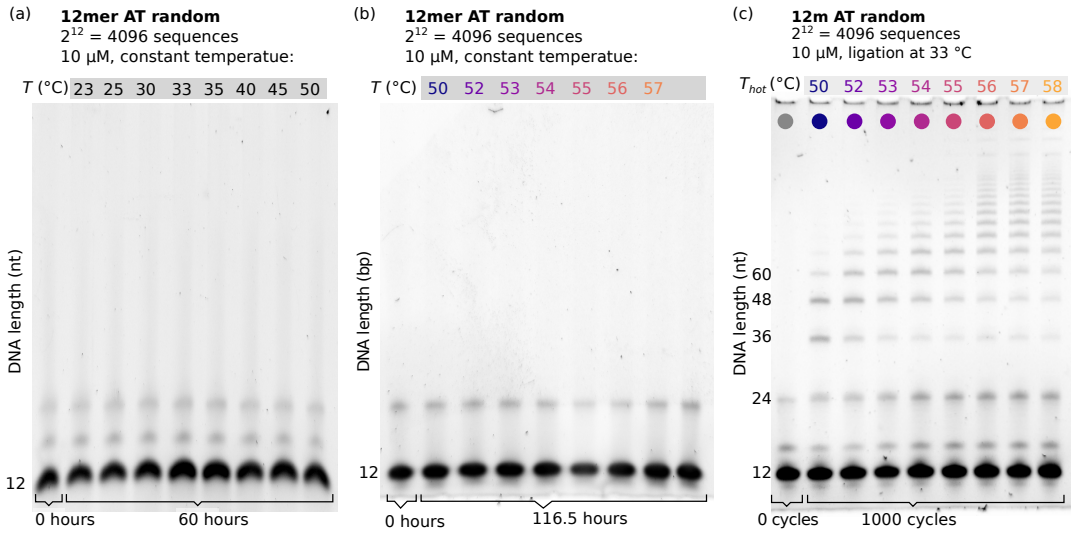


Figure A.6.: PAGE gels for experiments at constant temperature (a) and (b) and experiments exposed to 1000 temperature cycles. AT-only random sequence 12mer DNA does not show signs of multimer products from templated ligation during incubation at constant temperatures. In contrast, the experiments for temperature cycles show significant multimer production. In this case, the pattern of the PAGE gel change with temperatures.

A.6(c) again shows the PAGE gel from Fig. 6.25 in the main manuscript with the results for 1000 temperature cycles between T_{cold} and varied T_{hot}, which takes about 40 to 60 hours depending on the temperatures and the temperature ramp of the PCR thermos cycler device.

A.6.3. Concentration quantification on PAGE gels by Image Analysis

We used a custom LabView program for concentration quantification of bands on PAGE gels, as shown before [103]. The method has limitations, as described in depth below, but for the samples used in this study, the method is reliable, and results are reproducible.

First, each lane is marked with a top and a bottom cursor that span a rectangular ROI (region of interest, about 10-30 % of the lane width) on the lane, as shown in Fig. A.7(a). The intensity is read as mean intensity values averaged over the width of the ROI. The center region of each lane is the lowest lateral intensity change in the band and is therefore ideal for selecting ROIs to compare different lanes. Additionally, the areas in between lanes are also selected with separate ROIs (red). The inter-lane ROIs characterize the gel background and a possible inhomogeneous illumination. For each lane, the average background calculated from the left and the right inter-lane ROIs is calculated and subtracted to get the band intensity only, as shown in Fig. A.7 (b).

To finally quantify each band, the intensity of each lane is normalized to the reference lane. This step includes some of the limitations of this analysis:

- The total intensity per lane is homogeneous for each lane:
  - The total amount of DNA in the reference sample and in each lane is the same, just in different length distributions. We assume there is a similar chance for all DNA to be stained by SYBR gold, resulting in the same total intensity per lane.

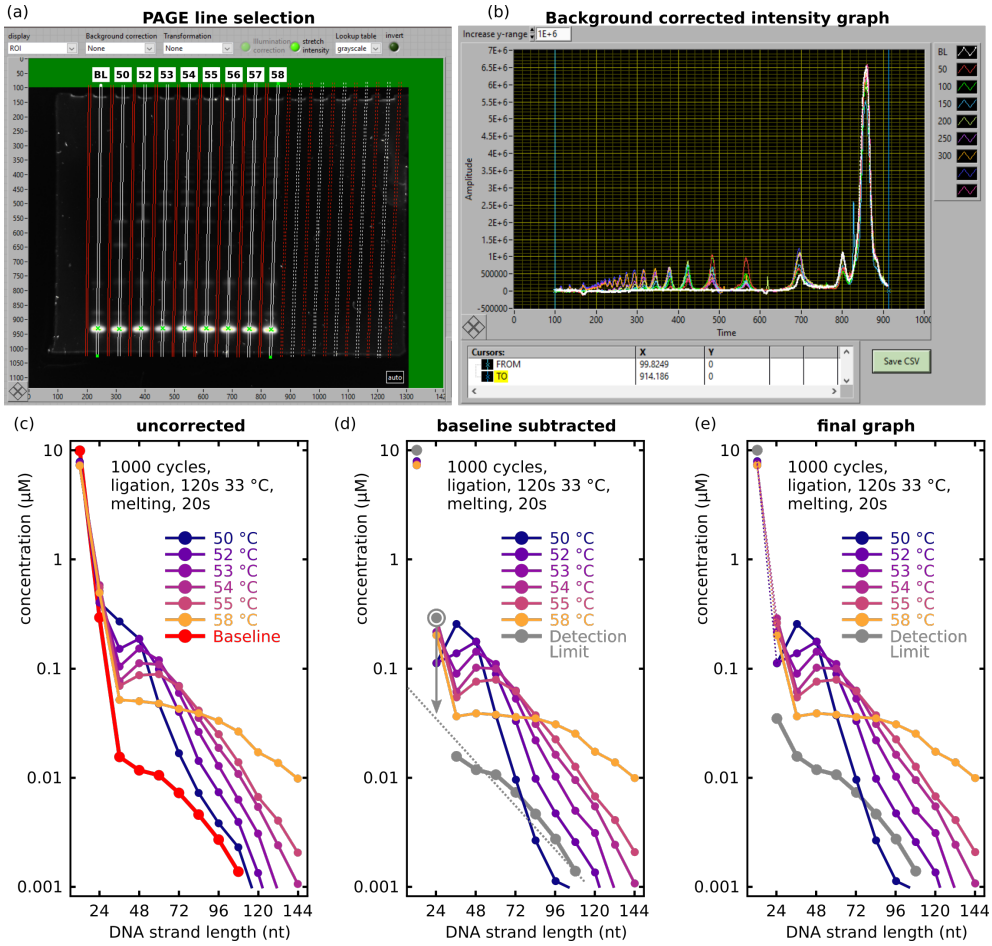


Figure A.7.: Concentration quantification workflow:
(a) Selection of the lanes on the gel image. (b) Background corrected (red lines in a) intensity over position graph. (c) Concentration estimation from the peak-areas in b. (d) Corrected concentration due to subtraction of the baseline signal, which is then called the detection limit. (e) Final graph, the position of the 24mer is extrapolated from the normal-log plot in d.

  - The increase in intensity at similar concentrations for longer bands is due to the increase in length. With the item above, an increase in length is similar to a linearly increasing probability of SYBR staining. This is not true for non-denaturing gels, dsDNA, and DNA with all four bases.
  - Differences in the total intensity of each lane are attributed to the pipetting error that occurs due to handling small volumes of fluid with high viscosity.

- There is a need for a reference sample of known length and concentration. Furthermore, the products need to be well defined as resulting from the monomers. This analysis is not suitable for pools with different illumination per length samples with similar product lengths.

- The concentration of DNA products in comparison to the monomers can only vary in the range of detection.

In the last step, the baseline is subtracted from all lanes to achieve the final concentrations. The band at a length of about 24 nt was identified as artifacts during the strand synthesis (see [103]). Therefore, the position of the 24mer baseline is extrapolated in a linear fashion from the normal-log plot, as shown in Fig. A.7(e).

Each lane is then accessible as an intensity over gel-position data structure. Due to the background correction described above, there is no tilting or region-specific shift in the baseline. The intensities in each lane are only slightly too high due to the lane being slightly lighter in the fluorescent image than the rest of the gels. Each band is then either fit by a Gaussian curve, or all data points in the region of the band, from baseline to baseline, are simply summed up. The concentration is then a function of band-peak intensity and the position in the lane (meaning the length of the reaction product).

Contents

List of Figures

List of Tables

Bibliography

[1] M. v. Smoluchowski. "Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen." In: *Zeitschrift für Physikalische Chemie* 92U.1 (Jan. 1918). ISSN: 2196-7156, 0942-9352. DOI: 10.1515/zpch-1918-9209.

[2] F. Crick. "The origin of the genetic code." In: *J. Mol. Biol.* 38.3 (Dec. 1968), pp. 367–379. ISSN: 00222836. DOI: 10.1016/0022-2836(68)90392-6.

[3] J. Sulston, R. Lohrmann, L. E. Orgel, and H. T. Miles. "Nonenzymatic synthesis of oligoadenylates on a polyuridylic acid template." In: *Proceedings of the National Academy of Sciences* 59.3 (Mar. 1968), pp. 726–733. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.59.3.726.

[4] J. J. Hopfield. "Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity." In: *Proceedings of the National Academy of Sciences* 71.10 (Oct. 1974), pp. 4135–4139. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.71.10.4135.

[5] D. T. Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions." In: *Journal of Computational Physics* 22.4 (Dec. 1976), pp. 403–434. ISSN: 00219991. DOI: 10.1016/0021-9991(76)90041-3.

[6] M. Eigen and P. Schuster. "A principle of natural self-organization: Part A: Emergence of the hypercycle." In: *Naturwissenschaften* 64.11 (Nov. 1977), pp. 541–565. ISSN: 0028-1042, 1432-1904. DOI: 10.1007/BF00450633.

[7] D. T. Gillespie. "Exact stochastic simulation of coupled chemical reactions." In: *The Journal of Physical Chemistry* 81.25 (Dec. 1977), pp. 2340–2361. ISSN: 0022-3654, 1541-5740. DOI: 10.1021/j100540a008.

[8] P. W. Anderson. "Suggested model for prebiotic evolution: the use of chaos." In: *Proceedings of the National Academy of Sciences* 80.11 (June 1983), pp. 3386–3390. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.80.11.3386.

[9] "Chapter 2 Diffusion-Controlled Reactions in Solution." In: *Comprehensive Chemical Kinetics*. Ed. by C. H. Bamford, C. F. H. Tipper, and R. G. Compton. Vol. 25. Diffusion-Limited Reactions. Elsevier, Jan. 1985, pp. 3–46. DOI: 10.1016/S0069-8040(08)70252-8.

[10] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. "Improved free-energy parameters for predictions of RNA duplex stability." In: *Proceedings of the National Academy of Sciences* 83.24 (Dec. 1986), pp. 9373–9377. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.83.24.9373.

[11] G. von Kiedrowski. "A Self-Replicating Hexadeoxynucleotide." In: *Angewandte Chemie International Edition in English* 25.10 (1986), pp. 932–935. ISSN: 0570-0833, 1521-3773. DOI: 10.1002/anie.198609322.

[12] A. Kanavarioti and D. H. White. "Kinetic analysis of the template effect in ribooligoguanylate elongation." In: *Origins of Life and Evolution of the Biosphere* 17.3-4 (Sept. 1987), pp. 333–349. ISSN: 0169-6149, 1573-0875. DOI: 10.1007/BF02386472.

[13] W. S. Zielinski and L. E. Orgel. "Autocatalytic synthesis of a tetranucleotide analogue." In: *Nature* 327.6120 (May 1987), pp. 346–347. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/327346a0.

[14] W. S. Zielinski and L. E. Orgel. "Oligoaminudeoside phosphoramidates. Oligomeilzation of dimers of 3'-amino-3'-deoxy-nucleotides (GC and CG) in aqueous solution." In: *Nucleic Acids Research* 15.4 (1987), pp. 1699–1715. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/15.4.1699.

[15] G. Wachtershauser. "An all-purine precursor of nucleic acids." In: *Proc. Natl. Acad. Sci. USA* 85.4 (Feb. 1988), pp. 1134–1135. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.85.4.1134.

[16] G. F. Joyce. "RNA evolution and the origins of life." In: *Nature* 338.6212 (Mar. 1989), pp. 217–224. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/338217a0.

[17] S. A. Kauffman. *The origins of order: self-organization and selection in evolution*. New York: Oxford University Press, 1993. ISBN: 978-0-19-505811-6.

[18] D. Sievers and G. von Kiedrowski. "Self-replication of complementary nucleotide-based oligomers." In: *Nature* 369.6477 (May 1994), pp. 221–224. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/369221a0.

[19] E. Szathmáry and J. M. Smith. "The major evolutionary transitions." In: *Nature* 374.6519 (Mar. 1995), pp. 227–232. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/374227a0.

[20] K. R. Birikh, P. A. Heaton, and F. Eckstein. "The Structure, Function and Application of the Hammerhead Ribozyme." en. In: *European Journal of Biochemistry* 245.1 (Apr. 1997), pp. 1–16. ISSN: 0014-2956, 1432-1033. DOI: 10.1111/j.1432-1033.1997.t01-3-00001.x.

[21] E. L. Hegg, K. A. Deal, L. L. Kiessling, and J. N. Burstyn. "Hydrolysis of Double-Stranded and Single-Stranded RNA in Hairpin Structures by the Copper(II) Macrocycle Cu([9]aneN $_3$)Cl $_2$." In: *Inorg. Chem.* 36.8 (Apr. 1997), pp. 1715–1718. ISSN: 0020-1669, 1520-510X. DOI: 10.1021/ic960955b.

[22] Y. Li and R. R. Breaker. "Kinetics of RNA Degradation by Specific Base Catalysis of Transesterification Involving the 2'-Hydroxyl Group." In: *J. Am. Chem. Soc.* 121.23 (June 1999), pp. 5364–5372. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja990592p.

[23] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." In: *Journal of Molecular Biology* 288.5 (May 1999), pp. 911–940. ISSN: 00222836. DOI: 10.1006/jmbi.1999.2700.

[24] M. A. Gibson and J. Bruck. "Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels." In: *The Journal of Physical Chemistry A* 104.9 (2000), pp. 1876–1889. DOI: 10.1021/jp993732q.

[25] G. Varani and W. H. McClain. "The G·U wobble base pair: A fundamental building block of RNA structure crucial to RNA function in diverse biological systems." en. In: *EMBO reports* 1.1 (July 2000), pp. 18–23. ISSN: 1469-221X, 1469-3178. DOI: 10.1093/embo-reports/kvd001.

[26] S. Howorka, L. Movileanu, O. Braha, and H. Bayley. "Kinetics of duplex formation for individual DNA strands within a single protein nanopore." In: *Proceedings of the National Academy of Sciences* 98.23 (Nov. 2001), pp. 12996–13001. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.231434698.

[27] S. Redner. *A Guide to First-Passage Processes*. Cambridge: Cambridge University Press, 2001. ISBN: 978-0-521-65248-3. DOI: 10.1017/CBO9780511606014.

[28] N. C. Stellwagen, S. Magnusdottir, C. Gelfi, and P. G. Righetti. "Measuring the translational diffusion coefficients of small DNA molecules by capillary electrophoresis." In: *Biopolymers* 58.4 (2001), pp. 390–397.

[29] J. S. Reader and G. F. Joyce. "A ribozyme composed of only two different nucleotides." In: *Nature* 420.6917 (2002), pp. 841–844.

[30] J. L. Mergny and L. Lacroix. "Analysis of Thermal Melting Curves." In: *Oligonucleotides* 13.6 (Dec. 2003), pp. 515–537. ISSN: 1545-4576, 1557-8526. DOI: 10.1089/154545703322860825.

[31] L. E. Orgel. "Prebiotic Chemistry and the Origin of the RNA World." In: *Critical Reviews in Biochemistry and Molecular Biology* 39.2 (Jan. 2004), pp. 99–123. ISSN: 1040-9238, 1549-7798. DOI: 10.1080/10409230490460765.

[32] J. SantaLucia and D. Hicks. "The Thermodynamics of DNA Structural Motifs." In: *Annual Review of Biophysics and Biomolecular Structure* 33.1 (2004), pp. 415–440. ISSN: 1056-8700, 1545-4266. DOI: 10.1146/annurev.biophys.32.110601.141800.

[33] P. Mignon. "Influence of the - interaction on the hydrogen bonding capacity of stacked DNA/RNA bases." en. In: *Nucleic Acids Research* 33.6 (Mar. 2005), pp. 1779–1789. ISSN: 1362-4962. DOI: 10.1093/nar/gki317.

[34] R. F. Gesteland, T. Cech, and J. F. Atkins, eds. *The RNA world: the nature of modern RNA suggests a prebiotic RNA world*. 3rd ed. Cold Spring Harbor monograph series 43. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press, 2006. ISBN: 978-0-87969-739-6.

[35] P. Yakovchuk. "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix." en. In: *Nucleic Acids Research* 34.2 (Jan. 2006), pp. 564–574. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkj454.

[36] C. Fernando, G. Von Kiedrowski, and E. Szathmáry. "A Stochastic Model of Nonenzymatic Nucleic Acid Replication: "Elongators" Sequester Replicators." In: *Journal of Molecular Evolution* 64.5 (May 2007), pp. 572–585. ISSN: 0022-2844, 1432-1432. DOI: 10.1007/s00239-006-0218-4.

[37] D. Andrieux and P. Gaspard. "Nonequilibrium generation of information in copolymerization processes." In: *Proc. Natl. Acad. Sci. USA* 105.28 (2008), pp. 9516–9521.

[38] E. G. D. Cohen. "Properties of nonequilibrium steady states: a path integral approach." In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.07 (July 2008), P07014. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/07/P07014.

[39] M. A. Nowak and H. Ohtsuki. "Prevolutionary dynamics and the origin of evolution." In: *Proc. Natl. Acad. Sci. USA* 105.39 (2008), pp. 14924–14927.

[40] D. Jost and R. Everaers. "A Unified Poland-Scheraga Model of Oligo- and Polynucleotide DNA Melting: Salt Effects and Predictive Power." In: *Biophysical Journal* 96.3 (Feb. 2009), pp. 1056–1067. ISSN: 00063495. DOI: 10.1529/biophysj.108.134031.

[41] M. Manapat, H. Ohtsuki, R. Bürger, and M. A. Nowak. "Originator dynamics." In: *J. Theor. Biol.* 256.4 (2009), pp. 586–595.

[42] K. Schlosser and Y. Li. "DNAzyme-mediated catalysis with only guanosine and cytidine nucleotides." In: *Nucleic Acids Res* 37.2 (2009), pp. 413–420.

[43] I. Schoen, H. Krammer, and D. Braun. "Hybridization kinetics is different inside cells." In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21649–21654. ISSN: 0027-8424. DOI: 10.1073/pnas.0901313106.

[44] Z. Adamczyk, K. Sadlej, E. Wajnryb, M. L. Ekiel-Jeżewska, and P. Warszyński. "Hydrodynamic radii and diffusion coefficients of particle aggregates derived from the bead model." In: *J. Colloid Interface Sci.* 347.2 (July 2010), pp. 192–201. ISSN: 00219797. DOI: 10.1016/j.jcis.2010.03.066.

[45] E. Kervio, A. Hochgesand, U. E. Steiner, and C. Richert. "Templating efficiency of naked DNA." In: *Proceedings of the National Academy of Sciences* 107.27 (July 2010), pp. 12074–12079. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0914872107.

[46] J. Kim and M. Mrksich. "Profiling the selectivity of DNA ligases in an array format with mass spectrometry." In: *Nucleic Acids Res.* 38.1 (Jan. 2010), e2–e2. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkp827.

[47] M. L. Manapat, I. A. Chen, and M. A. Nowak. "The basic reproductive ratio of life." In: *Journal of Theoretical Biology* 263.3 (Apr. 2010), pp. 317–327. ISSN: 00225193. DOI: 10.1016/j.jtbi.2009.12.020.

[48] C. B. Mast and D. Braun. "Thermal Trap for DNA Replication." In: *Phys. Rev. Lett.* 104 (18 May 2010), p. 188102. DOI: 10.1103/PhysRevLett.104.188102.

[49] S. Rajamani, J. K. Ichida, T. Antal, D. A. Treco, K. Leu, M. A. Nowak, J. W. Szostak, and I. A. Chen. "Effect of Stalling after Mismatches on the Error Catastrophe in Nonenzymatic Nucleic Acid Replication." In: *Journal of the American Chemical Society* 132.16 (Apr. 2010), pp. 5880–5885. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja100780p.

[50] P. Reineck, C. J. Wienken, and D. Braun. "Thermophoresis of single stranded DNA." In: *Electrophoresis* 31.2 (2010), pp. 279–286. ISSN: 1522-2683. DOI: https://doi.org/10.1002/elps.200900505.

[51] J. P. Schrum, T. F. Zhu, and J. W. Szostak. "The Origins of Cellular Life." In: *Cold Spring Harbor Perspectives in Biology* 2.9 (Sept. 2010), a002212–a002212. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a002212.

[52] D. H. Turner and D. H. Mathews. "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure." In: *Nucleic Acids Research* 38.suppl_1 (Jan. 2010), pp. D280–D282. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkp892.

[53] C. Deck, M. Jauker, and C. Richert. "Efficient enzyme-free copying of all four nucleobases templated by immobilized RNA." In: *Nature Chemistry* 3.8 (Aug. 2011), pp. 603–608. ISSN: 1755-4330, 1755-4349. DOI: 10.1038/nchem.1086.

[54] K. A. Dill and S. Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. 2nd ed. London ; New York: Garland Science, 2011. ISBN: 978-0-8153-4430-8.

[55] B. Obermayer, H. Krammer, D. Braun, and U. Gerland. "Emergence of Information Transmission in a Prebiotic RNA Reactor." In: *Physical Review Letters* 107.1 (June 2011). ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.107.018101.

[56] J. W. Szostak. "An optimal degree of physical and chemical heterogeneity for the origin of life?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1580 (Oct. 2011), pp. 2894–2901. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2011.0140.

[57] I. I. Cisse, H. Kim, and T. Ha. "A rule of seven in Watson-Crick base-pairing of mismatched sequences." In: *Nature Structural & Molecular Biology* 19.6 (June 2012), pp. 623–627. ISSN: 1545-9993, 1545-9985. DOI: 10.1038/nsmb.2294.

[58] 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." In: *Nucleic Acids Research* 40.10 (May 2012), pp. 4711–4722. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gks065.

[59] I. García-Robles, J. Sánchez-Navarro, and M. de la Peña. "Intronic hammerhead ribozymes in mRNA biogenesis." In: *Biological Chemistry* 393.11 (Nov. 2012), pp. 1317–1326. ISSN: 1437-4315, 1431-6730. DOI: 10.1515/hsz-2012-0223.

[60] C. Hammann, A. Luptak, J. Perreault, and M. de la Pena. "The ubiquitous hammerhead ribozyme." en. In: *RNA* 18.5 (May 2012), pp. 871–885. ISSN: 1355-8382. DOI: 10.1261/rna.031401.111.

[61] J. W. Szostak. "The eightfold path to non-enzymatic RNA replication." In: *Journal of Systems Chemistry* 3.1 (Dec. 2012), p. 2. ISSN: 1759-2208. DOI: 10.1186/1759-2208-3-2.

[62] K. Adamala and J. W. Szostak. "Competition between model protocells driven by an encapsulated catalyst." In: *Nature Chemistry* 5.6 (June 2013), pp. 495–501. ISSN: 1755-4330, 1755-4349. DOI: 10.1038/nchem.1650.

[63] C. J. Kershaw and R. T. O'Keefe. "Splint Ligation of RNA with T4 DNA Ligase." In: *Recombinant and In Vitro RNA Synthesis*. Ed. by G. L. Conn. Vol. 941. Totowa, NJ: Humana Press, 2013, pp. 257–269. DOI: 10.1007/978-1-62703-113-4_19.

[64] K. Leu, E. Kervio, B. Obermayer, R. M. Turk-MacLeod, C. Yuan, J.-M. Luevano, E. Chen, U. Gerland, C. Richert, and I. A. Chen. "Cascade of Reduced Speed and Accuracy after Errors in Enzyme-Free Copying of Nucleic Acid Sequences." In: *Journal of the American Chemical Society* 135.1 (Jan. 2013), pp. 354–366. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja3095558.

[65] C. B. Mast, S. Schink, U. Gerland, and D. Braun. "Escalation of polymerization in a thermal gradient." In: *Proceedings of the National Academy of Sciences* 110.20 (May 2013), pp. 8030–8035. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1303222110.

[66] T. E. Ouldridge, P. Šulc, F. Romano, J. P. K. Doye, and A. A. Louis. "DNA hybridization kinetics: zippering, internal displacement and sequence dependence." In: *Nucleic Acids Research* 41.19 (Oct. 2013), pp. 8886–8895. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkt687.

[67] B. Rauzan, E. McMichael, R. Cave, L. R. Sevcik, K. Ostrosky, E. Whitman, R. Stegemann, A. L. Sinclair, M. J. Serra, and A. A. Deckert. "Kinetics and Thermodynamics of DNA, RNA, and Hybrid Duplex Formation." In: *Biochemistry* 52.5 (Feb. 2013), pp. 765–772. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/bi3013005.

[68] J. Vignard, G. Mirey, and B. Salles. "Ionizing-radiation induced DNA double-strand breaks: A direct and indirect lighting up." In: *Radiother. Oncol.* 108.3 (Sept. 2013), pp. 362–369. ISSN: 01678140. DOI: 10.1016/j.radonc.2013.06.013.

[69] E. Kervio, B. Claasen, U. E. Steiner, and C. Richert. "The strength of the template effect attracting nucleotides to naked DNA." In: *Nucleic Acids Research* 42.11 (June 2014), pp. 7409–7420. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gku314.

[70] S. Tanaka, H. Fellermann, and S. Rasmussen. "Structure and selection in an autocatalytic binary polymer model." In: *EPL (Europhysics Letters)* 107.2 (July 2014), p. 28004. ISSN: 0295-5075, 1286-4854. DOI: 10.1209/0295-5075/107/28004.

[71] P. G. Higgs and N. Lehman. "The RNA World: molecular cooperation at the origins of life." In: *Nature Reviews Genetics* 16.1 (Jan. 2015), pp. 7–17. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3841.

[72] M. Jauker, H. Griesser, and C. Richert. "Copying of RNA Sequences without Pre-Activation." In: *Angewandte Chemie International Edition* 54.48 (Nov. 2015), pp. 14559–14563. ISSN: 14337851. DOI: 10.1002/anie.201506592.

[73] M. Kreysing, L. Keil, S. Lanzmich, and D. Braun. "Heat flux across an open pore enables the continuous replication and selection of oligonucleotides towards increasing length." In: *Nature Chemistry* 7.3 (Mar. 2015), pp. 203–208. ISSN: 1755-4330, 1755-4349. DOI: 10.1038/nchem.2155.

[74] A. V. Tkachenko and S. Maslov. "Spontaneous emergence of autocatalytic information-coding polymers." In: *The Journal of Chemical Physics* 143.4 (July 2015), p. 045102. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4922545.

[75] W. J. Cannan and D. S. Pederson. "Mechanisms and Consequences of Double-Strand DNA Break Formation in Chromatin: DOUBLE-STRAND DNA BREAK FORMATION IN CHROMATIN." In: *J. Cell. Physiol.* 231.1 (Jan. 2016), pp. 3–14. ISSN: 00219541. DOI: 10.1002/jcp.25048.

[76] E. Kervio, M. Sosson, and C. Richert. "The effect of leaving groups on binding and reactivity in enzyme-free copying of DNA and RNA." In: *Nucleic Acids Research* 44.12 (July 2016), pp. 5504–5514. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw476.

[77] G. J. S. Lohman, R. J. Bauer, N. M. Nichols, L. Mazzola, J. Bybee, D. Rivizzigno, E. Cantin, and T. C. Evans. "A high-throughput assay for the comprehensive profiling of DNA ligase fidelity." In: *Nucleic Acids Res.* 44.2 (Jan. 2016), e14–e14. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv898.

[78] Y. J. Matsubara and K. Kaneko. "Optimal size for emergence of self-replicating polymer system." In: *Physical Review E* 93.3 (Mar. 2016), p. 032503. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.93.032503.

[79] N. Prywes, J. C. Blain, F. Del Frate, and J. W. Szostak. "Nonenzymatic copying of RNA templates containing all four letters is catalyzed by activated oligonucleotides." In: *eLife* 5 (June 2016), e17756. ISSN: 2050-084X. DOI: 10.7554/eLife.17756.

[80] R. Rao and M. Esposito. "Nonequilibrium Thermodynamics of Chemical Reaction Networks: Wisdom from Stochastic Thermodynamics." In: *Phys. Rev. X* 6 (4 Dec. 2016), p. 041064. DOI: 10.1103/PhysRevX.6.041064.

[81] T. Walton and J. W. Szostak. "A Highly Reactive Imidazolium-Bridged Dinucleotide Intermediate in Nonenzymatic RNA Primer Extension." In: *Journal of the American Chemical Society* 138.36 (Sept. 14, 2016), pp. 11996–12002. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.6b07977.

[82] H. Fellermann, S. Tanaka, and S. Rasmussen. "Sequence selection by dynamical symmetry breaking in an autocatalytic binary polymer model." In: *Physical Review E* 96.6 (Dec. 2017), p. 062407. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.96.062407.

[83] J. M. Huguet, M. Ribezzi-Crivellari, C. V. Bizarro, and F. Ritort. "Derivation of nearest-neighbor DNA parameters in magnesium from single molecule experiments." In: *Nucleic Acids Research* 45.22 (Dec. 2017), pp. 12921–12931. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkx1161.

[84] S. Jung, D. Lee, S. W. Kim, and S. Y. Kim. "Persistence Length and Cooperativity Estimation of Single Stranded DNA using FCS Combined with HYDRO Program." en. In: *Journal of Fluorescence* 27.4 (July 2017), pp. 1373–1383. ISSN: 1053-0509, 1573-4994. DOI: 10.1007/s10895-017-2072-8.

[85] L. M. R. 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." In: *Nature Communications* 8.1 (Dec. 2017), p. 1897. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02065-3.

[86] L. Li, N. Prywes, C. P. Tam, D. K. O'Flaherty, V. S. Lelyveld, E. C. Izgu, A. Pal, and J. W. Szostak. "Enhanced Nonenzymatic RNA Copying with 2-Aminoimidazole Activated Nucleotides." In: *Journal of the American Chemical Society* 139.5 (Feb. 2017), pp. 1810–1813. ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.6b13148.

[87] J. W. Szostak. "The Narrow Road to the Deep Past: In Search of the Chemistry of the Origin of Life." In: *Angewandte Chemie International Edition* 56.37 (Sept. 2017), pp. 11037–11043. ISSN: 14337851. DOI: 10.1002/anie.201704048.

[88] A. Tupper, K. Shi, and P. Higgs. "The Role of Templating in the Emergence of RNA from the Prebiotic Chemical Mixture." In: *Life* 7.4 (Oct. 2017), p. 41. ISSN: 2075-1729. DOI: 10.3390/life7040041.

[89] L. H. Gonçalves da Silva and D. Hochberg. "Open flow non-enzymatic template catalysis and replication." In: *Physical Chemistry Chemical Physics* 20.21 (2018), pp. 14864–14875. ISSN: 1463-9076, 1463-9084. DOI: 10.1039/C8CP01828F.

[90] E. Hänle and C. Richert. "Enzyme-Free Replication with Two or Four Bases." In: *Angewandte Chemie International Edition* 57.29 (July 2018), pp. 8911–8915. ISSN: 14337851. DOI: 10.1002/anie.201803074.

[91] A. Mariani, C. Bonfio, C. M. Johnson, and J. D. Sutherland. "pH-Driven RNA Strand Separation under Prebiotically Plausible Conditions." In: *Biochemistry* 57.45 (Nov. 2018), pp. 6382–6386. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/acs.biochem.8b01080.

[92] Y. J. Matsubara and K. Kaneko. "Kinetic Selection of Template Polymer with Complex Sequences." In: *Physical Review Letters* 121.11 (Sept. 2018), p. 118101. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.121.118101.

[93] T. E. Ouldridge. "The importance of thermodynamics for molecular systems, and the importance of molecular systems for thermodynamics." In: *Natural Computing* 17.1 (Mar. 2018), pp. 3–29. ISSN: 1567-7818, 1572-9796. DOI: 10.1007/s11047-017-9646-x.

[94] M. Sosson and C. Richert. "Enzyme-free genetic copying of DNA and RNA sequences." In: *Beilstein Journal of Organic Chemistry* 14 (Mar. 2018), pp. 603–617. ISSN: 1860-5397. DOI: 10.3762/bjoc.14.47.

[95] A. V. Tkachenko and S. Maslov. "Onset of natural selection in populations of autocatalytic heteropolymers." In: *The Journal of Chemical Physics* 149.13 (Oct. 2018), p. 134901. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.5048488.

[96] 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." In: *Chem. Sci.* 10 (22 2019), pp. 5807–5814. DOI: 10.1039/C9SC00770A.

[97] 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." In: *Angewandte Chemie International Edition* 58.37 (2019), pp. 13155–13160. DOI: 10.1002/anie.201907909.

[98] R. Mizuuchi and N. Lehman. "Limited Sequence Diversity Within a Population Supports Prebiotic RNA Reproduction." In: *Life* 9.1 (Feb. 2019), p. 20. ISSN: 2075-1729. DOI: 10.3390/life9010020.

[99] M. Sosson, D. Pfeffer, and C. Richert. "Enzyme-free ligation of dimers and trimers to RNA primers." In: *Nucleic Acids Research* 47.8 (May 2019), pp. 3836–3845. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz160.

[100] S. Toyabe and D. Braun. "Cooperative Ligation Breaks Sequence Symmetry and Stabilizes Early Molecular Replication." In: *Physical Review X* 9.1 (Mar. 2019), p. 011056. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.9.011056.

[101] L. Zhou, S. C. Kim, K. H. Ho, D. K. O'Flaherty, C. Giurgiu, T. H. Wright, and J. W. Szostak. "Non-enzymatic primer extension with strand displacement." en. In: *eLife* 8 (Nov. 2019), e51888. ISSN: 2050-084X. DOI: 10.7554/eLife.51888.

[102] L. Geyrhofer and N. Brenner. "Coexistence and cooperation in structured habitats." In: *BMC Ecology* 20.1 (Dec. 2020), p. 14. ISSN: 1472-6785. DOI: 10.1186/s12898-020-00281-y.

[103] P. W. Kudella, A. V. Tkachenko, S. Maslov, and D. Braun. *Ligation of random oligomers leads to emergence of autocatalytic sequence network*. en. preprint. Biophysics, Aug. 2020. DOI: 10.1101/2020.08.18.253963.

[104] H. Subramanian and R. A. Gatenby. "Evolutionary advantage of anti-parallel strand orientation of duplex DNA." In: *Sci. Rep.* 10.1 (Dec. 2020), p. 9883. ISSN: 2045-2322. DOI: 10.1038/s41598-020-66705-3.

[105] L. Zhou, D. K. O'Flaherty, and J. W. Szostak. "Template-Directed Copying of RNA by Non-enzymatic Ligation." In: *Angewandte Chemie* 132.36 (Sept. 2020), pp. 15812–15817. ISSN: 0044-8249, 1521-3757. DOI: 10.1002/ange.202004934.

[106] J. H. Rosenberger, T. Göppel, P. W. Kudella, D. Braun, U. Gerland, and B. Altaner. "Self-Assembly of Informational Polymers by Templated Ligation." In: *Phys. Rev. X* 11 (3 Sept. 2021), p. 031055. DOI: 10.1103/PhysRevX.11.031055.

[107] L. Zhou, D. Ding, and J. W. Szostak. "The virtual circular genome model for primordial RNA replication." en. In: *RNA* 27.1 (Jan. 2021), pp. 1–11. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.077693.120.

Acknowledgments

I want to thank Ulrich Gerland for the supervision and the great freedom he allowed me during my research. I want to thank Bernhard Altaner and Tobias Göppel for their excellent cooperation, which was crowned by the publication of our joint paper. In this context, I would also like to thank Patrick Kudella and Dieter Braun. I thank Stefan Recksiegel for taking care of the computer cluster. Furthermore, I would like to thank Giovanni Giunta and Stephan Kremser for their cooperation in teaching.

I want to thank Milean for her support during a very turbulent and busy time. I want to thank Mona for being a great roommate. I would also like to thank my family in Bad Heilbrunn, who took me in when I was in poor health. I want to thank Chris, Chiara, Maxi, Andrea, and Samuele for being my loyal friends, for letting me crash at their places countless times, for doing yoga, for drinking tea, for hiking, for climbing, for letting me use their car, for talking, etc. Without that support, this thesis had not have been possible. I also thank my mother for taking me in during this moment of upheaval. Thank you all very much for your support.