

Università di Torino
Dipartimento di Fisica
Laurea Magistrale in Fisica dei Sistemi Complessi

Master Thesis



Dimensionality reduction impact on DNA hybridization kinetics
and on DNA-controlled proto-metabolic efficiency

Candidate: Marco Tuccio *University of Turin*
Matricola: 964841

supervisor: Prof. Steen Rasmussen, *Southern Denmark University, Santa Fe Institute,
European Centre for Living Technology,*

co-supervisor: Prof. Michele Caselle, *University of Turin*

Abstract

The quantification of sequence-dependent DNA hybridization kinetics is of fundamental importance in a wide array of DNA nanotechnology applications where functionality and performance critically depend on binding rates. Moreover, it is key to the investigation of protocellular systems where the non-enzymatic self-replication of DNA-encoded information defines underlying metabolism efficiency and shapes evolutionary trajectory. These designs are based on membrane anchoring of DNA species involved in a Lesion Induced DNA Amplification (LIDA) scheme. Therein, hybridization between involved DNA single strands happens in a two-dimensional fashion through translational diffusion of DNA anchored upon the outer layer of a lipid vesicle. We develop the so-called Zipping Graph Model, a computational mesoscale model of DNA hybridization kinetics which incorporates polymer physics based modeling into involved hybridization reaction pathways. The model is optimized against experimental data, achieving state-of-the-art performance in its predictive capabilities and strong correlation between predicted and experimental hybridization rates. We extend its descriptive capabilities to address the hybridization kinetics of membrane-anchored DNA. For conditions of interest in protocellular metabolism applications, we predict, as a consequence of membrane-anchoring, a 10 fold acceleration in hybridization rates and a 27% increase in self-replication efficiency of the LIDA scheme relative to the three-dimensional case. Moreover we predict a negative impact of increasing strands length on membrane-anchored hybridization rates, in stark contrast with length-independence of the three-dimensional hybridization rates.

Foreword

I'd like to begin this thesis dissertation, by thanking all the people that have been with me during this adventure. This thesis has been quite of a ride and I would've never been able to do that alone. I don't think this little note can ever fill in the amount of gratitude needed, I'll try my best.

I'd like to start with who at the first place accepted to work with me. I'd like to thank my Supervisor Prof. Steen Rasmussen for his open-minded guidance, good heart, golden advices, funny anecdotes and for being my scientific and spiritual guide during this journey. Thank you for your trust in me, for the amazing Donostia month and I promise, next time the risotto will be better. I want to thank my scientific partner-in-crime Kristoffer Thomsen, for your invaluable advice that helped me go through this thesis, I don't know how you are capable of listening to my stream of consciousness wanderings into scientific arguments about this thesis without getting instantly bored. You are a really good friend for me. I want to leave here a special thanks to a person that contributed from far away in the past to this achievement of mine, I want to thank my math and science teacher from middle school Francesco Aliotta, you showed me the beauty of science in the best way possible. Just know that this is mostly your fault to begin with. Thanks Giovanni, for teaching me the tips and tricks of object oriented python programming.

I want to thank my brothers and flatmates Raffaele, Pino and Riccardo, for just constantly reminding me to chill down and enjoy the thesis journey, as well as providing countless amounts of coffee in my room and especially Riccardo for always being able to provide the right means of distraction from my attempts at concentration. I'd like to thank my friend Michele, for making ramen broth even if we had to make it together but I had some short noticed thesis-related deadlines. My friends Walid, Sara, Luca and everyone I had the honor to share years of my life with in my bachelor. My hometown friends Salvo, Antonio, Angelo and Cassa, for just going through being friends with me for all these years, as-well as all my other friends from home, thanks.

I want to thank Giordano, my brother-cousin, first for being the one responsible of me getting interested in Physics in the first place, second for sharing with me so much of who you are. Thanks to my Grandfather Rosario, for making me curious about nature and things. My Mother Ernesta, the sweetest woman in the universe. No advanced scientific theory will never be able to explain how great you are, have been, and always will be. Thanks for your support throughout all my life. I wouldn't be here without you metaphorically and *literally*. The only true Sister I have, Elena. The most badass, caring, sweet and crazy sister one could ever hope to get and yeah I was just lucky enough to be blessed by being your brother. Thank you for just being who you are. Sounds general but to me it sounds very specific, I hope it also does for you.

Francesca, you are at the last place here and for very good reasons. You are at the bottom of this list, like my love for you comes from the bottom of my heart. You are the last, the last I think about before going to sleep. You are at the end, at the end of everything that makes sense to me. I hope the universe will start again and life will follow, so that we can meet again in eternity. I wouldn't be here if it wasn't for you. So thank you Francesca, my sweet half. Alike DNA strands, we are complementary and will be forever entangled. You are one of a kind.

Contents

1	Introduction	4
1.1	Minimal living systems	5
1.1.1	Essential components of protocellular metabolism	5
1.1.2	DNA co-factor based ruthenium-complex proto-metabolism	6
1.2	Aim of the present work	7
2	Thermodynamics of interacting DNA strands	9
2.1	Nearest-neighbor model	9
2.1.1	Mismatches and Abasic Defects	12
2.2	Relation between thermodynamics and chemical reaction kinetics	13
2.3	How to design a DNA based replicator?	15
3	Lesion induced DNA amplification	17
3.1	ODE modeling of chemical reaction networks	19
3.2	Representing LIDA	20
3.2.1	Approximate bounds on reaction rates	23
3.3	Analysis of LIDA	25
3.3.1	Example LIDA realization	25
3.3.2	Parameter influence on replication half-life	28
3.3.3	Considerations on strand design	30
3.4	Membrane LIDA design	31
3.5	Simulation of M-LIDA surface kinetics	34
3.6	Conclusions on LIDA	36
4	Fundamental processes of DNA hybridization kinetics	38
4.1	Two-step model	38
4.2	Modeling the nucleation rate	40
4.2.1	Gyration Radius of DNA polymers	41
4.2.2	Zimm Diffusivity	42
4.2.3	Steric Hindrance	42
4.2.4	Activation Energy	44
4.3	Modeling the zipping rate	45
4.3.1	Example results	47

4.4	The influence of membrane anchoring on DNA hybridization	48
4.4.1	Nucleation Rate of Membrane Anchored DNA	48
4.5	Preliminary results on membrane-anchored DNA hybridization kinetics .	53
4.6	Consequences on LIDA and M-LIDA kinetics	55
5	Transitions from misaligned meta-stable complexes	57
5.0.1	Inchworming transition	58
5.0.2	Pseudoknotting transition	59
5.1	Characterization of off-register transitions	60
5.1.1	Inchworming	61
5.1.2	Pseudoknotting	61
5.1.3	Cumulative sliding rate	63
6	Zipping graph model of DNA hybridization	64
6.1	Snapshot of the model	64
6.2	Object-oriented implementation	68
6.2.1	Kinetic graph generation: <code>class Kinetwork</code>	70
6.3	Simulation methods in <code>class Simulator</code> and analysis of results	71
6.4	Fit against experimental data	74
6.5	Comparison with known models	77
6.6	Zipping graph model of membrane-anchored DNA	79
6.7	Zipping graph based predictions on membrane-anchored DNA hybridization kinetics	82
6.8	Zipping graph on the LIDA system	84
7	Discussion and prospects	86
7.1	Considerations on LIDA and membrane-LIDA	86
7.2	Considerations on DNA hybridization modeling	86
7.3	Considerations on the Zipping Graph Model	87
8	Conclusion	91
9	Acknowledgement	92
	Bibliography	93

1 Introduction

Life as we know it is an intricate and dynamic system, exhibiting characteristics such as growth, reproduction, response to stimuli, and adaptation to the environment. Understanding the fundamental principles that govern living organisms and their processes can provide valuable insights and inspire innovations in various fields such as medicine, biotechnology, and artificial intelligence.

The origin and emergence of life remain to this day an unsolved mystery and a fundamental scientific problem that many prominent physicists and scientists from various fields have tackled. After putting the foundations for Quantum Mechanics, Erwin Schrödinger devoted his scientific effort in understanding life, resulting in his famous book "*What is Life?*" published in 1944.

In his book *The Theory of Self-Reproducing Automata*, John Von Neumann discussed the concept of a self-replicating machine, laying the foundation for understanding how living-matter could emerge from non-living matter. *"If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is"* – John Von Neumann. Manfred Eigen, proposed the idea of autocatalysis being the core chemical principle behind the transition from geochemistry to biochemistry in ancient earth.

Ilya Prigogine introduced the concept of self-organization in non-equilibrium thermodynamical systems, proposing the key concept of life being a far-from-equilibrium self-organizing phenomenon that relies on the equilibrium self-assembly of dissipative structures. Philip Anderson, pioneered the study of complexity in physical systems and coined the phrase "more is different" to emphasize the role of emergent phenomena in complex systems, unpredictable from properties of their individual components.

Murray Gell-Mann, famous for his fundamental contributions on the theory of quarks, developed the concept of Complex Adaptive Systems, from which the biological transition of matter can be seen as the cusp at which simple chemical systems become complex adaptive self-controlling systems.

In order to probe the fundamental principles behind Life, from a spirit beautifully described by the words of Feynman "*What I cannot create, I do not understand*", two entire scientific fields called **Artificial Life** and **Synthetic Biology** are devoted to reaching the milestone of bottom-up de-novo creation of life in the laboratory.

The core machinery of life is driven by informational molecules, which store and transmit genetic information. Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are the most well-known nucleic acids responsible for carrying information at the chemical level. Other chemically-plausible alternative nucleic acid like polymers such as xeno nucleic acids (XNA) and peptide nucleic acids (PNA) have also been identified for their possible role in early biological systems.

One of the key questions in the origins of life field is in understanding how genetic information, encoded in DNA or similar nucleic-acid like polymers, can be replicated in the absence of enzymes.

1.1 Minimal living systems

Minimal living systems refer to the most basic form of living systems that exhibit the essential characteristics of life. According to Rasmussen et al. [1, 2], a system can be considered "alive" if it is capable of utilizing free energy to metabolize environmental resources into components through which such system can grow and divide, effectively restarting its life cycle. Furthermore, if the metabolism of such system is partially regulated by inheritable information that can mutate between generations, then a population of such systems has the potential for selection and evolution.

In this context, a **protocell** is defined as the most basic physico-chemical system that satisfies the criteria for minimal living systems. It consists of a simple structure that can carry out basic metabolic processes and exhibit essential characteristics of living organisms, such as growth, reproduction, and evolution.

1.1.1 Essential components of protocellular metabolism

A fundamental protocellular system is the combination of three main functional blocks: an *information system* for self-control and information inheritance, a *metabolism* capable of transduction of environmental energy and matter into building blocks needed for the protocell to exist and a *container* needed for physical localization of non-equilibrium chemical processes and control of interaction with the environment.

In particular, the metabolism comprises an energy transducer of which the efficiency is modulated by a co-factor. If this co-factor possesses combinatorial properties and can replicate, it enables two critical processes:

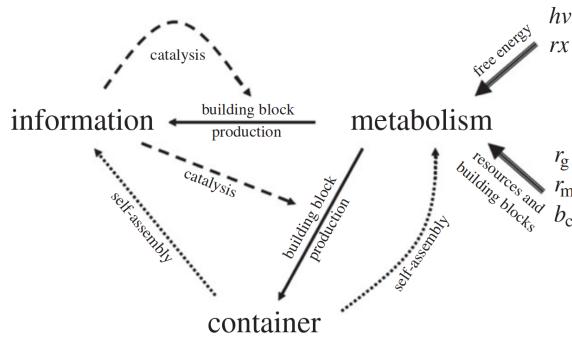


Figure 1.1: Organizational and functional closure in a protocell involves utilizing environmental resources and free energy for primitive metabolism and constructing building blocks for self-assembling containers and informational molecules. The container maintains proximity between metabolism and the informational system, which facilitates catalytic enhancement of metabolic rate kinetics and inheritance of catalytic capabilities across generations.

- The co-factor can serve as an information carrier for the simple living system, allowing for the transfer of genetic information between generations.
- A selection between co-factors becomes possible, resulting in the improvement of metabolic efficiency within the system.

Through its crucial role in regulating the efficiency and rate of the energy transducer, the co-factor becomes the a key controller of the whole metabolism and of the whole protocell.

1.1.2 DNA co-factor based ruthenium-complex proto-metabolism

It is well-established that DNA molecules can act as co-factors in protocellular metabolism¹ [3, 4], acting as an electron donor in a simple protocellular metabolism that implements a ruthenium-complex based energy transducer. In particular, specific DNA sequence is critical to the efficiency of DNA charge-transfer properties, that regulate the entire protocellular metabolism.

Moreover, in the right conditions, DNA co-factors are capable of self-replication, resulting in sequence variability. This will dictate adaptation and evolutionary trajectories for populations of protocells. From an ecological perspective and for a population of protocells in which each protocell "species" possesses a different co-factor sequence, an evolutionary trajectory will emerge which is ultimately dependant on the selection of genetically encoded information in DNA co-factors.

¹in the presence of 8-oxo-guanine, a particular form of the guanine nucleobase found in DNA.

We can then identify the two main contributors in the ecological fitness landscape of a protocell population, ultimately linked to their DNA co-factor sequences:

- The **efficiency of charge transfer** from the energy transducer to the protocellular metabolism.
- The **efficiency of self-replication** of the DNA sequence.

The efficiency of charge transfer defines "*how much per molecule*", while the efficiency of self-replication defines "*how many molecules*". The resulting metabolic efficiency, connected to the speed at which protocells replicate, will depend on the total per-protocell time-dependent charge transfer efficiency $E_p(t)$.

Denoting p a specific protocell and s_p its co-factor sequence, denoting $\delta_p(t)$ the number of self-replicating DNA co-factors at time t of protocell p and $\epsilon_{eff}(s_p)$ the time independent charge transfer efficiency of a single co-factor of protocell p , we compute the total charge transfer efficiency $E_p(t)$ of the metabolism of protocell p :

$$E_p(t) = \epsilon_{eff}(s_p) \delta_p(t) \quad (1.1)$$

Spatial localization in such protocellular systems may be achieved by covalent bonding of DNA co-factors (of hydrophilic nature) to an hydrophobic lipid tail. This will create a so-called DNA-lipid conjugate, which being amphiphilic, will strongly anchor to the outer-layer of a vesicle membrane, resulting in membrane-anchoring mediated spatial localization. Thus, the vesicle surface acts as a 2D container rather than the 3D interior lumen of the vesicle. The advantage of this is simplicity, as resources and waste can easily be exchanged with the environment, without the need to pass through a membrane.

1.2 Aim of the present work

In the present work we explore the **kinetics** of non-enzymatic self-replication of DNA co-factors in the ruthenium-complex and DNA co-factor based protocellular systems. Kinetic characterization is directly equivalent to characterization of the co-factor concentration time evolution $\delta_p(t)$.

This thesis is organized as follows. Chapter 2 delves into the thermodynamics of interacting DNA strands, providing the foundation for understanding non-enzymatic replication of DNA systems. Chapter 3 describes the Lesion Induced DNA Amplification system behind DNA self-replication, along with its detailed deterministic mathematical description and computational analysis of behavior. In chapters 4 and 5, we explore the fundamental processes of DNA hybridization kinetics, with a focus on how membrane anchoring can influence these processes.

Chapter 6 presents the development and implementation of the Zipping Graph Model, a mesoscale computational model for DNA hybridization that provides insights into membrane-anchored DNA hybridization kinetics. Finally, chapter 7 discusses the results and prospects for further research in this area.

2 Thermodynamics of interacting DNA strands

The first steps in our investigation of DNA based non-enzymatic replication systems, entails the knowledge of thermodynamic properties of DNA.

Thermodynamic characteristics of DNA are crucial for understanding the equilibrium properties of interacting DNA systems, which have wide implications for both basic biological and biotechnological research.

The A, C, T, G alphabet from which we can compose sequences gives us the possibility to design systems with desired equilibrium structural properties thanks to the specificity of Watson-Crick-Franklin (WCF) base pairing association between A-T bases and C-G bases. We'll focus our description on DNA but all the methods detailed below can be generalized to any possible nucleic acid like structure, given the availability of experimental data.

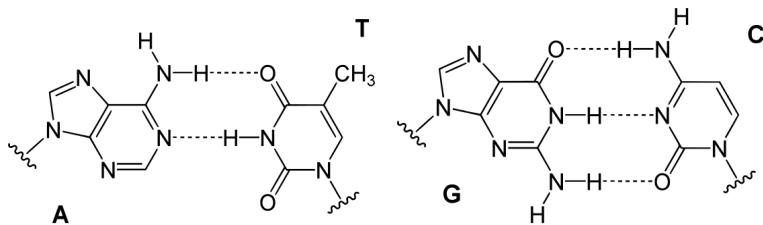


Figure 2.1: Chemical structure of Watson-Crick-Franklin base pairing. G · C pairs have three hydrogen and A · T pairs have two. The overall stability of hydrogen bonded pairs depends on their nearest base pairing context, due to the interplay between aromatic $\pi - \pi$ stacking and hydrogen bonding between subsequent pairs [5]. This fact underlies the Nearest Neighbor Model of DNA Thermodynamics.

2.1 Nearest-neighbor model

The equilibrium properties of nucleic acid systems can be studied through the well known nearest-neighbor (NN) model [6]. We define a nucleic acid strand σ as an ordered set with elements $\sigma_i \in \{A, C, T, G\}$ in the case of DNA strands. The first element of a strand indicates the 5' physical position of the real corresponding nucleic acid, the 3' will correspond to the last element. If we now select an ensemble of L strands $\phi = \{\sigma_1, \dots, \sigma_L\}$, if every possible secondary structure for the given ensemble can be counted, either intermolecular or intramolecular¹, we can construct a state space $\Gamma(\phi)$ over ensemble ϕ .

¹Intermolecular: between two or more separate single strands. Intramolecular: of a single strand with itself

Each element $c \in \Gamma(\phi)$ of our newly constructed state space corresponds to a specific secondary structure, referred to as a **complex**.

A Complex is defined as a tuple containing a set of strands and a set of base pairs, and can be rendered through one of the multiple notations available to efficiently represent complexes and convey intuitive understanding of their structural properties, such as: *dot-bracket notation*, *polymer graphs*, *DU+ notation* or through *structure matrices*.

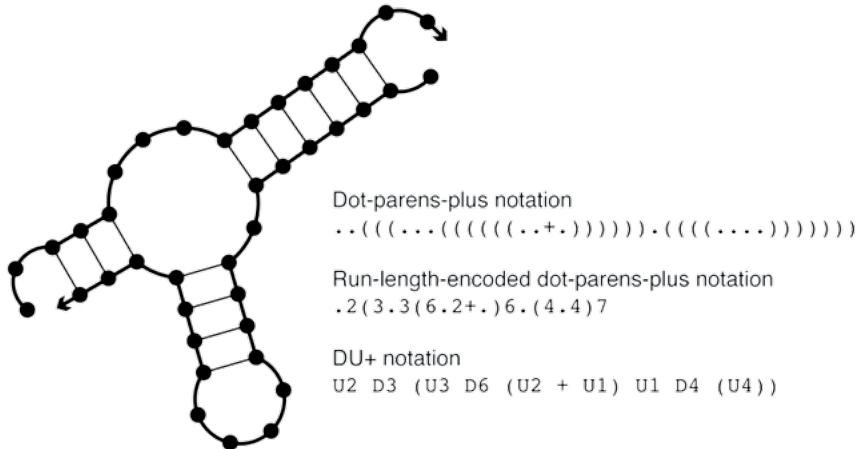


Figure 2.2: Example secondary structure notations, figure from Nupack [7]

For each complex $c \in \Gamma(\phi)$, its free energy $\Delta G(c)$ is the sum of free energies of its constituent loops plus a strand association penalty $\Delta G^{initiation}$ applied $(l_c - 1)$ times for a complex of l_c strands:

$$\Delta G(c) = (l_c - 1)\Delta G^{initiation} + \sum_{loop \in c} \Delta G(loop) \quad (2.1)$$

The strand association penalty is mainly due to entropy loss of the entire system due to the reduction of available states. Loops are microscopic structural conformations that a complex can hold. A visual representation of possible loops for a secondary structure can be found in figure 2.3 below.

The sequence dependent free energy parameter associated with a given loop depends on the equilibrium temperature of the bath in which the system is immersed. This free energy, being of Gibbs type, can be decomposed in enthalpic and entropic contributions as:

$$\Delta G(loop, T) = \Delta H(loop) - T\Delta S(loop) \quad (2.2)$$

The experimental parameters we use to derive free energies are then enthalpies of formation $\Delta H(loop)$ and entropy loss of formation $\Delta S(loop)$ [6].

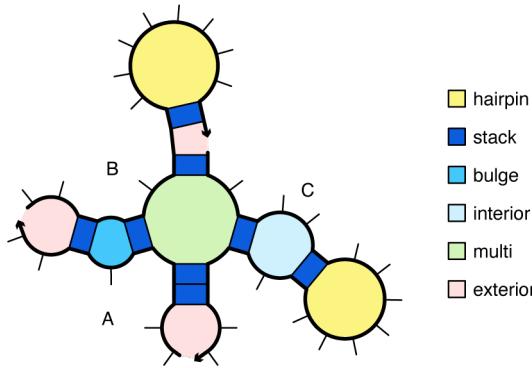


Figure 2.3: Example of possible loops composing a secondary structured complex. Here a complex of three strand. Figure from Nupack [7]

Depending on temperature, loops may be stabilizing ($\Delta G(\text{loop}) < 0$) or destabilizing ($\Delta G(\text{loop}) > 0$) and can be decomposed into canonical loop types:

- **Stacked base pairs:** two consecutive WCF base pairs which aren't separated by unpaired bases. These are the main source of stability of a complex. A subspace $\bar{\Gamma}(\phi)$ is defined as the space where only stacked base pairs interactions are taken into account, called the duplex space. We'll use this space extensively in the next sections.
- **Loops:** The stability of looped structures such as bulges, hairpins, interior loops and multiloops, depends largely on loop size and are almost always destabilizing [6].
 - **Hairpin loops:** a stretch of unpaired bases closed by a stretch of stacked base pairs.
 - **Interior loops:** unpaired bases are simultaneously present on both strands which are closed on both sides by stacked base pairs.
 - **Bulge loops:** same as interior loops but unpaired bases are present in only one of the two involved strands
 - **Dangling ends:** Non-base paired loops at the end of the complex also have a free energy contribution, usually of stabilizing nature due to conformational protection. These are also often called exterior loops.

Each secondary structure can be connected to a compatible ensemble of tertiary structures, which depends on the three dimensional coordinates of nucleotides. The most notable example of such tertiary structures is represented by pseudoknotted structures, for which an example is depicted in figure 2.4.

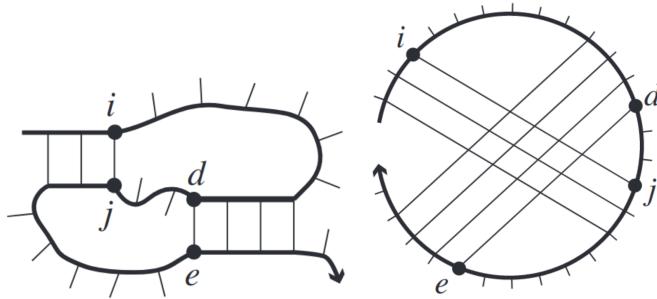


Figure 2.4: Intrastrand pseudoknotted tertiary structure (left) with polymer graph representation (right). A pseudoknotted structure can be recognised by checking crossover between lines in a polymer graph. Figure from [8].

2.1.1 Mismatches and Abasic Defects

A notable case of interest for the design of non-enzymatic DNA replication systems, is represented by the presence of defects in a complex. Given two single strands σ_1 and σ_2 we can distinguish two main cases:

- **Mismatched bases:** a complex will have a mismatched defect at position i if $\sigma_1(i+1) \cdot \sigma_2(i-1)$ and $\sigma_1(i+1) \cdot \sigma_2(i+1)$ and $\neg \sigma_1(i) \cdot \sigma_2(i)$ where « \cdot » stands for a correct WCF pair. The contribution of a mismatch to the stability of a strand strongly depends on which bases are mismatched and on the sequence context around the mismatch. Experimental values for mismatched have been determined for all possible combinations and are reported in [9–13]. The destabilizing effect of mismatches is mainly due to structural alterations, which in turn affect the free energy landscape of the whole strand [14].
- **Abasic sites:** structural defect where a nucleotide is completely missing. It represents a major defect in a DNA strand and has a very high destabilizing contribution [15], significantly altering the thermodynamic stability of their host complex. Although its presence doesn't alter structural properties of the global B-form conformation of an equivalent duplex, it induces a significant enthalpic destabilization of the complex. This can be boiled down to saying that an abasic site induces an almost temperature independent and only sequence dependant destabilization. It is especially dependent on the identity of the "flanking" neighboring base pairs and it depends only secondarily on the identity nucleotide opposite to the abasic defect. To compute the free energy of a complex c with an abasic defect we first compute the $\Delta G(c^{repaired})$ of the equivalent repaired complex (one where the abasic site is replaced with the correct WCF base). We'll then add the sequence context dependant abasic free energy penalty as $\Delta G(c) = \Delta G(c^{repaired}) + \Delta\Delta G(abasic)$ with $\Delta\Delta G(abasic) > 0$ causing a destabilizing contribution²

²Sequence dependant values for $\Delta\Delta G(abasic)$ can be found in [15].

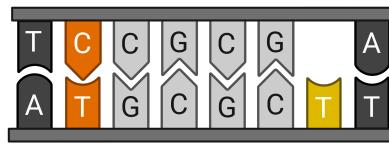


Figure 2.5: An example of a $C \leftarrow T$ mismatch (orange) and an abasic site flanked by $C \cdot G$ and $A \cdot T$ with opposing T base (yellow)

2.2 Relation between thermodynamics and chemical reaction kinetics

Now that we know how to fully characterize each possible complex, we can start characterizing transitions between different complexes. Given an ensemble of L single strands $\phi = \{\sigma_1, \dots, \sigma_L\}$ and their state space $\Gamma(\phi)$ of complexes, from statistical mechanics we can start tackling the problem of how the probability of finding a certain complex state is distributed by computing the canonical partition function over $\Gamma = \Gamma(\phi)$ as:

$$Z = \sum_{c \in \Gamma} \exp\left(-\frac{\Delta G(c)}{kT}\right) \quad (2.3)$$

Where k refers to the Boltzmann Constant. The universal gas constant R has to be used when $\Delta G(c)$ is measured in units of $\frac{Kcal}{mol \cdot K}$, which is often the case for tabulated $\Delta G(\text{loop})$ values. This partition function is of course impossible to compute analytically.

If we wish to calculate it over an ensemble Γ_t containing all the secondary and tertiary structures, the problem becomes NP -hard. The description is hence only commonly limited to combinatorial secondary structures [8]. Dynamic programs and efficient algorithms to compute Z have been developed over the years and can be found in [8, 16] and references thereafter. An efficient implementation is publicly available through the *NUPACK* software suit [7].

Knowing the partition function we can calculate the likelihood of our system to be in a given complex state as:

$$p_c = \frac{1}{Z} \exp\left(-\frac{\Delta G(c)}{kT}\right) \quad (2.4)$$

If then we assume the two complex states $i, j \in \Gamma$ to be reversibly kinetically connected, e.g. can transition from one to the other and vice versa, taken $\Delta G_{ji} = \Delta G(j) - \Delta G(i)$ we can calculate the ratio between the equilibrium probabilities of complexes as:

$$\frac{p_j}{p_i} = \exp\left(-\frac{\Delta G(j) - \Delta G(i)}{kT}\right) \quad (2.5)$$

We can link this to the chemical equilibrium constant of a simple kinetic connection like:
 $i \xrightleftharpoons[k_{ji}]{k_{ij}} j$ Denoting I and J the relative molar concentrations of complexes i and j , if we impose the equilibrium solution on any of the two associated kinetic ODE

$$\frac{dI}{dt} = -k_{ij}I + k_{ji}J = 0 \quad (2.6)$$

Results in $k_{ij}I = k_{ji}J$ giving us the condition of *detailed balancing*:

$$\frac{p_j}{p_i} \equiv \frac{J}{I} \equiv K_E = \frac{k_{ij}}{k_{ji}} = \exp\left(-\frac{\Delta G_{ji}}{kT}\right) \quad (2.7)$$

Where K_E is the *Equilibrium Rate Constant* for a reversible chemical reaction. Detailed balancing is in general sufficient to ensure consistency with the classical thermodynamics of closed systems of ideal mixtures with mass action kinetics.

Although this condition gives us information about the ratio of kinetic rates for a reversible reaction, it doesn't give us any information about the speed at which equilibrium is reached from an initial out of equilibrium state. To have information about the speed of reaction we need to establish a *rate method* through which we set the speed of reaction. The most common way to set a speed of reaction is to use the Metropolis method [17] for which given a reaction rate κ_r associated with the reaction, the rates of transition will be:

$$\text{if } \Delta G(i) > \Delta G(j) \quad (2.8)$$

$$\begin{aligned} k_{ij} &= \kappa_r \\ k_{ji} &= \kappa_r \exp\left(-\frac{\Delta G_{ij}}{kT}\right) \\ \text{if } \Delta G(i) < \Delta G(j) : \quad (2.9) \end{aligned}$$

$$k_{ij} = \kappa_r \exp\left(\frac{\Delta G_{ji}}{kT}\right)$$

$$k_{ji} = \kappa_r$$

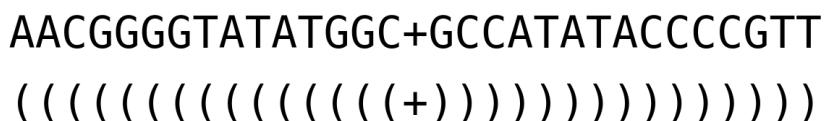
The metropolis method puts then an upper bound on how fast a reaction can happen, the question still remains open on how to calculate the reaction rate κ_r , question that will be addressed in the case of DNA in Chapters 4, 5 and 6. An entire theoretical framework called *Reaction-Rate Theory* has been developed during the last century to answer this particular question, a general review of reaction-rate theory can be found in [18].

2.3 How to design a DNA based replicator?

A replicator can be broadly defined as a physical entity that is capable of exploiting environmental resources and converting these into more copies of itself. DNA is the quintessential replicator molecule, copying itself together with the information to build the biological "machines" it is at the centre of.

Nowadays, DNA replication relies on the complex dynamical interplay of the molecular machinery decoded and built from the information genetically embedded in it. The core mechanism of this process is the conversion of single nucleotides scattered in solution into an ordered DNA oligomer, performed by complex molecular machinery in a base by base fashion.

How could DNA replication happen with the minimum aid of such advanced of molecular machinery characteristic of modern life? DNA is a very stable molecule, under ideal conditions it can be preserved for thousands of years. While its robustness is one of its main features, it is also a limit for its ability to self-replicate: once a duplex is formed, it will keep this stable state for a very long time. Energy has to be consumed to open it back and copy it and enzymes do that extremely efficiently. For example, the very short duplex:



Has already a very high equilibrium rate constant at 25°, owing it great stability:

$$\Delta G = -23.729 \text{ Kcal/mol} \quad (2.10)$$

$$K_E = \exp\left(-\frac{\Delta G}{RT}\right) = 2.17 \cdot 10^{17} \quad (2.11)$$

Such duplex will be extremely slow at replication without the aid of enzymes, since for replication to happen there needs to be a way to "open" the double strand and without the aid of enzymes there will be only really low concentrations of single stranded DNA available for replication.

Synthetic self-replicating systems, moreover, should demonstrate sigmoidal amplification in order to be functional as biomimetic models. DNA "as-is" cannot achieve such amplification due to very strong product inhibition (as previously highlighted), preventing DNA to achieve sigmoidal growth in its concentration. One way to avoid product inhibition is

the use of step-wise procedures like thermal cycling used in Polymerase Chain Reaction amplification of DNA samples and relies on the presence of modern-life enzymes.

As a matter of fact a PCR cycle, even if it achieves amplification of concentration of an initially low sample of target DNA (hence replication of it) doesn't qualify as a system where DNA is a self-replicating molecule, as it relies on artificially induced cyclical conditions on the temperature. For a system to be an autonomous self-replicator, the replication process must proceed with the only two following conditions: availability of the replicator and availability of resources.

A design principle that can be implemented to achieve sigmoidal amplification of a replicator, would imply the overcoming of the product-inhibition barrier, which is explored in the next chapter.

3 Lesion induced DNA amplification

In this chapter we'll introduce and study the Lesion Induced DNA Amplification (LIDA) system. We'll see how to represent it in sections 3.1 and 3.2. We'll explore its behavior in a specific case through numerical simulation in section 3.3. We'll then see how it can be applied as an information replication system for synthetic minimal living systems in the Membrane-LIDA design 3.4 and the characteristics of such design.

We'll study in particular **template-directed** replication. The principle behind it stands on finding the right balance between stability and instability, in order to achieve spontaneous cyclic replication.

We start by giving a general definition behind nucleic acid based template directed replicator systems.

An initial concentration of template DNA T_l can be amplified if an arbitrary number of shorter complementary oligonucleotides C_i of DNA are present and have the following properties, taken $\bar{\Gamma}(\{T_l, C_i\})$ the "duplex space" that refers to complexes containing only stacked bases, mismatches and abasic defects between strands T_l and C_i and for a collection of complementary duplexes $\phi = (C_1, \dots, C_N)$:

- For the duplex space $\bar{\Gamma}(\{T_l, C_i\})$ and $\forall C_i \in \phi$, the most stable hybridized duplex between T_l and C_i defined as $c_{min}(i) = \min c \in \Gamma(\{\bar{T}_l, O_i\})(\Delta G(c))$:
 - is stable $\rightarrow c_{min}(i) < 0$
 - is not perfectly matching: given the number of nucleotides in C_i $l = |C_i| \in \mathbb{N}$ and the number of stacked base pairs in complex c_{min} $b \in \mathbb{N} \rightarrow b < l - 1$
- The ordered concatenation of the collection of sequences relative to their position of hybridization with the template strand doesn't present overlaps and is equal to the complementary sequence of the template strand.

When these conditions are satisfied, the system will have a stable state in which all the short complementary oligomers are hybridized with the template. If an auxiliary chemical system is present to perform the ligation of the missing backbone phosphodiester bonds between adjacent oligomers, these will be joined in a unique strand of length equal to the initial template but complementary.

If enough defects such as **mismatches** and **abasic defects** are present in the complementary strand¹ the duplex will dissociate in its two single strands, avoiding product inhibition. The Template strand will then transform solubilized oligomers in more copies of its complementary self through the so-described *template directed* mechanism. The functioning of this principle can be pictured as a reaction cycle as in fig. 3.1 and is analysed in [19].

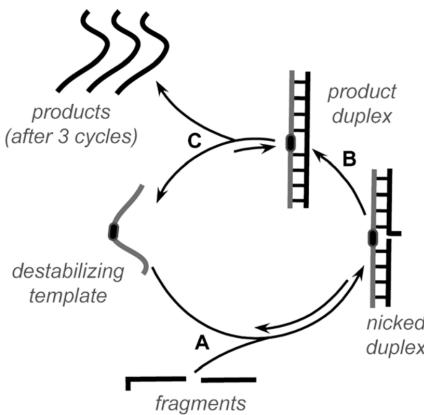


Figure 3.1: Isothermal turnover in DNA-templated ligation reactions using destabilizing templates, figure from [19].

If we join two complementary cycles of template directed replication, we obtain the **Lesion Induced DNA Amplification** system, LIDA [20, 21], capable of template-directed isothermal sigmoidal amplification of an initially low concentration of replicating DNA strands [22, 23].

In this chapter we'll explore the time evolution of a LIDA system through numerical integration of the associated system of Ordinary Differential Equations representing the rate of concentration change of species involved. We are interested in understanding the role of the various involved parameters in setting the overall replication rate of our initial template.

We start our investigation by determining the appropriate level of abstraction at which we wish to represent the system, we do this by defining the process underlying each single reactions involved and describing its kinetics at that level. In the case of LIDA, single reactions involved are hybridization and de-hybridization between two single strands of DNA and a reduced reaction representing the catalytic mechanism underlying ligation of the nicked duplex. This will define the structure of the system, which to be completely defined will need to have rate constants associated to each reaction, which will be discussed in 3.2.

¹such that the resulting equilibrium rate K_E of the newly formed *duplex* state is considerably lower than the equilibrium rate of the equivalent "repaired duplex" K_E^{repaired} in which we substitute defects with correct WCF base pairing

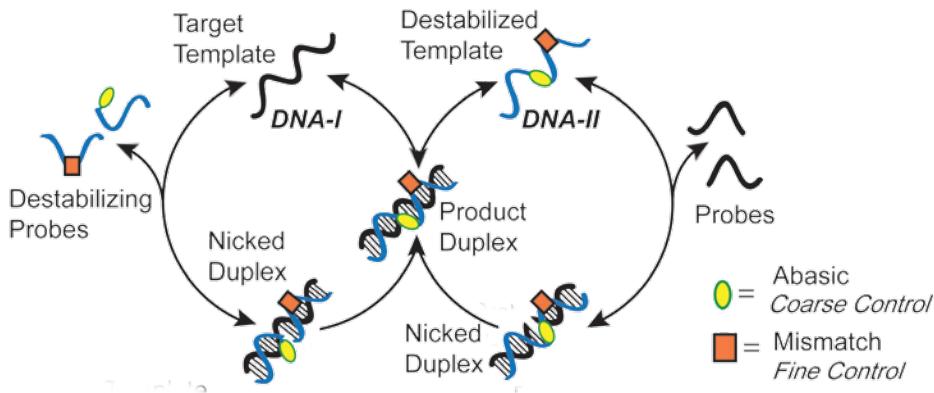


Figure 3.2: LIDA system at the hybridization level of abstraction as depicted in [20].
 (the figure has been modified from the original to highlight the reversibility of hybridization steps)

3.1 ODE modeling of chemical reaction networks

In order to get an abstract mathematical representation of the LIDA system as in figure 3.2 we'll briefly see in this section how to represent a generic chemical reaction network. A Chemical Reaction Network Ω with M involved species in N reactions is a tuple:

$$\begin{aligned} \Omega &= \{\mathbf{S}, \mathbf{T}\} && \text{Chemical Reaction Network} \\ \mathbf{S} &= (\mathbf{s}_1, \dots, \mathbf{s}_M) && \text{set of basis vectors for } \mathbb{R}^M \text{ representing species} \\ \mathbf{T} &= (\mathbf{R}_1, \dots, \mathbf{R}_N) && \text{set of } N \text{ reactions} \end{aligned}$$

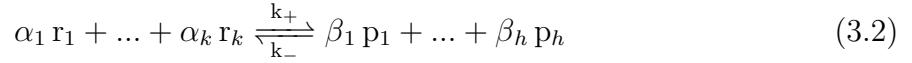
A time-dependent concentration vector $\mathbf{C} \in \mathbb{R}^M$ will be a linear combination of species vectors:

$$\mathbf{C}(t) = \sum_{s \in \mathbf{S}} \rho_s(t) \mathbf{s} \quad (3.1)$$

where $\rho_s(t) \geq 0$ refers to the time-dependent molar concentration of the relative species. Each element $\mathbf{R}_i \in \mathbf{T}$ of the reaction set is defined as:

$$\begin{aligned} \mathbf{R} &= \{\mathbf{r}, \mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, k_+, k_-\} \\ \mathbf{r}, \mathbf{p} &\subseteq \mathbf{S} && \text{reagents and products sets for reaction } \mathbf{R} \\ \boldsymbol{\alpha} &\in \mathbb{N}_+^k && \text{Stoichiometric coefficients of reactants} \\ \boldsymbol{\beta} &\in \mathbb{N}_+^h && \text{Stoichiometric coefficients of products} \\ k_+ &\in \mathbb{R}_+ && \text{reaction rate constant from } \mathbf{r} \text{ to } \mathbf{p} \\ k_- &\in \mathbb{R}_+ && \text{reaction rate constant from } \mathbf{p} \text{ to } \mathbf{r} \end{aligned}$$

With $k = \dim(\mathbf{r})$ and $h = \dim(\mathbf{p})$ This is an equivalent representation of the chemical formula:



For each reaction $R \in \mathbf{T}$ we construct the following vectors:

- A reaction vector $\mathbf{v}_R \in \mathbb{R}^M$ as a linear combination of stoichiometric coefficients and elements of the reactants and products sets as:

$$\mathbf{v}_R = \sum_{i=0}^h \alpha_i p_i - \sum_{j=0}^k \beta_j r_j \quad (3.3)$$

where products show up as positive coefficients and reactants as negative coefficients.

- time dependant reactants and products mass action propensities:

$$z_+(R, t) = k_+ \prod_{r_i \in \mathbf{r}} \rho_{r_i}(t) \quad \text{forward propensity of reaction R at time t} \quad (3.4)$$

$$z_-(R, t) = k_- \prod_{p_j \in \mathbf{p}} \rho_{p_j}(t) \quad \text{backward propensity of reaction R at time t} \quad (3.5)$$

We construct the *MxN Stoichiometric Matrix* Σ as the row vector in reactions space of column reaction vectors in species space:

$$\Sigma = [\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_N] \quad (3.6)$$

The stoichiometric matrix is a compact mathematical representation of our Chemical Reaction Network. It is organized such that each column corresponds to a specific reaction and each row corresponds to a specific species. By construction of a forward propensities vector $\mathbf{f}(t) = (z_+(\mathbf{R}_1, t) \ z_+(\mathbf{R}_2, t) \ \dots \ z_+(\mathbf{R}_N, t))$ and the diagonal matrix $\mathbf{F}(t) = \text{diag}(\mathbf{f}(t))$, doing the same for a $\mathbf{B}(t)$ backward propensities diagonal matrix, we can finally write the ODE system relative to the given Chemical Reaction Network Ω with the initial condition on concentration c_0 as:

$$\frac{d\mathbf{C}(t, c_0)}{dt} = \Sigma(\mathbf{F}(t) - \mathbf{B}(t)) \quad (3.7)$$

3.2 Representing LIDA

Through the previously explained framework of mathematical representation, we'll write a system of ODEs for the LIDA system and numerically integrate it. We assign a nomenclature to the involved species as done in figure 3.3 below.

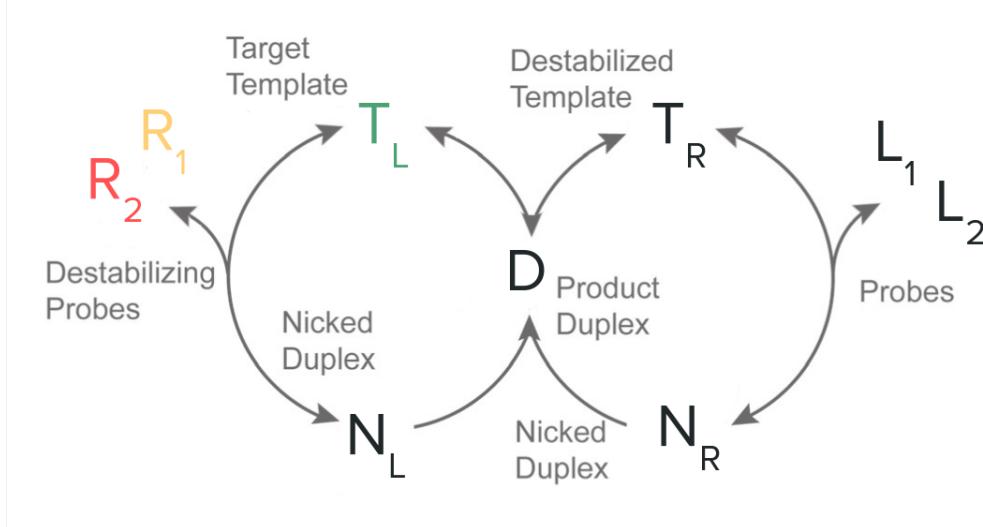


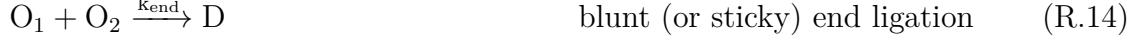
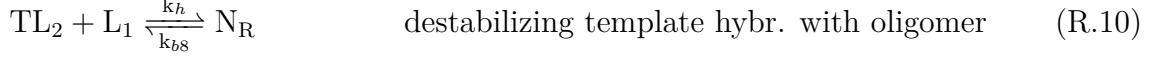
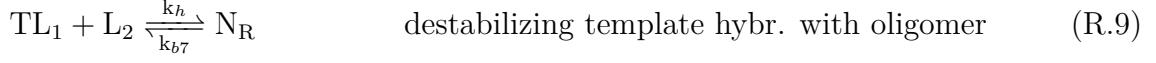
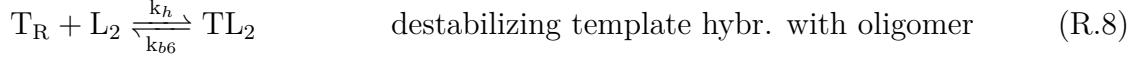
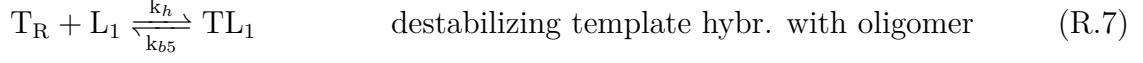
Figure 3.3: Schematic representation of the LIDA system with assigned nomenclature. T_L refers to our target template, R_1 and R_2 refer to destabilizing oligomers, everything else follows the same pattern as in figure 3.2.

Then our CRN $\Omega_{LIDA} = \{S_{LIDA}, T_{LIDA}\}$ will have the following species set as a basis space for \mathbb{R}^{15} :

$$S_{LIDA} = (T_L, R_1, R_2, TR_1, TR_2, N_L, T_R, L_1, L_2, TL_1, TL_2, N_R, D, O_1, O_2) \quad (3.8)$$

$$S_{LIDA} = \left(\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right)$$

Transitions set will contain the following reactions:



Note that we are assigning the same forward hybridization rate k_h to every hybridization reaction, this is a first approximation we take in order to begin our study of the system, which we'll later discard. Reactions (R.1—R.11) refer to the main LIDA process, while reactions (R.12—R.14) refer to the background processes responsible for the slow formation of templates in the absence of a relative initial concentration [20]. The main reactions responsible for this background process are *blunt end ligation* or *sticky end ligation* depending on sequence overlaps. The possibility of blunt or sticky end ligation depends on how the nick duplex repair reaction is catalyzed in the scheme. In particular the background reaction (R.14) can only happen if catalyzation is realized through ATP powered modern biological enzymes such as *T4 DNA Ligase* as done by [20]. Due to the wide functional capacity of *T4 Ligase* [24, 25] the formation of T_l templates can happen through the route:

$$(R.12) + (R.13) \rightarrow (R.14) \rightarrow (R.6).$$

Instead when the non-enzymatic ruthenium complex based ligation scheme is used [4],

thanks to the higher functional specificity of this reaction system, blunt or sticky end ligation cannot happen and we dismiss (R.14) from our model.

The corresponding stoichiometric matrix of the LIDA system Σ_{LIDA} will be:

$$\Sigma = \begin{matrix} & R1 & R2 & R3 & R4 & R5 & R6 & R7 & R8 & R9 & R10 & R11 & R12 & R13 & R14 \\ T_L & -1 & -1 & 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ R_1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ R_2 & 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 \\ TR_1 & +1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ TR_2 & 0 & +1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ N_L & 0 & 0 & +1 & +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ T_R & 0 & 0 & 0 & 0 & 0 & +1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ L_1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ L_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & -1 & 0 \\ TL_1 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ TL_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & -1 & 0 & 0 & 0 & 0 \\ N_R & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & -1 & 0 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 & +1 & 0 & 0 & +1 \\ O_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 & -1 \\ O_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 \end{matrix} \quad (3.9)$$

As a general heuristic for writing stoichiometric matrices, for each reaction column one can fill in negative stoichiometric coefficients of its reactants and positive stoichiometric coefficients of its products as a consequence of equation (3.3).

3.2.1 Approximate bounds on reaction rates

As can be seen from the outlined reactions (R.1-R.14), in order to fully define the kinetics of the LIDA system considered, we need to define two rates:

- 1. The hybridization rate of DNA single strands, from which to also recover the relative de-hybridization rates from thermodynamics (sec. 2.2)
- 2. The rate of ligation of the missing backbone phosphodiester bond for nicked duplexes in reactions (R.5) and (R.11)

We start our description from hybridization rates and then move on the ligation rate.

We make here an initial approximation on the hybridization kinetic rate by assuming it has a constant value of $k_h = 2 \cdot 10^7$ as experimentally found in stopped flow experiments [26]. This approximation inherently assumes that hybridization rates are not dependent on

the sequence of the involved strands, neither on more broad spatial characteristics of the system. We'll later need to eliminate this assumption in order to construct a better model for LIDA and make realistic predictions. For a given duplex participating in LIDA, we can compute its free energy through the NN model as detailed in 2.2. As an example, if we wish to get kinetic parameters for the hybridization reaction $T_L + R_1 \xrightleftharpoons[k_-]{k_+} TR_1$ we will have:

$$k_+ = k_h$$

$$k_- = k_h \exp\left(-\frac{\Delta G(TR_1)}{RT}\right)$$

The ligation reaction, is represented by reactions (R.5) (R.11). In general these reactions are representing reduced mechanisms of the more complex catalysis mechanism underlying ligation of a nicked duplex [27]. The catalyzed reaction of nick reparation can be either enzyme mediated as in [20] or follow nonenzymatic DNA ligation schemes, such as photo-catalytic processes involving a ruthenium photosensitizer outlined in [4]. When *T4 DNA Ligase* is used, ligation can be simplified through enzymatic Ping-Pong kinetics [24]. It is moreover show that T4 has the ability to seal nicks in DNA even in the presence of mismatches in the proximity of the nick.

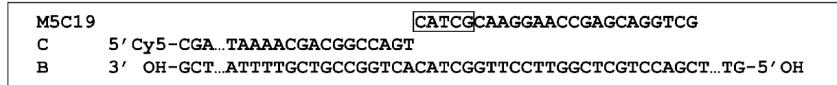


Figure 3.4: As shown in [24], T4 Ligase can seal nicks with a mismatched region up to 5 mismatched base pairs. These particular strands have been repaired with an observed kinetic rate of $k_{repair} = 4.7 \cdot 10^{-3} s^{-1}$. We'll use this value as a lower bound on possible ligation kinetics

We don't fix the ligation rate k_{lig} to a certain value but strive to keep it inside plausible ranges, since it depends on many experimental factors and can be fine tuned:

$$k_{lig} \in [4.7 \cdot 10^{-3}, 0.5] s^{-1}$$

Non-enzymatic ligation has been reported to achieve ligation rates inside the same range [4], so the reported k_{lig} range can be descriptive of both enzymatic or non-enzymatic ligation reaction schemes.

3.3 Analysis of LIDA

We first wish to get a general understanding of how the LIDA system works by treating it as an abstract dynamical system. We first ought to understand from its mathematical structure if it can admit multiple equilibria or not. To answer this initial question we can assess from formal theorems in chemical reaction networks theory [28] that if every nonzero stoichiometric coefficient is equal to unity, then the chemical reaction network admits only one equilibrium. This condition is satisfied as from Σ_{LIDA} (3.9).

3.3.1 Example LIDA realization

We'll base our investigation of the LIDA system on the same sequences used for its experimental characterization [20]. The system has one abasic defect and one A · G mismatch.

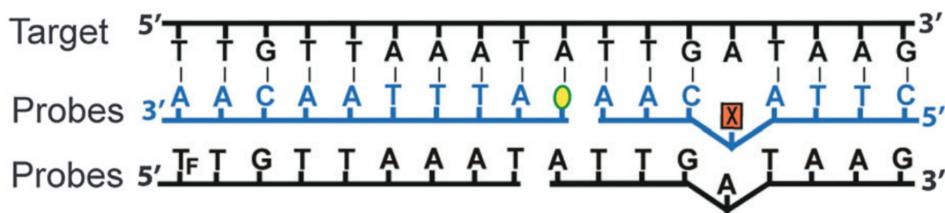


Figure 3.5: Target template and oligomers sequences used in our description. The **X** label refers to a generic mismatch. We'll use a A · G mismatch for our free energy calculations. Figure from [20]

The system has been implemented in *MATLAB 2022b*. Various implemented numerical integrators have been tested for performance. Due to observed numerical stiffness of the implemented ODE system, best performance has been achieved by using the variable step size *ode15s* integrator. We first compare the results of our model against experimental results outlined in [20]. Results are obtained for initial template concentration $\rho(T_l, 0) = 14nM$, $1.4nM$, $140pM$ and initial oligomers concentration $\rho(oligo, 0) = \rho(R_1, 0) = \rho(R_2, 0) = \rho(L_1, 0) = \rho(L_2, 0) = 2.8\mu M$.

In order to faithfully compare our simulations with data, we compute the total %Yield of the system as done in [20], from which we confront our results:

$$\%Yield(t) \simeq \frac{\rho(T_l, t) + \rho(D, t)}{\rho(T_l, t) + \rho(D, t) + \rho(oligo, t)} \quad (3.10)$$

Accordance with experimental data has been found even with our rough estimates on hybridization reaction rates of $k_h = 2 \cdot 10^7 (Ms)^{-1}$ by optimization of the unknown effective ligation rate to a value of $k_{lig} = 8 \cdot 10^{-2} s^{-1}$, which is realistic and inside possible bounds [24].

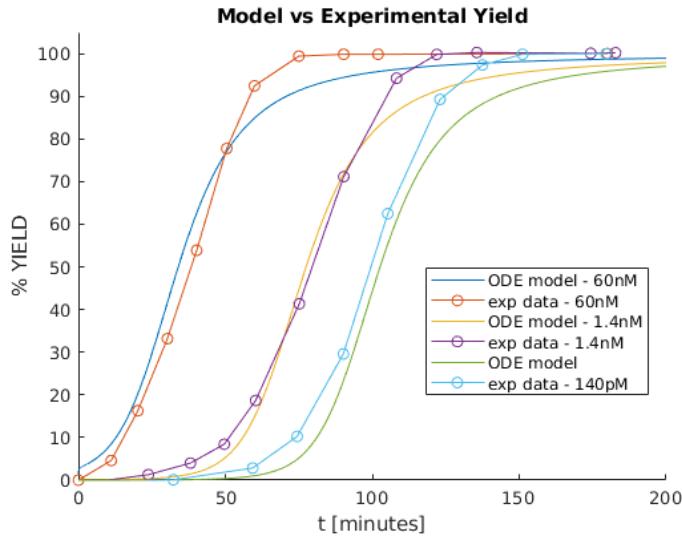


Figure 3.6: LIDA percent yield of target template over time against experimentally measured data from [20]

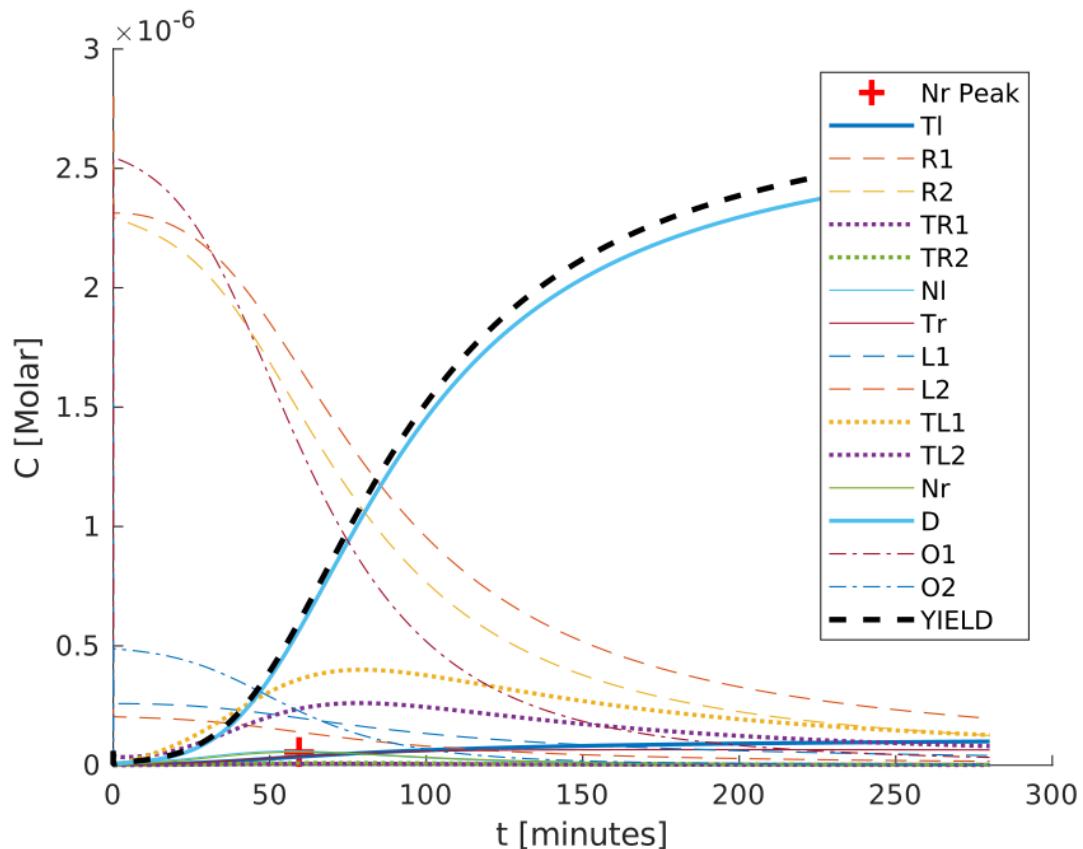


Figure 3.7: Realization of the LIDA system with rate constants as in section 3.2.1 and initial oligomer concentration $\rho(\text{oligo}, 0) = 0.1\text{mM}$ with shown time evolution of concentration for all involved species. Detailed explanation of the main steps involved can be found in the figures below.

Figure 3.7 shows a generic realization of the LIDA system, where the separation of time scales inherent in the system is visible. Reaction pathways (R.12) and (R.13) that form background "oligomer duplexes" reach equilibrium after time $\tau \simeq 1/\rho(\text{oligo}, 0)(k_h + k_b) \simeq 0.1\text{s}$ [18] while the half-life of replication relative to D and T_l needs a time in the order of hours.

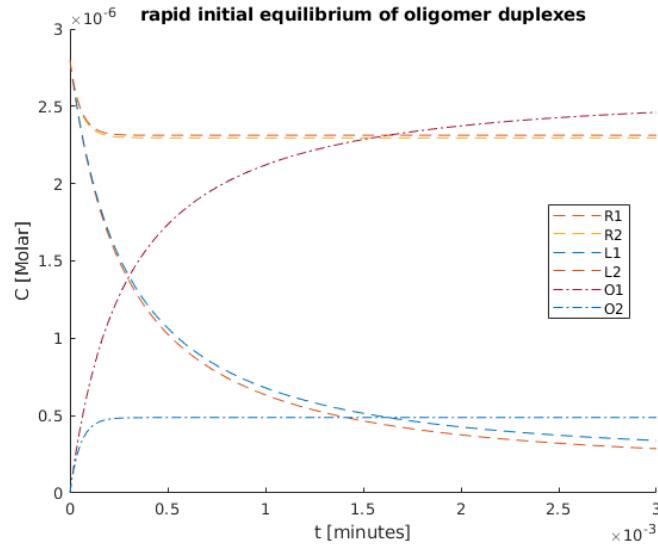


Figure 3.8: Rapid equilibrium for oligomer duplexes. O_2 reaches a fast equilibrium with $K_E(R13) = 1.68 \cdot 10^6$ due to the presence of the mismatch. O_1 presents a higher equilibrium constant $K_E(R12) = 4.86 \cdot 10^7$, this is due to the external position of the abasic defect, acting as a dangling end.

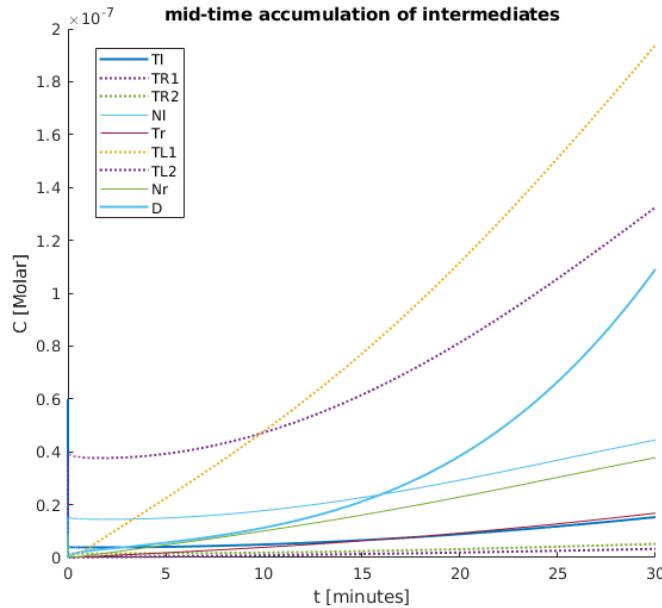


Figure 3.9: Accumulation of intermediates at concentrations comparable with the initial template concentration.

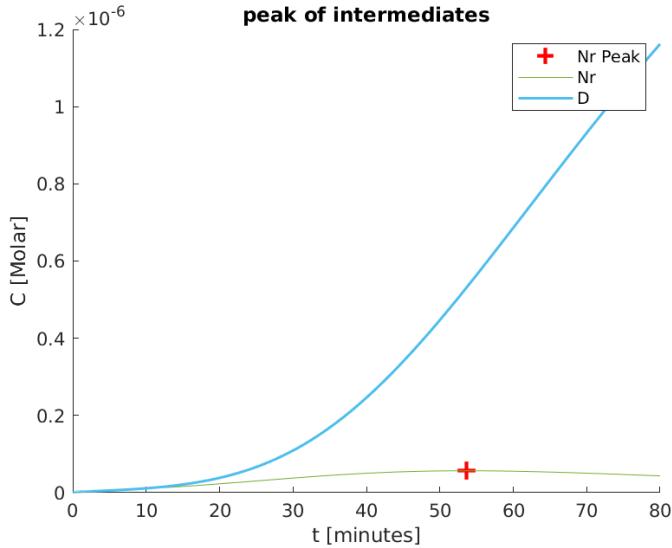


Figure 3.10: A peak is observed in the concentration of intermediates. We found that the peak of concentration of the "right" Nicked duplex N_r always coincides with the zero for the second derivative of the curve representing the time evolution of duplex concentration. We use this point as indicative of the half-life of replication $\tau_{1/2}$

3.3.2 Parameter influence on replication half-life

The dependence of replication half-life $\tau_{1/2}$ on the various parameters of interest has been explored through extensive numerical integration of LIDA. We start by performing Local Sensitivity Analysis (LSA) by calculating time-dependent partial derivatives of concentrations of interest on parameters. In our case we treat kinetic rate constants as parameters upon which output concentration curves will depend, generically for a model with output x and parameters y and z we calculate the quantities:

$$\frac{y}{x} \frac{\partial x}{\partial y}, \frac{z}{x} \frac{\partial x}{\partial z} \quad (3.11)$$

By performing LSA on the concentration $\rho(N_r, t)$ of N_r over which we calculate the replication half-time $\tau_{1/2}$ as its peak. As in figure 3.11 we find strongest dependence on ligation rate k_{lig} and duplex de-hybridization rate k_{bd} , thus we choose to explore these parameters.

We then construct a grid-based algorithm in order to characterize parameter-dependant values of the half-life $\tau_{1/2}$. Each point in the grid will correspond to a specific combination

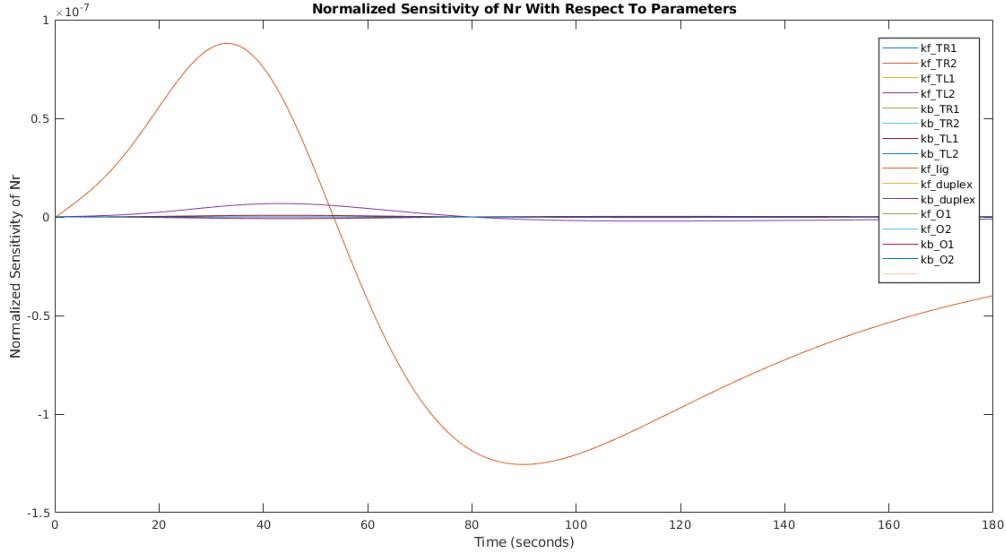


Figure 3.11: Local Sensitivity Analysis on Right Nicked Duplex shows strongest dependence on ligation rate k_{lig} and duplex de-hybridization rate $k_{bd_{duplex}}$

of the parameters we identified as being the most influential, k_{lig} and k_{bd} (see figure 3.11):

$$\begin{aligned} k_{lig} &= \text{logspace}(10^{-5}, 10^0, 40) && 40 \text{ log spaced points from } 10^{-5} \text{ to } 10^0 \\ k_{bd} &= \text{logspace}(10^{-8}, 10^8, 40) && 40 \text{ log spaced points from } 10^{-8} \text{ to } 10^8 \\ n_{simulations} &= 1600 && \text{total simulations performed} \end{aligned}$$

The resulting surface log-log-log plots inform us on the functional dependence of $\tau_{1/2}$ on k_{lig} and $k_{bd_{duplex}}$: From plot 3.13a is evident that, for each fixed $k_{bd_{duplex}} = k_{bd}$ the LIDA system shows a power-law like dependence on k_{lig} that can be summarized as:

$$\tau_{1/2}(k_{lig}) = a(k_{bd})k_{lig}^{-b(k_{bd})} \quad (3.12)$$

Here $a(k_{bd})$ and $b(k_{bd})$ are parameters dependant on duplex de-hybridization kinetics, with a related to overall scaling while b is related to the slope-exponent.

The half-life depends exponentially on duplex dehybridization rate k_{bd} as shown in 4.4b. From this observation we have a confirmation of the baseline idea behind the LIDA system: dialling in more destabilization in the involved strands will increase the efficiency of the overall system. The stability of duplexes will exponentially inhibit replication rates, bringing the system in the regime of product inhibition.

Decreasing duplex stability such that the resulting equilibrium rate constant is of order $K_E \simeq 10^4$, will not decrease $\tau_{1/2}$. In this parameter region duplex dehybridization is replaced by nicked duplex ligation as the rate limiting factor. Moreover it is observed that the transition from exponential to null dependence on k_{bd} doesn't depend on k_{lig} as the ligation rate k_{lig} becomes the rate limiting factor. In such regimes, an increase of the ligation rate is needed in order to increase LIDA efficiency. However, describing such case in is outside the scope of this thesis.

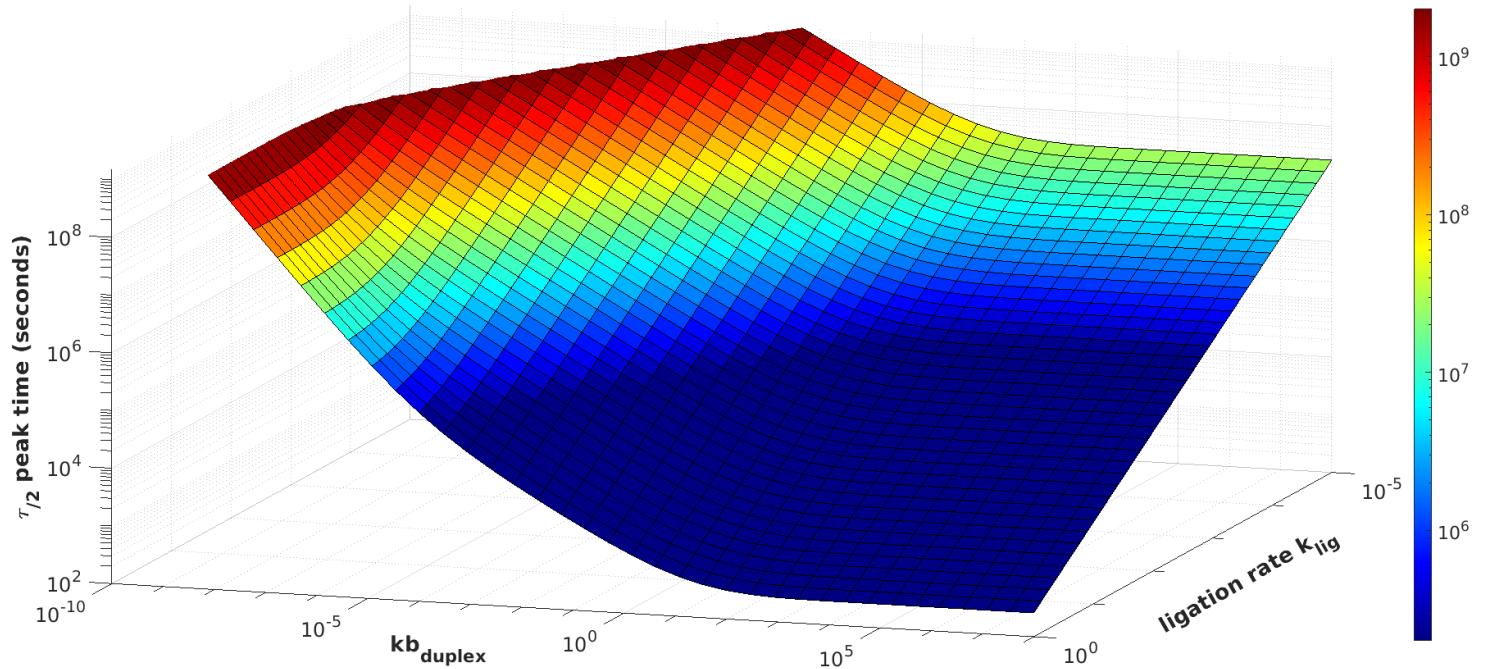


Figure 3.12: surface plot of LIDA replication half-life surface by varying k_{lig} and duplex destabilization k_{bd}

3.3.3 Considerations on strand design

From our analysis we can now propose important factors in the design of DNA sequences involved in a LIDA system. First we note the importance of the position of the abasic defect at the 5' most end of oligomer $R1$ (see figure 3.5). This is a fundamental design element since the abasic destabilization only affects the duplex and doesn't affect as much the stability of the involved oligomer. There's no thermodynamic knowledge about the destabilization induced by an abasic defect nearby a mismatch, we propose it as an interesting experimental investigation that might improve our understanding of destabilization in DNA.

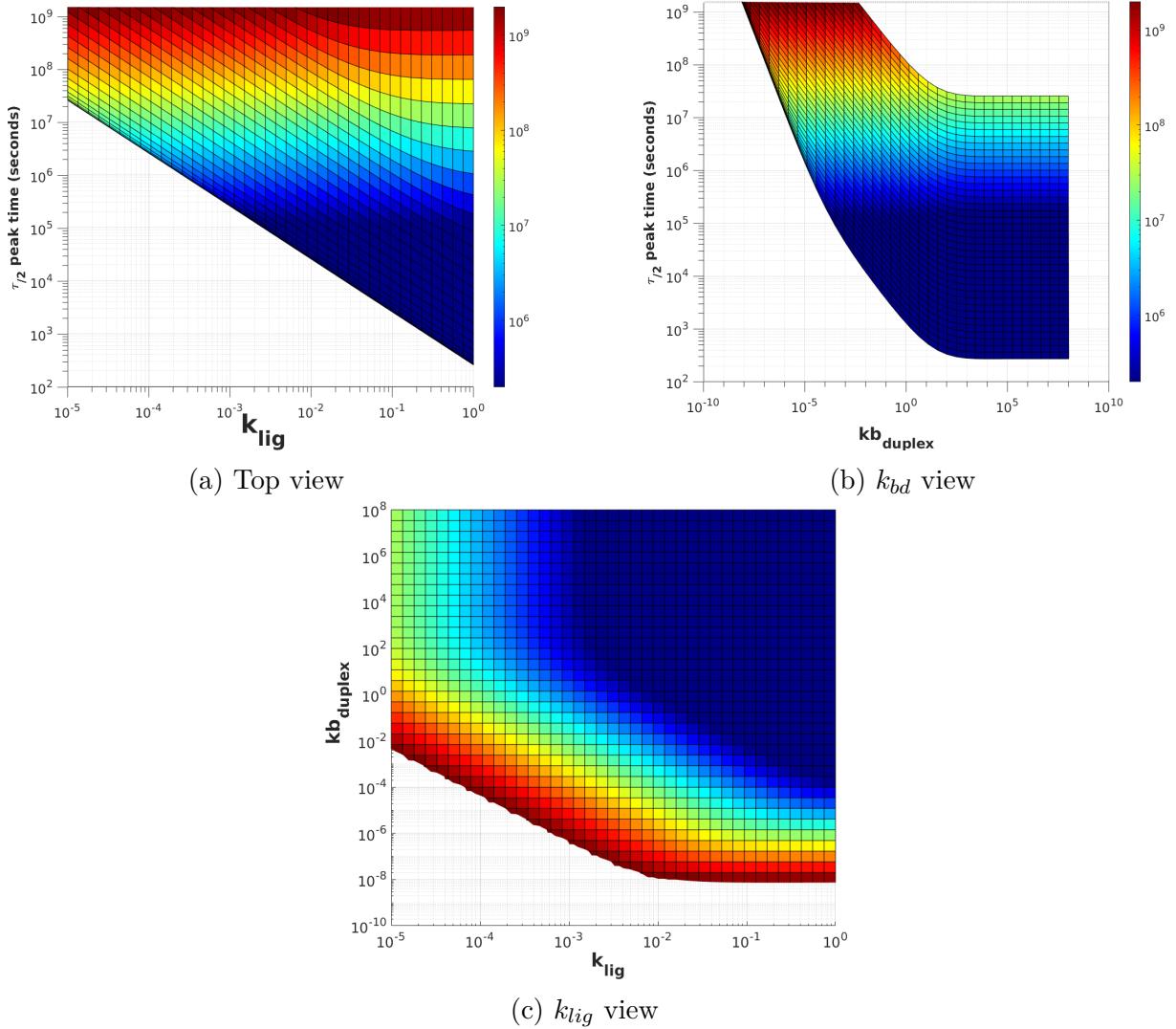


Figure 3.13: side views

3.4 Membrane LIDA design

The possibility to chemically integrate functional groups into oligonucleotides, most notably with lipophilic anchors, enabled a widespread usage of DNA as a viable alternative to proteins with respect to functional activity on membranes. In protocell design, anchoring an information-replication scheme on the outer layer of a bilayer lipid membrane will embed it in a spatially confined region, enabling control and heritability of information.

The use of LIDA as the information-replication mechanism in protocell design is proposed by Rasmussen and outlined in [1].

Membrane-LIDA, M-LIDA for short, is a modification of the LIDA system where involved DNA species are chemically bound to lipophilic "tail" in order achieve high affinity to membrane anchoring. The design of M-LIDA implies the anchoring of every involved

species except one of the right oligomers and one of the left oligomers. If we bind the lipophilic tail at the 5' of an oligonucleotide, after membrane-insertion its 3' end will be the farthest from the surface. We call bottom the lipophilic end and top the opposite "hydrophilic" end. If we anchor T_l at the 5' end, the two incoming oligomers R_1 and R_2 will be complementary to the bottom half of T_l while the other will be with the top half. The same logic applies to T_r and its complementary oligomers L_1 and L_2 . We'll take the convention now to refer to oligomers R_1 and L_1 as the bottom ones while R_2 and L_2 will be the top ones. From this follows that R_1 and L_1 will be membrane-anchored while R_2 and L_2 will be free in solution.

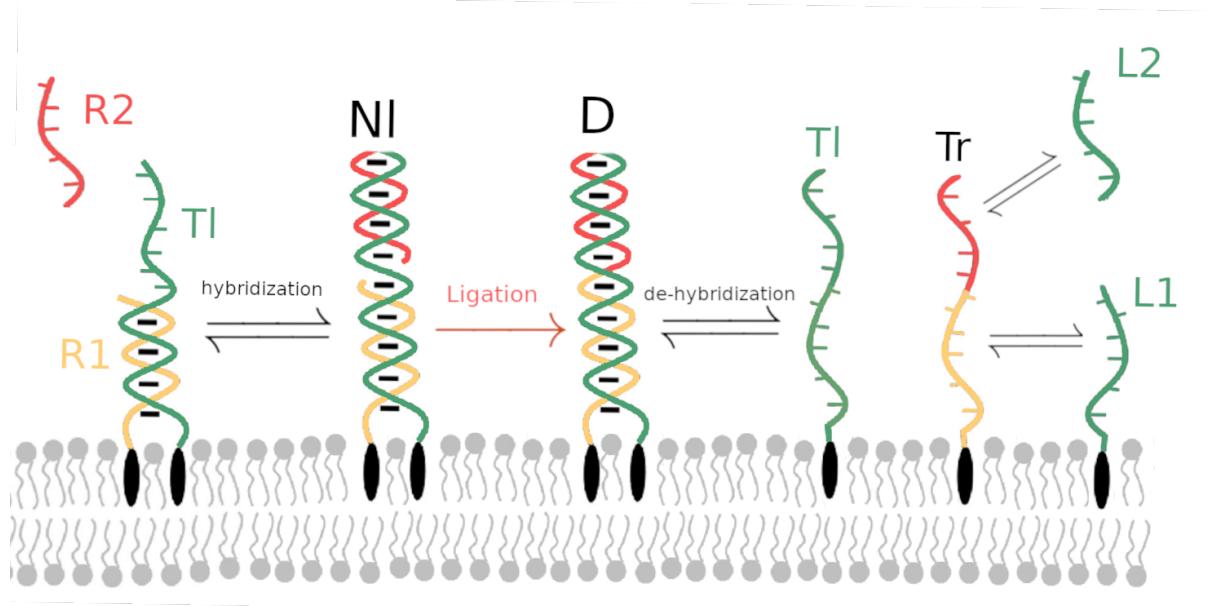


Figure 3.14: Illustration of some of M-LIDA reactions on a surface. Bulk R_2 hybridizes with TR_1 making the left nicked duplex N_l , which ligates in the duplex D . After dehybridization the new strands will enter two new cycles of replication with their complementary oligomers.

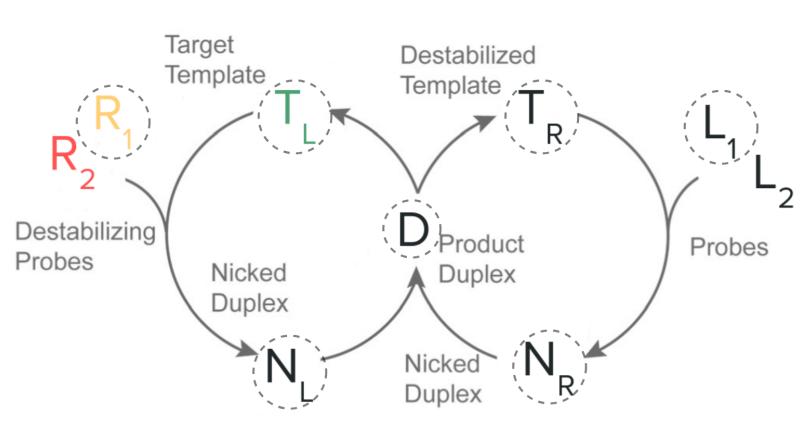


Figure 3.15: Circled species are membrane-anchored

DNA-lipid conjugates are synthetic structures where a DNA simplex is covalently bound to a lipophilic tail adopted to have the desired insertion affinity relative to membrane. DNA-

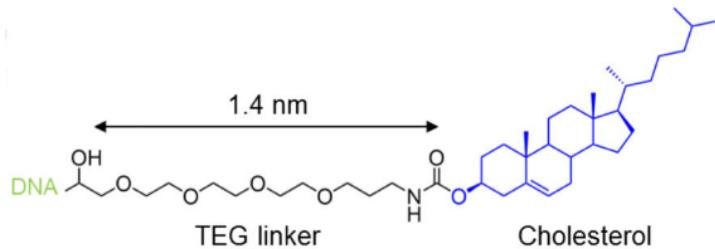


Figure 3.16: An example of a DNA-lipid conjugate structure with cholesterol tail.

lipid conjugate insertion affinity for membrane insertion depends on the hydrophobicity of the used lipid, on its affinity with the target membrane and on the electrical charge of membrane lipids head groups, due to possible electrostatic interactions with the negatively charged DNA. A very strong affinity can be in general assumed for DNA-cholesterol conjugates in non-charged membranes [29]. We'll use this fact to simplify our treatment of M-LIDA by assuming the given volume concentration of DNA to be fully adsorbed on the outer layer of vesicles membrane, this can be easily implemented experimentally by efficient design [2].

The M-LIDA design will comprise then **four** main types of reaction:

- **Two-dimensional DNA Hybridization:** membrane-anchored and translationally diffusing DNA-lipid conjugates of complementary sequence will undergo hybridization.
- **Target-Probe DNA Hybridization:** this reaction represents hybridization between a surface-bound *probe* single strand and the complementary free in solution *target* single strand. It is also widely known in the literature as *surface DNA hybridization*², this reaction is not to be confused with the purely two-dimensional DNA hybridization reaction.
- **Usual DNA Hybridization** in solution.
- **The ligation reaction** to repair the nicked duplex.

We'll treat Target-Probe DNA Hybridization as being the same as usual bulk DNA Hybridization, as it shows deviations from it only in the regime of crowded membranes and only in the case of targets complementary to near-surface sequences of the probe [30].

²Causing several revealed-fake Gotcha! moments to the literature-researching author

We first see how to treat such modifications from a modeling perspective. We'll see that proper transformations on kinetic rate constants apply and use these for probing half-life surfaces at the variation of the two-dimensional hybridization rate constant.

3.5 Simulation of M-LIDA surface kinetics

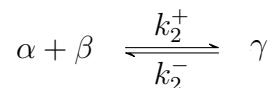
We describe here our simulation approach on treating the Membrane-LIDA (M-LIDA) system. The main difference here stands in the treatment of surface kinetics.

So far we have described deterministic chemical kinetics at the molar concentration level of description, in such framework the units of measure of rates and concentrations are based on *Molarity* denoted M and measured in moles per liter $\frac{\text{mol}}{\text{dm}^3}$.

The question is: *how do we measure surface concentration?*

One of the most used ways is to use the concept of area fraction. We choose not to use this concept as it is related to the adsorption on solid surfaces with little lateral motion of adsorbed species involved. We on the other hand choose the other most straightforward way to measure a surface concentration x , which is in units of moles per decimeter square, namely $\frac{\text{mol}}{\text{dm}^2}$.

Now consider the following surface bimolecular reaction between species α and β performing only two-dimensional translational diffusive motion resulting in surface species γ :



Now we represent the time evolution of the surface concentration of species α through the usual kinetic ordinary differential equation (for simplicity we denote concentration using the same species symbol):

$$\frac{d\alpha}{dt} = -k_2^+ \alpha \beta + k_2^- \gamma \quad (3.13)$$

In the limit of full adsorption, as in our case [2], we know that the totality of molecules of species α will be adsorbed on the surface available.

If then we have a system of volume V and with enough available surface area S such for the limit of full adsorption to hold³, having an initial volume concentration A of molecules

³the specific nature of this limit depends on a multitude of system wide conditions. We do not wish to model such nuanced and complex aspects and just put ourselves in the situation of full adsorption as a general assumption which surely holds for the case of membrane-anchoring of DNA-lipid conjugates for the experimental setups of our interest [1–3]

to be absorbed into their corresponding surface state α , we can translate between the two by considering fixed volume V and fixed surface S as:

$$\alpha = A \frac{V}{S} \quad (3.14)$$

The result is shown by considering the total number of molecules resulting from the volume concentration and connecting it to the surface concentration computed from that same number of molecules. With B and G referring to volume concentration relative to surface concentrations β, γ , we can then plug (3.14) into (3.13) to obtain:

$$\frac{dA}{dt} \frac{V}{S} = -k_2^+ A \frac{V}{S} B \frac{V}{S} + k_2^+ G \frac{V}{S} = \frac{dA}{dt} = -k_2^+ \frac{V}{S} AB + k_2^+ G \quad (3.15)$$

From here we see that in the limit of full adsorption, is possible to translate the two-dimensional kinetic equations into equivalent descriptions on three-dimensional concentrations by scaling the kinetic rates of surface bimolecular association reactions as:

$$k_3^+ = k_2^+ \frac{V}{S} \quad (3.16)$$

Through this transformation we manage to simplify our description by describing species populations by means of the same volumetric molar quantities.

We can now investigate the joint role of the two hybridization rate constants involved, the surface one $k_{h,2}$ and the bulk one $k_{h,3}$. We consider for parameter exploration the resulting two-dimensional rate as a three-dimensional equivalent of the "true" two-dimensional one.

We elaborate for the sake of clarity: we consider a k_2^+ range which is the three-dimensional concentration equivalent of the true two dimensional rate \tilde{k}_2^+ , denoted by a tilde. The two are related as in (3.16). The parameters values explored here for the two-dimensional rate can be realized by arbitrary experimental conditions that result in an equivalent k_2^+ from an initial arbitrary combination of \tilde{k}_2^+ , volume V and available surface S , always on the assumption of total adsorption.

Figure 3.17 reports the exploration of parameters: two-dimensional hybridization rate $k_2^+ = [2 \cdot 10^7, \dots, 2 \cdot 10^{12}] (M s)^{-1}$ and three-dimensional hybridization rate $k_3^+ = [2 \cdot 10^6, \dots, 2 \cdot 10^8] (M s)^{-1}$ at three different duplex de-hybridization rates $k_{bd} = (0.003, 0.03, 30) s^{-1}$ producing three-different surfaces. Ranges are chosen on the basis of observations for two-dimensional diffusion limited rates being several orders of magnitude faster than three-dimensional ones 4.4.1. We observe exponential dependence of replication half-life $\tau_{1/2}$ on the two-dimensional hybridization rate constant $k_{h,2}$ while we observe weaker exponential dependence on the bulk target-probe hybridization rate $k_{h,3}$.

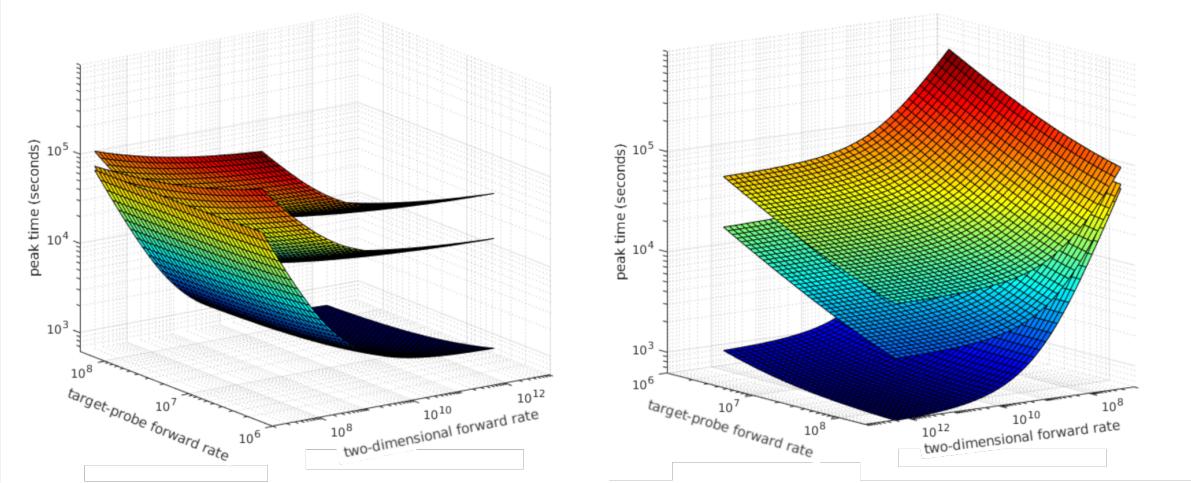


Figure 3.17: Two perspectives of the same surface plot for target-probe hybridization rate k_3^+ versus surface hybridization rate k_2^+ . For the considered parameter range, strong exponential dependence of replication half-life is observed on two-dimensional hybridization rates. Weak exponential dependence is observed on three-dimensional ones. We explore parameters: two-dimensional hybridization rate $k_2^+ = [2 \cdot 10^7, \dots, 2 \cdot 10^{12}]$, three-dimensional hybridization rate $k_3^+ = [2 \cdot 10^6, \dots, 2 \cdot 10^8]$. Three surfaces are produced at different values of duplex stability k_{bd} , in order from top to bottom we have $k_{bd} = (0.003, 0.03, 30)$

We then conclude that if we predict membrane-anchoring produces an acceleration in DNA hybridization, we expect exponentially lower LIDA replication half-life. This exponential acceleration is still limited by the presence of a plateau (fig. 3.17) where ligation rate becomes the rate-limiting factor of the LIDA system.

3.6 Conclusions on LIDA

We are faced with a decision regarding the direction of our exploration of the properties of LIDA. One possible avenue is to undertake an in-depth investigation of the parameter space of the system. This could be accomplished through the use of more complex and efficient methods of characterizing and reducing high-dimensional parameter spaces, e.g. Diffusion Maps. The aforementioned techniques have been successfully utilized in a variety of scientific disciplines and have shown promise in elucidating detailed aspects of complex systems such as LIDA. However, it is important to acknowledge that the representations generated by these methods are influenced by the assumptions we make regarding kinetic rates, and are therefore susceptible to bias. Consequently, the insights obtained from this type of analysis may be limited to the mathematical structure of LIDA and may not be directly applicable to the development of theoretical models for experimental prediction.

Moreover, our true research goal is focused on investigating how the reduction in dimensionality of the M-LIDA design impacts the kinetics of LIDA replication. The process of LIDA replication is, in turn, influenced by the effects that dimensionality reduction imparts on the hybridization rate constants of DNA. To gain a better understanding of these phenomena.

For such reasons, the rest of our investigative effort will be devoted into an in-depth analysis of DNA Hybridization (next chapter) and in the construction of a physics grounded mesoscopic computational model of DNA hybridization kinetics, called Zipping Graph Model, constructed with the aim of producing first predictions on the effects of dimensionality reduction.

4 Fundamental processes of DNA hybridization kinetics

We develop here a baseline understanding of the physics underlying DNA hybridization. The goal of this chapter is to start from a three-dimensional description of the fundamental processes of DNA hybridization and then apply this understanding to the two dimensional case.

The simple model described in this chapter doesn't have accurate predictive power on DNA hybridization kinetics of short sequences and we don't expect it to. We use it as a toy model that can help us navigate through the fundamental physical aspects underlying DNA Hybridization. The physical insights developed in this chapter about the kinetics of fundamental processes involved in DNA hybridization will be fundamental in the construction of the Zipping Graph Model in chapter 6.

4.1 Two-step model

From a purely kinetic standpoint, DNA hybridization is *not* an elementary reaction, it rather follows a more complex reaction sequence.

We start by considering here the simplest possible reaction sequence that we can associate with DNA Hybridization: the nucleation-zipping sequence, based on the foundational theoretical understanding of DNA hybridization kinetics developed by Wetmur in 1968 [31].

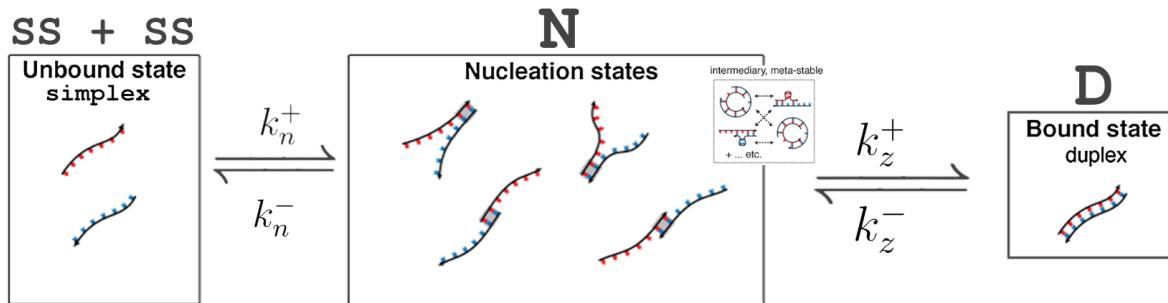
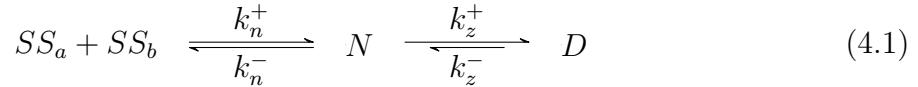


Figure 4.1: Cartoon illustration of the two step model presented in this chapter. Two initially unbound strands perform first nucleation attempts and subsequently close their configuration reaching the duplex state. This process is in reality underlined by mid-way meta-stable states, which are not considered in the two-step model. Figure adapted from [32].

DNA Hybridization kinetics is modeled as a two step process involving an initial slow nucleation step where a first number of base pairs is formed, followed by a fast zipping step where all the remaining base pairs are rapidly closed. Nucleation is known to be the rate-limiting step in DNA hybridization. We denote k_n^+ and k_n^- the forward and backward nucleation rates, k_z^+ and k_z^- the forward and backward zipping rates:



We are treating here the case of two completely complementary DNA single strands SS_a and SS_b of the same sequence length (i.e. number of nucleotides) that undergo hybridization with each other. That is to say that their sequences A with A_0 in the 5'-end of SS_a and B with B_0 the 3'-end of SS_b form correct WCF base pairs at each position.

The nucleation reaction step is a bimolecular reaction where two initially separated single strands, after collision, generate a nucleation complex. This is a metastable state with a low number of stacked base pairs, which can either separate back into the single stranded state with rate k_n^- or proceed with zipping into the full duplex state through the unimolecular zipping reaction with rate k_z^+ .

We coarse-grain this two-step process to a cumulative "observed" kinetic rate of hybridization. We say observed in the sense that in reality, experimental results on the kinetic rates of hybridization k_h measure an *observed* rate constant of hybridization which should correspond to the coarse-grained resulting rate from an underlying more complex model.

Calling $k_{obs} (\equiv k_h)$ this rate, in the two step model it depends on the specific characteristics of the nucleation and zipping rate constants involved.

To get the observed rate constant of transition from the single stranded state $SS_a + SS_b$ to the duplex state D we consider that the waiting time τ_α out of a generic chemical state α is an exponentially distributed random variable (Gillespie [33]):

$$f_{\tau_\alpha}(t) = \lambda_\alpha e^{-\lambda_\alpha t} \quad (4.2)$$

$$\lambda_\alpha = \sum_{\beta \neq \alpha} k_{\alpha \rightarrow \beta} \quad (4.3)$$

Moreover the probability of transition from state α to a specific state γ out of all other possible states is obtained by normalizing the rate constant associated to this transition as:

$$\mathbb{P}_{\alpha \rightarrow \gamma} = \frac{k_{\alpha \rightarrow \gamma}}{\sum_{\beta \neq \alpha} k_{\alpha \neq \beta}} \quad (4.4)$$

Then for the Nucleation Zipping scheme (4.1), with S denoting the single stranded state, we can compute the observed kinetic rate as:

$$\begin{aligned} k_{obs} &= \mathbb{P}_{N \rightarrow D} \left(\frac{1}{E[\tau_S] + E[\tau_N]} \right) \\ &= \frac{k_z^+}{k_z^+ + k_n^-} \left(\frac{1}{\frac{1}{\lambda_S} + \frac{1}{\lambda_N}} \right) \\ &= \frac{k_z^+}{k_z^+ + k_n^-} \left(\frac{1}{\frac{1}{k_n^+} + \frac{1}{k_z^+ + k_n^-}} \right) \\ k_{obs} &= \frac{k_z^+ k_n^+}{k_z^+ + k_n^+ + k_n^-} \end{aligned} \quad (4.5)$$

4.2 Modeling the nucleation rate

We model the nucleation step as a *sterically hindered diffusion-controlled* reaction. A diffusion-controlled reaction is by definition a fast bimolecular chemical reaction with a very low activation energy. In such reactions, kinetics are dominated by the rate at which reactants encounter within a specified reaction distance.

The theory of diffusion-controlled reaction rates starts from Smoluchowski [34].

He considered a reaction that instantly happens upon contact between spherical reactants A and B with diffusivities D_a, D_b and reaction radiiuses R_a, R_b . The diffusion-controlled rate is computed by calculating the flux over a surface of radius $R = R_a + R_b$ of particles with diffusion coefficient $D = D_a + D_b$ with vanishing boundary condition and appropriate initial conditions. For $t > 0$ the diffusion-controlled rate constant k_{dl} is equal to:

$$k_{dl}(t) = 4\pi DR \left(1 + \frac{R}{\sqrt{\pi Dt}} \right) \quad (4.6)$$

For long times (4.6) converges to the a constant function $k_{dl} = 4\pi DR = 4\pi(D_A + D_B) + (R_A + R_B)$. When polymers are involved in a bimolecular diffusion-controlled reaction in three-dimensional space, the following results apply [35]. By considering the non-markovian motion of single monomers bound in a polymer chain, authors [35] conclude that two reacting polymer chains need to approach each other at an encounter radius comparable

to the sum of their *gyration radiiuses*. By taking diffusivity equal to sum of centre of mass diffusivities of the two polymers we get:

$$k_{dl-chains} = 4\pi D_{CM} R_G \quad (4.7)$$

With diffusion coefficient $D_{CM} = (D_{CM,A} + D_{CM,B})$ and $R_G = (R_g(a_1, N_1) + R_g(a_2, N_2))$ the reaction radius. To fully characterize nucleation, we need to characterize the gyration radius of single stranded DNA and its centre of mass diffusivity.

4.2.1 Gyration Radius of DNA polymers

The gyration radius refers to the spherical volume occupied *on average* by a polymer. For an ideal linear polymer chain with N monomers of length b , the ensemble average¹ gyration radius is equal to:

$$\langle R_g^2(N, b) \rangle = \frac{Nb^2}{6} \quad (4.8)$$

DNA polymers can be modeled by the **worm like chain** model, sometimes called *Krakty-Porod* model [36], it is used to model stiff polymers where conformational flexibility is mainly due to small fluctuations of the overall contour of the chain and not to sudden directional changes between single bonds.

A worm like polymer with n monomers of length l can be renormalized to a freely jointed chain (a chain with zero orientational correlation between monomers) with N *Kuhn* monomers of *Kuhn* length b of two times the persistence length l_p , namely $b = 2l_p$.

The persistence length measures the distance at which correlation in position decays: $\langle \vec{r}_i \cdot \vec{r}_{i+l_p} \rangle = 0$. It is a key characteristic to define a worm-like polymer and has been measured for single stranded DNA [37] and is reported below:

$$l_d = 0.676 \text{ nm} \quad \text{monomer length ssDNA} \quad (4.9)$$

$$l_h = l_d/2 \quad \text{monomer length dsDNA} \quad (4.10)$$

$$l_p = 2.223 \text{ nm} \quad \text{persistence length} \quad (4.11)$$

$$l_b = l_d/4 \quad \text{backbone bond length} \quad (4.12)$$

Consider the contour length of the chain, namely its maximum possible extension, to be equal by definition to both the original worm like chain and the renormalized freely jointed

¹The ensemble average of the gyration radius is related to the ensemble average of the end-to-end distance as $\langle R_g^2 \rangle = \langle R_{end-end}^2 \rangle / 6$ Due to the lack of correlations, the end-to-end distance $\langle R_{end-end}^2 \rangle$ is effectively equal to the Mean Square Displacement of a random walk that takes N steps of length b : $\langle R_{end-end}^2 \rangle = Nb^2$

chain. We can use it to calculate the number of Kuhn monomers as:

$$R_{max} = nl_d = Nb = 2Nl_p \quad (4.13)$$

$$N = \frac{nl_d}{2l_p} \quad (4.14)$$

We can then write the gyration radius of a DNA single strand as:

$$R_g^2 = \frac{Nb^2}{6} = n \frac{l_d l_p}{3} \quad (4.15)$$

4.2.2 Zimm Diffusivity

The centre of mass diffusivity of single stranded DNA is well described by the Zimm theory of polymer dynamics [38]. The theory is a generalization of the beads-springs Rouse model, that takes into account hydrodynamic interactions between monomers immersed in solution. Starting from Einstein relation $D = kT/\zeta^2$ Zimm [39] computed the centre of mass diffusion coefficient of a polymer chain in solution as:

$$D_Z = \frac{8}{3\sqrt{6\pi^3}} \frac{kT}{\eta_s b N^{1/2}} \quad (4.16)$$

Where k is the Boltzmann constant, T the temperature, b the *Kuhn* length of the renormalized chain and N its number of *Kuhn* monomers. The exponent of N is fixed at $1/2$ under our ideal linear chain approximation.

4.2.3 Steric Hindrance

So far, the validity of our description is contingent upon the assumption that the interacting DNA single strands will form a stable nucleation complex regardless of their relative orientation. However, it is clearly evident from the geometry of DNA molecules, that this assumption is not accurate, as the formation one or more base pairs is possible only with the correct relative orientation of the colliding nucleotides.

If only a limited region of one or both reactants is active, we need to scale down the overall rate by a *steric factor* ρ that considers the probability of a reactive encounter based on geometrical considerations on active sites of reactants. Steric factor ρ , which is a de-facto probability, will be a multiplicative factor on the purely diffusion-controlled

² ζ the friction coefficient

reaction rate. From our knowledge of ssDNA monomer length l_d we know that a segment of persistence length l_p will contain $l_p/l_d \approx 3$ nucleotides. This tells us that we can approximate 3 consecutive nucleotides as a stiff cylinder. We compute steric hindrance ρ by defining a limited active surface area on such cylinders and by calculating the probability of correct relative orientation upon collision. We define the steric factor $\rho(\theta, \phi)$ with θ azimuthal angle and ϕ inclination angle (fig 4.2) as follows:

$$\rho(\theta, \phi) = \left(\frac{\theta}{2\pi}\right)^2 \left(\frac{\phi}{2\pi}\right)^2 = \left(\frac{\theta\phi}{4\pi^2}\right)^2 \quad (4.17)$$

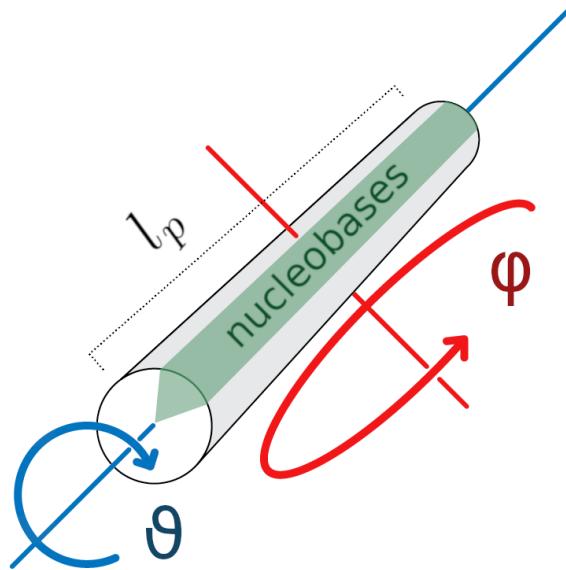


Figure 4.2: Rotational degrees of freedom of a stiff cylinder representing a persistence segment of ssDNA. In green the reactive surface area corresponding to nucleobases. θ is the azimuth angle and ϕ the angle of inclination. l_p is the persistence length of ssDNA (4.11)

For a ssDNA with n nucleotides, the nucleation rate is then obtained by inserting equations (4.17) (4.16) (4.15) into (4.7), substituting the *Kuhn* number of monomers N (4.14) and length b (4.13), we get:

$$k_{nuc}(N, T, \theta, \phi) = \tilde{\rho}(\theta, \phi) \frac{kT}{\eta_s} \sqrt{n} \quad (4.18)$$

$$\tilde{\rho}(\theta, \phi) = \frac{8}{9} \left(\frac{\phi\theta}{2\pi^2}\right)^2 \sqrt{\frac{l_d^3}{3\pi l_p}} \quad (4.19)$$

We underline that the formula we derived, is in accordance with the widely known result that nucleation kinetics are proportional to the square root of the number of nucleotides³ [31].

³Advanced results [40] show more precisely that the nucleation rate is proportional to $k_n \propto N^\nu$. This result

4.2.4 Activation Energy

Lastly, in order to fully characterize the nucleation rate, we need to consider the existence of a small reaction barrier that limits the speed at which base pairs can actually form.

Generally, when an activation energy barrier E_b is present in a reaction with collision rate c , the resulting reaction rate k will be given by Arrhenius as:

$$k = c \exp\left(-\frac{E_b}{kT}\right) \quad (4.20)$$

In a gas, the frequency factor may be computed from the Maxwell-Boltzmann distribution, whereas for macromolecular reactions happening in solvents, as is our case, the frequency factor coincides with the diffusion limited reaction rate.

The activation barrier associated with base pairing events is related to hydrogen bonds. Molecular Dynamics simulations on DNA base pairing [42] show the presence of a (small) potential barrier associated with the energy required to displace solvated water molecules from the nucleobases sites in order to form the new hydrogen bonds relative to base pairing. This energy barrier has been found to be on average $E_b \sim 3kT$.

From the above considerations, we see that in a more detailed model of DNA Hybridization, the zipping rate would depend on whether the base pair is G · C or A · T. The natural conclusion is that even if G · C rich sequences are more stable, they are also slightly slower in hybridizing than A · T rich sequences, due to having three hydrogen bonds instead of two. For now we limit our description to the averaged activation barrier $E_b \sim 3kT$.

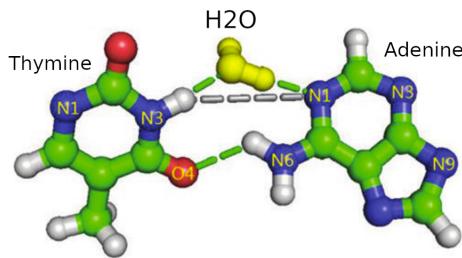


Figure 4.3: A water molecule impedes the formation of a hydrogen bond in an A · T base pair, slowing down the process. Figure from [42]

is obtained in our model by abandoning the ideal chain approximation and introducing considerations on the real chain behavior of ssDNA. *Flory Theory* [36] accounts for real chain effects such as intra-chain repulsion energy between monomers and deformation entropy loss. The general result of *Flory theory* is that polymer size is universally power-law dependant on the number of monomers as $R \sim N^\nu$, which is in accordance with [40]. For a detailed review on such topics the reader can refer to the following review [41]. We'll continue our treatment in the ideal chain approximation, which doesn't introduce much error since our modelling endeavour is aimed at the hybridization of short single stranded DNA in the LIDA system.

Through the well known Arrhenius relation with A frequency factor and E_b activation energy $k = Ae^{E_b/kT}$ we compute the rate of nucleation:

$$k_n^+ = k_{nuc}(N, \theta, \phi, T) \exp\left(-\frac{E_b}{kT}\right) \quad (4.21)$$

We finish our considerations on the nucleation reaction by calculating the backward nucleation rate as (see section 2.2):

$$k_n^- = k_n^+ \exp\left(-\frac{\mu \Delta G_{nuc}}{RT}\right) \quad (4.22)$$

Here ΔG_{nuc} is an average free energy over all possible stacked base pair energies and μ is the *minimum nucleation factor*, an integer used to define the threshold over which a nucleation is considered stable and could go on zipping.

4.3 Modeling the zipping rate

Zipping is the second stage of hybridization of DNA. It involves the processive joining of subsequent base pairs until complete formation of the duplex. It is a faster process than nucleation [41].

Experimental studies cannot help us determine the zipping rate directly, as these are primarily focused on establishing the second-order character of hybridization due to its bi-molecularity. The zipping process is a simplified model of the second stage of DNA hybridization, but it is sufficiently accurate for short strands where bubble formation can be neglected⁴.

The zipping of one base pair is related to the diffusive motion of monomers immediately following the last joined base pair (monomers in position $n+1$ when position n is base paired). The joining of a new base pair is moreover slowed down by an activation barrier relative to the displacement of solvating water molecules, as discussed for nucleation.

We can calculate the diffusion-controlled rate for two DNA monomers by using the long time limit Smoluchowski diffusion-controlled rate (4.6), where we calculate the diffusion constant through the standard Stokes-Einstein relation $D = kT/6\pi\eta r$ where the radius r refers to the hard-core radius of a DNA nucleotide and has value $r = 1\text{ nm}$. The activation barrier is equal to the activation barrier defined for nucleation $E_b = 3kT$.

We get the rate of base pair formation as an Arrhenius rate with frequency factor A

⁴Bubble-dependant effects described by the Poland-Scheraga model are only relevant for modeling the hybridization of long DNA single strands [41, 43]

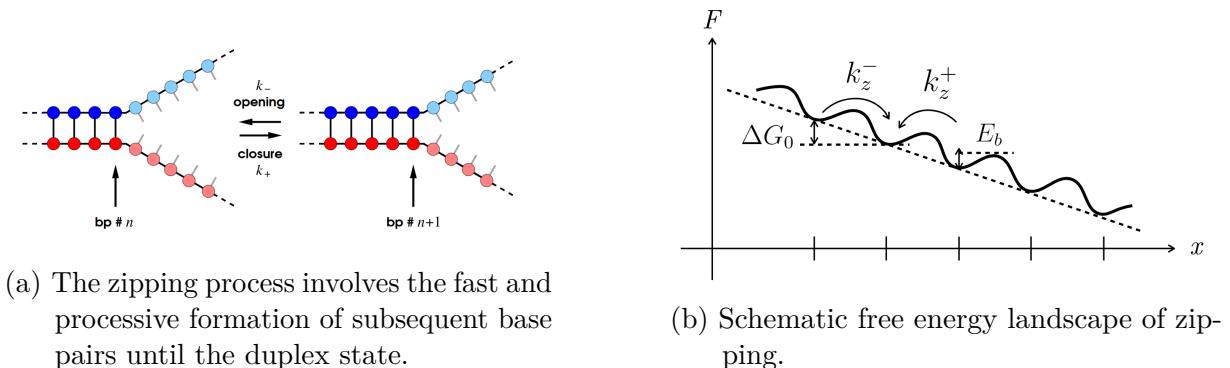


Figure 4.4: (b) The x axis refers to the number of formed base pairs, the F axis refers to the complex free energy. ΔG_0 refers to the increase in free energy due to the joining of a new base pair. E_b refers to the activation energy to make a new base pair, related to solvation (see *Activation Energy* paragraph of section 4.2)

equal to the diffusion-controlled reaction rate between two monomers with equal radius r and with equal diffusivity D , which in this special case reduces to (4.24) below:

$$k_{bp}^+ = k_{bp,dl} \exp\left(-\frac{E_b}{RT}\right) \quad (4.23)$$

$$k_{bp,dl} = 4\pi(D + D)(r + r) = \frac{8}{3} \frac{kT}{\eta} \quad (4.24)$$

An objection on such formulas, comes by saying that these represent a bimolecular reaction, while we are considering zipping as an unimolecular reaction from the nucleation complex to the duplex complex.

By simple dimensional analysis of such formulas, we indeed get a rate measured in $(volume)/(time)$ while for an unimolecular reaction rate we need to have $1/(time)$.

Regarding such correct observations, we argue that the joining of two base pairs is a bimolecular reaction, if seen from the perspective of participating monomers. This is an unimolecular reaction only from the point of view of the whole complex, in the sense that the complex is a unique entity that is transitioning from one state to another without changing the overall number of species in the system.

In order to transform the bimolecular joining of two base pairs in a volume-independent unimolecular frequency, we calculate the number concentration of the two monomers inside the spherical volume they inhabit.

Generally, the number concentration n_m of N_m molecules in a volume V is simply given by $n_m = N_m/V$. We can calculate the volume inside which the two monomers are constrained by considering that the monomer length in double stranded DNA is half the one in the single stranded state $l_h = l_d/2 = 0.334 \text{ nm}$. We are here assuming that

the monomer length tends to l_h as the zipping process goes onward. By calculating the spherical volume with radius l_h and get the number density n_m of zipping monomers:

$$V = \frac{4}{3}\pi l_h^3 = 0.161748 \text{ nm}^3$$

$$n_m = 2/V = 1.2365 \cdot 10^{25} \text{ dm}^{-3}$$

Which if plugged into the base pair formation rate k_{bp}^+ results into:

$$k_{bp}^+ = n_m \frac{8}{3} \frac{kT}{\eta_{H_2O}} \exp\left(-\frac{E_b}{kT}\right) = 7.793 \cdot 10^7 \frac{1}{\text{s}} \quad (4.25)$$

$$k_z^+ = \frac{32}{\pi l_d^3 N \eta_{H_2O}} \exp\left(-\frac{E_b}{kT}\right) = 7.793 \cdot 10^7 \frac{1}{\text{s}} \quad (4.26)$$

with water viscosity $\eta_{H_2O} = 0.8701 \cdot 10^{-3} \text{ kg(ms)}^{-1}$, at a temperature $T = 26^\circ\text{C}$ and activation energy $E_b = 3kT$.

Finally, if B is the number of base pairs to be formed from the nucleation complex to the duplex complex, the cumulative zipping rate will be [41]:

$$k_z^+ = \frac{k_{bp}^+}{B} \quad (4.27)$$

4.3.1 Example results

The model so far described is largely based on established knowledge from polymer physics, reaction-rate theory and detailed theoretical understanding of DNA hybridization kinetics. It correctly reproduces orders of magnitude answers on hybridization rates, but cannot be used for the accurate prediction . For short DNA oligomers, these are experimentally found to range from $1 \cdot 10^6$ to $5 \cdot 10^7$ [26, 44]. As a short example, we calculate k_h for a 15-mer, by setting our parameters to be:

$$\theta = \phi = \frac{\pi}{2}$$

$$\mu = 1$$

The hybridization rate equals:

$$k_h = 4.763 \cdot 10^6 \frac{1}{M \text{s}}$$

The description of DNA hybridization so far reproduces orders of magnitude results. We fail to characterize the wide variability of experimentally measured DNA hybridization rates, effectively ranging from order 10^5 to order 10^7 . This wide variability crucially depends on more complex aspects of the process.

4.4 The influence of membrane anchoring on DNA hybridization

We now modify the two-step model so far described to start peeking at the orders of magnitude effects that membrane-anchoring of DNA-lipid conjugates can have on the kinetics of DNA Hybridization.

We stress here the fact that no experimental data is available on membrane anchored DNA hybridization kinetics⁵.

We assume that surface anchoring *doesn't* affect the structure of the core processes of DNA hybridization, namely Nucleation and Zipping, but only affects the kinetics of these steps. In particular we hypothesize here that two-dimensionality *only* affects the nucleation step. Even with membrane anchoring, the zipping step still involves the motion of monomers in three dimensional space and in the same viscous fluid (water) as it was in three dimensions.

We'll have to consider how then two-dimensionality affects the nucleation step, and see what's the overall effect of it on the observed kinetic rate constant. To do that, we start by considering diffusion-controlled reaction kinetics in two dimensions.

4.4.1 Nucleation Rate of Membrane Anchored DNA

The logic behind model building here is the same as in the three dimensional case. Nucleation is a sterically hindered diffusion-controlled reaction with a very low activation energy. We then need to understand *diffusion-controlled reactions in two dimensions*, and how the steric hindrance is modified by membrane anchoring.

It has been suggested [45] that reactions between ligands and cell surface receptors can be speeded up by nonspecific adsorption of the ligand to the cell surface followed by two-dimensional surface diffusion to the receptor, a mechanism referred to as "reduction-of-dimensionality" (RD) rate enhancement.

The reduction of dimensionality enhancement depends mostly on the interplay between slower surface diffusion on membranes and search efficiency increase in 2D space against 3D space.

⁵One could in fact search the literature on DNA hybridization kinetics *on surfaces* and actually find a plethora of experimental results. Unfortunately, what in the literature is known as *Surface DNA Hybridization* is the case of a *surface immobilized probe* that hybridizes with a *solubilized target*, and is in fact modeled as an adsorption reaction on fixed sites represented by the surface probes, usually covalently bound to the surface of gold nano-particles.

Our scientific interest is here different, since we need to understand **chemical kinetics in two-dimensional space**.

It is also affected by adsorption stability, which as explained in 3.4, we can readily approximate to a probability $\mathbb{P} = 1$ of anchoring, e.g. all DNA-lipid conjugates will be membrane anchored. This helps us to greatly simplify our investigation by reducing it to a pure comparison between diffusion-controlled reaction rates.

Diffusion-controlled kinetics in two-dimensional space

We consider here two main results on diffusion-controlled reaction kinetics on surfaces. The first result comes from the adaptation of Smoluchowski [34] to a bi-dimensional case. The change in dimensionality, entails non-trivial modifications in the mathematical structure of the problem, which we'll not address here due to the great mathematical intricacy of the problem. It can be studied through the following references [46–50] and is summarized in an review by Torney and McConnell [51].

The main result can be understood as diffusing disks on an infinite plane surface, resulting in the following functional form for the diffusion-controlled rate $k_{dl}(t)$:

$$k(t) = \frac{4\pi D}{\ln\left(\frac{4Dt}{R^2}\right) - 2\gamma} \quad (4.28)$$

Where γ refers to the *Euler's constant* $\gamma = 0.5772156649\dots$ to not be confused with the well known Euler's number e related to natural logarithms and the exponential function. The result here presented is a long time expansion of the more convoluted exact result represented by the flux of density c over a circular boundary of radius $R = R_a + R_b$ denoted $\Phi(t)$:

$$\Phi(t) = k(t) = 4\pi D \left(r \frac{\partial c(t)}{\partial r} \right)_{r=R} = \frac{16Dc_0}{\pi} \int_0^\infty \frac{e^{-Du^2t}}{u[J_0^2(uR) + Y_0^2(uR)]} du \quad (4.29)$$

Where J_0 and Y_0 are order zero Bessel functions of first and second kind. For more information on such topics, the reader can refer to [50, 51].

In the case of surface reactions happening on the outer surface of a sphere, which is the case of M-LIDA, we refer to the theory of diffusion controlled reactions on spherical surfaces by Sano and Tachiya [52], readily applicable to reactions on micellar and vesicular surfaces. Taken two spheres of radius a reacting on the outer surface of a sphere of radius b and diffusing on it with diffusion coefficient D , the mean reaction time $\langle\tau\rangle$ is given by:

$$\langle\tau\rangle = \frac{(a+b)^2}{D} \left[\frac{2}{1 - \left(\frac{a}{a+b}\right)^2} \ln\left(\frac{a+b}{a}\right) - 1 \right] \quad (4.30)$$

The existence of a spatially limited spherical surface removes several layers of mathematical complexity in the treatment of the problem and also eliminates the time dependence on the rate. We'll see how to calculate the diffusion coefficient D in the next section.

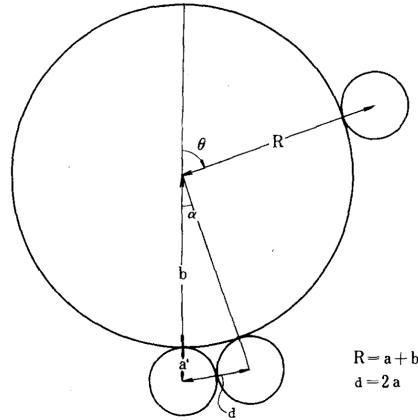


Figure 4.5: Geometric representation corresponding to the Sano-Tachiya diffusion-controlled rate on a spherical surface. We'll use in our description $a = R_g(n)$ gyration radius of a strand with n nucleotides. Figure from [52]

As in the three-dimensional case (4.7) we define the reaction radius a as the spherical volume of radius equal the gyration radius of the involved membrane-anchored DNA, hence $a(n) = R_g(n)$ as in (4.15). Moreover as in [53], we can approximate the gyration radius of a membrane-anchored DNA to the corresponding free-in-solution value one for strands long less than 40 nucleotides.

We then get from equation (4.30) an average time of collision. To get a kinetic rate constant, we can simply take the inverse. Moreover to use it in the context of surface molar concentrations⁶ measured in mol/dm^2 , we multiply by N_A the Avogadro number and S a reference surface value of 1 dm^2 , resulting in the diffusion-controlled reaction rate on a spherical surface k_s :

$$k_s = \frac{N_A S}{\langle \tau \rangle} \exp\left(-\frac{E_b}{kT}\right) \quad (4.31)$$

Where N_A refers to the Avogadro number and E_b to the minor activation energy contribution as explained in the 3D case 4.2.

⁶We measure molar surface concentration in units of $(\text{mol})/(\text{area})$ and not $(\text{mol})/(\text{volume})$ as in the three dimensional case. volume rates are usually measured in $\frac{1}{\text{dm}^3 \text{s}}$, there is no standard way in the literature on how to measure surface rates and concentrations. We choose to keep our description as similar as possible to the 3D one, by measuring surface rates in $\frac{\text{mol}}{\text{dm}^2 \text{s}}$ and surface concentrations in $\frac{\text{mol}}{\text{dm}^2}$

With respect to the three-dimensional case, we anticipate the following observation: in a purely diffusion-limited case where we don't consider steric constraints or activation energy and for parameters of interest⁷, we find a striking orders of magnitude difference in collision rates (i.e. purely diffusion-controlled rates).

By denoting c_2 and c_3 the collision rates in two and three dimensions, we find:

$$c_2 \propto 10^{14} \text{ } (M \text{ s})^{-1}$$

$$c_3 \propto 10^9 \text{ } (M \text{ s})^{-1}$$

This observation hints at the following question: how much does this orders of magnitude difference in collision rates affect the resulting hybridization rate constant?

We start providing a preliminary first answer to such question at the end of the present chapter, and start developing a suspect for a possible *acceleration* in kinetic rates of DNA hybridization.

Diffusivity of a membrane-anchored DNA-lipid conjugate

In order to evaluate the surface diffusion constant of a membrane anchored DNA, we first consider that the translational diffusion coefficient of a particle with characteristic radius a bound to a membrane with surface viscosity μ and surrounded by water (viscosity η_{H_2O}) is essentially independent of the particle shape and size if the membrane is viscous enough. i.e., when:

$$\frac{\mu}{a\eta_{H_2O}} > 4 \quad (4.32)$$

Condition easily satisfied for typical membrane viscosity values $\mu \sim 5 \cdot 10^{-10} \text{ Pa s m}$ and water viscosity of $\eta_{H_2O} = 0.8701 \cdot 10^{-3} \text{ kg (m s)}^{-1}$ and a equal to the gyration radius of a 15-mer $R_g(15) = 1.34 \text{ nm}$, getting $\mu/(a\eta_{H_2O}) \simeq 428$.

We can then reduce the centre of mass diffusivity of a membrane anchored DNA to the diffusivity of the anchor in the membrane. This can be calculated through the Saffman-Delbrück theory of Brownian motion in biological membranes [54].

A key quantity in the Saffman-Delbrück theory is the so called *Saffman-Delbrück reduced length* ϵ , which is equal to the ratio:

$$\epsilon = \frac{r_a}{L_{SD}} = \frac{a\eta_{H_2O}}{\mu} \quad (4.33)$$

Which we already know to be capped at $< 1/428$ in our case, since the radius of the anchor $r_a < R_g$ is smaller than the gyration radius of the membrane anchored DNA.

⁷e.g. viscosity of water/membrane, diffusion rate constants, equivalent radii

We can then particularize our description in the asymptotic behavior described when $\epsilon \ll 1$, for which the translational diffusion of a lipid is well approximated by the Saffman-Delbrück translational diffusion constant D_T^{SD} :

$$D_T^{SD} = \frac{kT}{4\pi\mu} \left(\log\left(\frac{2}{\epsilon}\right) - \gamma \right) \quad (4.34)$$

With γ always referring to Euler's constant. Calculations lead to membrane diffusivities of about two orders of magnitude lower than in solution. As an example, the Saffman-Delbrück diffusivity at a temperature $T = 26^\circ C$ and with the previously defined values for viscosity and with anchor radius $r_a = 2nm$, we get a diffusivity $D_T^{SD} \simeq 3.34 \cdot 10^{-6} \frac{cm^2}{s}$ which is two orders of magnitude lower than the corresponding diffusivity in water of $D_T^Z \simeq 3.29 \cdot 10^{-4} \frac{cm^2}{s}$. These results are in accordance with experimentally determined values for the surface diffusion of membrane anchored DNAs [55].

Steric constraints in the membrane-anchored case

Our last considerations are devoted to the **steric constraints**. Since the surface imposes orthogonal directionality in the inclination of anchored DNA we assume that the inclination angle ϕ can be neglected. In the two dimensional case the only steric constrain considered will be the azimuthal angle θ . We get a steric factor $\rho^{2D}(\theta)$:

$$\rho^{2D}(\theta) = \left(\frac{\theta}{2\pi} \right)^2 \quad (4.35)$$

Resulting membrane-anchored hybridization rate

We combine all our previous results to get the hybridization rate for membrane-anchored DNA. We call this k_m .

We first compute the forward nucleation rate for membrane-anchored DNA. First consider membrane diffusion constant as in Saffman-Delbrück (4.34), this is plugged into the Sano-Tachiya average collision time for spheres of radius equal to the gyration radius of the anchored DNA (eq. (4.15)) and sphere radius equal to an indicative $1\mu m$ approximating the radius of lipid vesicles.

The obtained collision time is then converted into a surface-molar rate constant through (4.31) which also considers the presence of a small activation barrier of $\sim 3kT$ as in the three-dimensional case.

This is then scaled down by surface steric constraint in (4.35). Denoting n the number of nucleotides in the strand and θ the azimuthal steric angle and S_{ref} a $1dm^2$ reference surface:

$$k_m(n, \theta) = \rho^{2D}(\theta) k_s(n) = \rho^{2D}(\theta) \frac{N_A S_{ref}}{\langle \tau \rangle_{st}} \exp\left(\frac{E_b}{kT}\right) \quad (4.36)$$

With $\langle \tau \rangle$ as in (4.30) with diffusing spheres radius equal to gyration radius of anchored DNA polymers, obtaining a rate measured in units of mol/dm^2

4.5 Preliminary results on membrane-anchored DNA hybridization kinetics

We can now get a preliminary result on membrane anchored DNA hybridization kinetics and confront this order of magnitude value against the three dimensional one.

After computation of the new membrane-anchored nucleation rate as discussed in the previous section, we get the observed reaction rate as in (4.5). In order to combine nucleation and zipping into (4.5) we need to translate the two-dimensional nucleation rate of (4.36) into an equivalent surface-rate in terms of volume molar concentrations. To do this we recall from section 3.5 the result reported in equation (3.16) that achieves such transformation.

For the protocellular system discussed in the Introduction it is of particular interest to calculate the volume to surface ratio V/S for systems consisting of vesicle anchored DNA co-factors. Consider a mixture of decanoic acid lipids in a volume V with volumetric concentration c_{dl} . For decanoic acid the *Critical Vesicle Concentration* (CVC), the point at which lipids start forming vesicles, is $cvc = 6.75 mM$. It is observed [3] that all the exceeding concentration of lipids over the CVC will form vesicles, such that the lipids concentration involved in vesicles will be $c_v = c_{dl} - cvc$ if $c_{dl} > cvc$ and $v = 0$ else.

Assuming the formed vesicles are unilamellar (only one double layer) then the number of vesicle forming fatty acids in the outer vesicle layer is the total number of fatty acids divided by two. Further, if the surface area of the polar head of a decanoic acid lipid molecule is $a_l = 37.3\text{\AA}^2$ we can estimate the total external vesicle surface area. If we start from a volume of $V = 1dm^3$, we get a total surface area of $S = 9.2671 \cdot 10^4 dm^2$ if we account a total concentration of DNA-lipid conjugates of $c_{dna} = 10 \mu M$ and a total concentration of lipids $c_{dl} = 15 mM$. The result is obtained by considering equation:

$$S = \frac{1}{2} V N_a a_l c_v = 9.2671 \cdot 10^4 dm^2 \quad (4.37)$$

Resulting in a total volume to surface area ratio

$$\frac{V}{S} = 1.08 \cdot 10^{-5} \text{ dm} \quad (4.38)$$

We use this result to translate the membrane nucleation rate into its equivalent volume molar concentration rate for this specific case of interest. The three-dimensional equivalent k_{3m} of the membrane-anchored two-dimensional nucleation rate k_m is obtained as in equation (3.16):

$$k_{3m} = k_m \frac{V}{S} \quad (4.39)$$

which can now be plugged into (4.5). Denoting then k_{3m}^+ and k_{3m}^- rates for forward and backward nucleation obtained by detailed balance as in section 2.2 we get the two-dimensional hybridization k_h^{2D} rate as:

$$k_h^{2D} = \frac{k_z^+ k_{3m}^+}{k_z^+ + k_{3m}^+ + k_{3m}^-} \quad (4.40)$$

In three dimensional space we previously found an hybridization rate constant of value $k_h^{3D} = 4.763 \cdot 10^6 \text{ (M s)}^{-1}$. By keeping all the shared parameters equal, namely the zipping rate $k_z^+ = 7.793 \cdot 10^7 \text{ s}^{-1}$, $\theta = \pi/2$, we arrive to the conclusion that, for this specific parameter set, the reduction of dimensionality results in a speed up of the two-dimensional DNA hybridization rate denoted k_h^{2D} :

$$k_h^{2D} \simeq 1.164 \cdot 10^7 \left(\frac{\text{mol}}{\text{dm}^3 \text{ s}} \right)^{-1} \quad (4.41)$$

This result is valid in the absence of surface crowding effects, i.e. when the surface crowding ratio:

$$\sigma = \frac{n_d a_d}{n_l a_l} \quad (4.42)$$

is small, where n_d and n_l denote the number of DNA and lipid molecules in the same vesicle while a_d and a_l denote their areas. For situations where $n_d a_d \ll n_l a_l$ (e.g. $n_d = n_l \cdot 10^{-2}$) and for number of monomers n such that the gyration radius $R_g(n) < 10 a_l$ we get at most an error of factor 2 [56]. In general we'll not consider crowding effects further in our analysis, but these will generally result in multiplicative corrective factors that depend on relative lipid-DNA concentrations of the experimental system of interest that act in non-trivial ways, such that the study of crowding is left for future works.

4.6 Consequences on LIDA and M-LIDA kinetics

We consider here the effects of the kinetic description so far developed of DNA hybridization in solution and in the membrane-anchored case. The effects on the M-LIDA design 3.4 consequential to the hybridization rates here derived are investigated by ODE simulation of LIDA as outlined in 3.2. We simulate the same DNA sequences involved as in figure 3.5.

Input hybridization kinetic parameters are as derived in previous sections 4.5 for the membrane-anchored hybridization rate constant and 4.3.1.

So obtained results show a very slight influence of dimensionality on LIDA replication half-life denoted τ , with τ_{3D} referring to the classical LIDA design and τ_{2D} referring to the membrane LIDA design.

We report:

$$k_h^{3D} = 4.73 \cdot 10^6 \text{ (M s)}^{-1} \quad k_h^{2D} = 1.164 \cdot 10^7 \text{ (M s)}^{-1}$$

$$\tau_{3D} = 59.36 \text{ min} \quad \tau_{2D} = 54.95 \text{ min}$$

Relative kinetic realizations are shown in figure below.

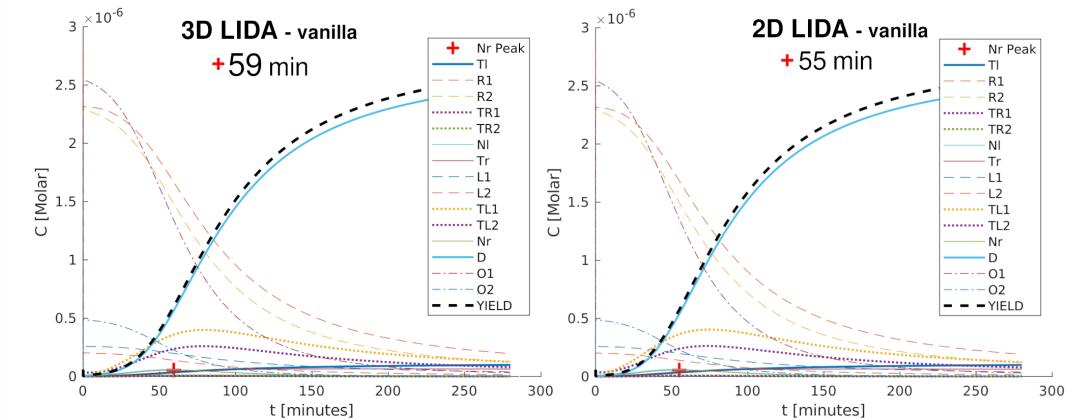


Figure 4.6: Confrontation of LIDA kinetics realizations with three-dimensional and two-dimensional hybridization rate constants. Crossed in red, the nicked duplex peak corresponding to replication half-life.

Conclusions

We conclude by underlining that this analysis provides us with an initial indication of the potential efficacy of the M-LIDA design compared to LIDA in a solution. However, it is important to acknowledge that the findings presented are limited in scope and only apply to a narrow range of conditions, considering the intricate nature of the system.

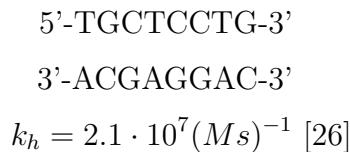
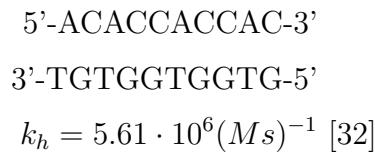
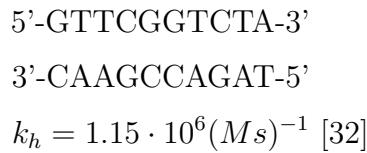
To further explore and expand upon these findings, we intend to investigate the kinetics of DNA hybridization in greater detail. Specifically, our objective is to construct a robust computational model capable of producing more accurate qualitative and quantitative predictions regarding the acceleration or inhibition of DNA hybridization kinetics resulting from the reduction of dimensionality conferred by membrane anchoring. By developing a model of this nature, we aim to attain a more comprehensive understanding of the complex mechanisms at play and uncover potential areas of optimization or improvement.

5 Transitions from misaligned meta-stable complexes

So far, the description of DNA hybridization kinetics outlined in the last chapter, could as well be applied to any imaginary linear polymer with the capacity to form duplexes, by adjusting the parameters to mimic the nature of this imaginary polymer.¹

We will from now on characterize the **sequence dependant** hybridization rates for complementary strands of the same length.

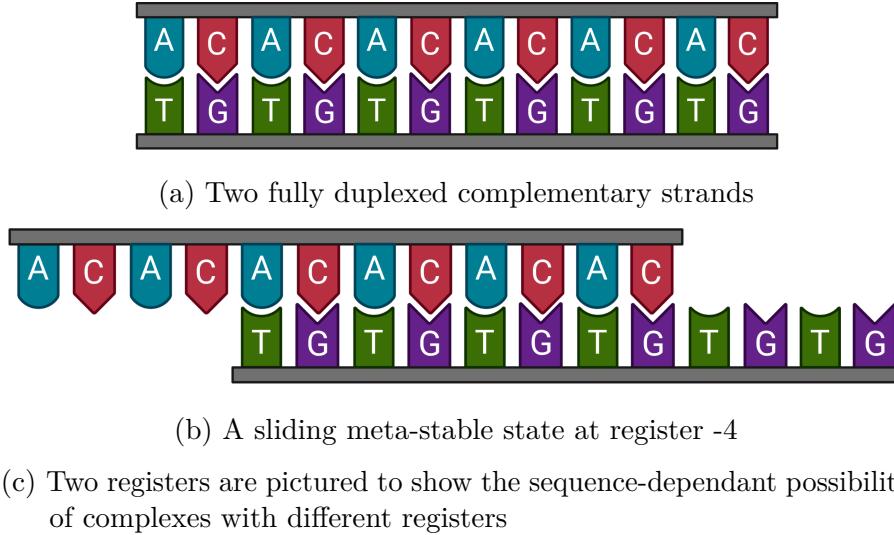
DNA Hybridization is in fact a **sequence dependant** process. As an example, three sequences are shown below to show the variability of DNA Hybridization kinetics. Such examples are ubiquitous in DNA Hybridization data and need an explanation.



We develop a description of the specific kinetic pathways of DNA Hybridization. We base our description on the early analysis of Porschke [57] who was the first to propose a third fundamental mechanism to consider over nucleation and zipping: a fast chain **sliding** reaction.

¹If the author has learned something general about Physics from during the Master of Science related to this thesis, it is that a good model is neither *too general* or *too specific* but it has to be *just right*. The author doesn't sustain the hypothesis to possess such ability, but it's quite self evident that the previous model is just *too general*

The sliding pathway involves the formation of misaligned structures, that is to say stable structures with a sequence offset with respect to the fully formed duplex. We call this sequence offset a *register*, denoted by r . It corresponds to bases pairing with complementary bases by offset r in the 5' direction. We also call an **off-register complex** a complex with $r \neq 0$. By this definition, the full duplex will correspond to $r = 0$, a positive register ($r > 0$) corresponds to an offset in the 5' direction and a negative register ($r < 0$) to an offset in the 3' direction.



More recent developments in coarse-grained molecular dynamics (MD) simulations [30, 58, 59] enabled an *in silico* view of hybridization pathways. Authors concluded two pathways are the most prominent alternative pathways of hybridization that lead from an off-register metastable state to the full duplex state: **pseudoknotting** and **inchworming**.

5.0.1 Inchworming transition

Due to thermal fluctuations, the outermost base pairs of a complex can fluctuate between the open (unzip) and closed (zip) state.

Outermost base pair opening can be followed by the next possibilities:

- 1. Opening of another subsequent base pair, continuing the unzipping process
- 2. Closing of the same base pair, returning to the previous state.
- 3. Formation of a new base pair with a nucleotide on another register, due to conformational fluctuations of the unpaired tails.

In case 3. a bulge/interior loop is formed (see sec. 2.1). In such case, bases adjacent to it will undergo a rapid base pair formation and dissociation process.

As a consequence of this rapid equilibration, a bulge/interior loop moves quite rapidly along the complex. The probability for bulge/interior loop movement is equal for both directions [57].

If a bulge/interior loop is formed at one end of a helix sequence and disappears at the opposing end, the bulge/interior loop diffusion will result in a shift of the nucleotide strands by a register amount equal to the difference between the register of the initial complex and the register of the newly formed complex 5.2.

In general the propensity at which an off-register will shift to another one depends on the difference in stability between these two. The resulting tendency is for less stable off-registers to shift towards more stable ones. The likelihood of a bulge/interior loop to form will be hindered by the register distance between off-registers, i.e., given an initial sliding with register r_i and a target slide with register r_j , the probability p_{iw} of initiating an inchworming process will be proportional to $p_{iw} \propto 1/|r_i - r_j|$.

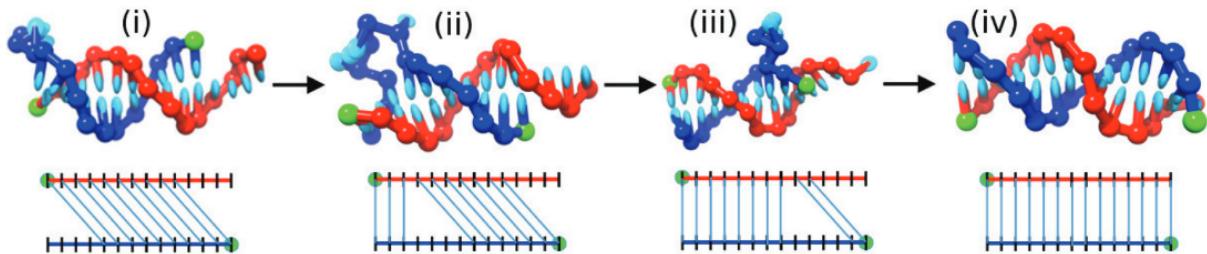


Figure 5.2: Molecular Dynamics simulation of an inchworming transition from the initial slide with register $r = 4$ to the duplex state $r = 0$.

(i) Initial binding of 10 base pairs at register 4. (ii) Owing to thermal fluctuations, three base pairs form a register 0 at the expense of register 4. (iii) The bulge loop is passed down the helix. (iv) After the final base pair in register 4 breaks, register 0 is able to form all its possible base pairs. Figure from MD simulations performed by Ouldridge [58]

5.0.2 Pseudoknotting transition

Taken n the total number of nucleotides of two hybridizing strands, short off-register complexes with $0 \ll |r| \leq n - 1$ will have very long single stranded tails. These tails can nucleate with each other and form new base pairs on a different register. In this case two registers will be formed at the same time, and will energetically compete with one another, resulting in the most stable one displacing the less stable.

Due to the symmetry of the tails, pseudoknotting transitions mostly happen between off-register complexes with register of opposite sign, namely $\text{sgn}(r_i) \neq \text{sgn}(r_j)$ for r_i and r_j the registers of the two off-registers involved. Direct pseudoknotting towards the duplex state is also permitted if the newly formed bases are at the correct register $r_j = 0$.

This results in a complete displacement of the old off-register complex and the direct formation of the duplex (fig 5.3).

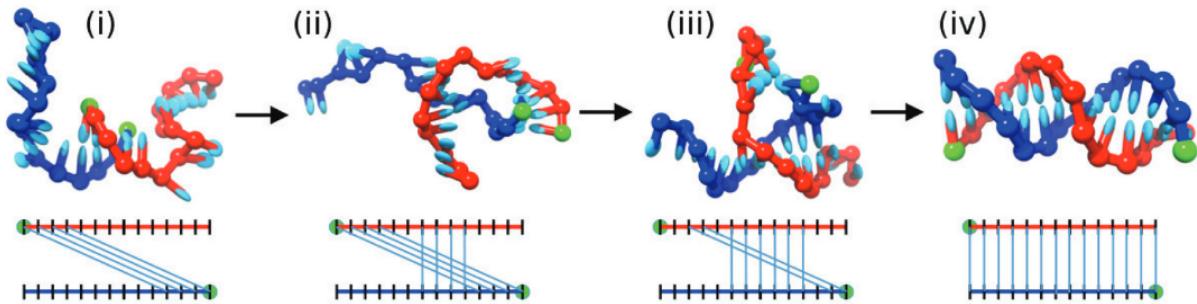


Figure 5.3: Molecular Dynamics simulation of an inchworming transition from the initial slide with register $r = 4$ to the duplex state $r = 0$.

(i) A complex is initially in register 10. (ii) Additional base pairs form at register 0. (iii) Further base pairs form at register 0 at the expense of register 10. (iv) The final base pairs at register 10 are broken and the duplex forms. Figure from MD simulations performed by Ouldridge [58]

5.1 Characterization of off-register transitions

We characterize here more formally the transition rates from an off-register complex ω to the duplex state at register $r = 0$.

We start by outlining here all the approximations that we'll apply in our model.

- We will only consider hybridization between strands with the same total number of nucleotides
- We will only consider off-register complexes c with at least one consecutive stacked base pair (i.e. at least two base pairs)
- We will only consider stable off-register complexes c with negative free energy $\Delta G(c) < 0$.
- We will neglect the role of intrastrand secondary structure (i.e. hairpins and intrastrand pseudoknots)²

We propose here that pseudoknotting and inchworming are two parallel pathways that an off-register complex can undergo to transition to the duplex state. We provide here a more formal characterization of inchworming and pseudoknotting from a kinetic standpoint.

²This limits our descriptive power only to short single strands with sequences such that hybridization between complementary intrastrand subsequences is not possible or is not more stable than the single stranded state.

Given two hybridizing strands σ_1 and σ_2 and their structure space Γ (chap. 2), our goal is to derive sliding transition probabilities from all possible off-register complexes to the duplex, to be later implemented in the Zipping Graph Model in the next chapter.

We define the **sliding rate** $k_s(\omega)$ of an off-register complex ω towards the duplex complex to be proportional to its stability. We expect a more stable sliding to have higher transition likelihood due to a higher permanence time in a state from which the whole system can transition to the duplex state.

$$k_s(\omega) = \kappa_s \exp\left(-\frac{\Delta G(\omega)}{RT}\right) \quad (5.1)$$

κ_s is a free parameter we call **basal sliding rate** that indicates the baseline rate of sliding transitions and is in all effects just a constant multiplicative parameter that will be fitted against data.

5.1.1 Inchworming

Inchworming can be initiated in two ways. Either two parallel tails are not bounded and can nucleate on a new register, or due to zip-unzip fluctuations of the outermost base-pairs, these will re-zip with a complementary base on another register. Both cases result in bulge/interior loop diffusion and shifting if the new register is more compactly connected and stable than the original one. Refer to figure 5.4 for an explanatory example.

We postulate an initial conjecture here: The probability of an off register complex ω at register r_ω to transition to the duplex state is inversely proportional to the absolute value of the register:

$$\mathbb{P}_{iw}(\omega) = \frac{1}{r_\omega} \quad (5.2)$$

connected to the *sliding rate* $k_s(\omega)$ of off-register complex ω at register r_ω by:

$$k_{iw}(\omega) = \mathbb{P}_{iw}(\omega) k_s(\omega) = \frac{k_s(\omega)}{r_\omega} \quad (5.3)$$

5.1.2 Pseudoknotting

We calculate the pseudoknotting probability by considering the gyration radius R_g (see eq. (4.15) sec. 4.2) associated to the single stranded tails of each complementary strand and compute the equivalent spherical volume for each. We fix the distance between the centres of such spheres to be equal to the length of the base paired section of the off-register complex and evaluate the volume of the overlapping region at the intersection between those spheres. A probability is obtained by normalizing the intersection volume by the union volume of the two spheres, an illustrative example is found in figure 5.5.

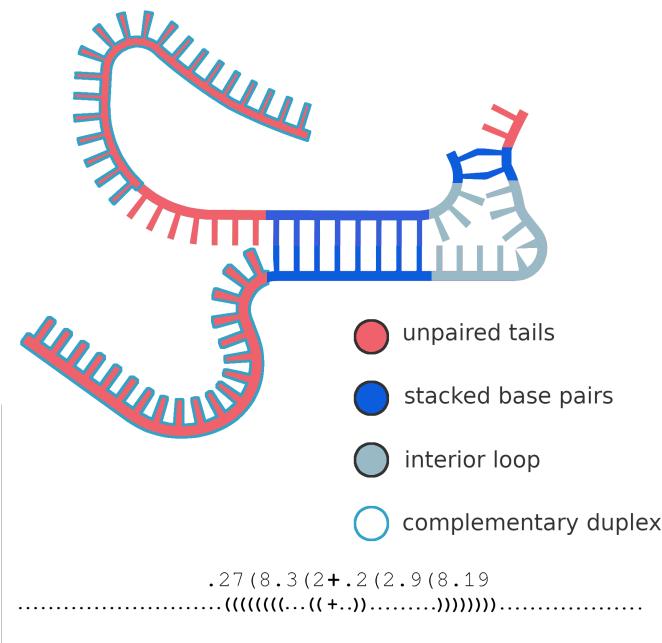


Figure 5.4: Two complementary strands long 40 nucleotides have a stable off-register complex at register $r = -8$. The **left side** tails, long 27 nucleotides for the top strand and 19 nucleotides for the bottom strand, can undergo direct on-register nucleation, initiating the transition from the off-register $r = -8$ to the on-register duplex $r = 0$. On the **right side** two complementary base pairs form at register $r = -2$, creating an *interior loop* (see sec. 2.1). **Below** two equivalent secondary structure notations referring to the pictured complex: **top)** run-length-encoded dot-bracket notation, **bottom)** plain dot-bracket notation

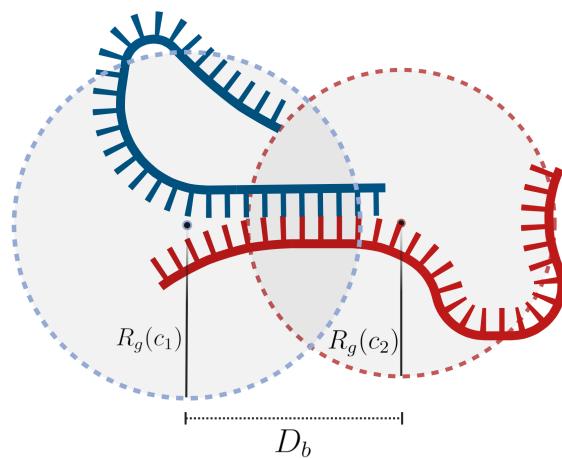


Figure 5.5: Two dimensional representation of the overlapping spherical caps representing the pseudoknotting probability. D_b is the duplexed segment length, $R_g(c_1)$ and $R_g(c_2)$ refer to gyration radiiuses of single stranded tails.

Consider an off-register complex ω with a paired segment long b nucleotides and single stranded tails long c_1 and c_2 nucleotides. The physical length of these traits is $D_b = b l_d/2$ for the paired segment and $R_g(c_1) = c_1 \frac{l_d l_p}{3}$, $R_g(c_2) = c_2 l_d$ for single stranded tails. We remind that $l_d = 0.676\text{ nm}$ is the monomer length of single stranded DNA and $l_d/2 = 0.338$ is the monomer length in double stranded DNA. $l_p = 2.223\text{ m}$ is the persistence length of single stranded DNA (see (4.9) in sec. 4.2).

Associating V_1 and V_2 to the spherical volumes relative to the gyration radiiuses $R_g(c_1)$ and $R_g(c_2)$, we get the union volume $V = V_{TOT} - V_{CAP}$ where $V_{TOT} = V_1 + V_2$ and V_{CAP} refers to the volume of the overlapping spherical caps, only possible if $D_b \leq R_g(c_1) + R_g(c_2)$. Denoting with $\mathbb{P}_{pk}(\omega)$ the probability of pseudoknotting of off-register complex ω :

$$\mathbb{P}_{pk}(\omega) = \begin{cases} \frac{V_{CAP}(\omega)}{V_{TOT}(\omega) - V_{CAP}(\omega)}, & \text{if } D_b > R_g(c_1) + R_g(c_2) \\ 0, & \text{else} \end{cases} \quad (5.4)$$

The pseudoknotting sliding rate k_{pk} of off-register complex ω will then be:

$$k_{pk}(\omega) = \mathbb{P}_{pk}(\omega) k_s(\omega) = \mathbb{P}_{pk}(\omega) \kappa_s \exp\left(-\frac{\Delta G(\omega)}{kT}\right) \quad (5.5)$$

5.1.3 Cumulative sliding rate

As explained at the beginning, inchworming and pseudoknotting are two parallel pathways that an off-register state can undertake. This means that the cumulative sliding rate, denoted by a tilde $\tilde{k}_s(\omega)$ of off-register complex ω is the sum of its corresponding pseudoknotting and inchworming rates:

$$\tilde{k}_s(\omega) = \kappa_s \left(\mathbb{P}_{pk} + \frac{1}{r_\omega} \right) \exp\left(-\frac{\Delta G(\omega)}{kT}\right) \quad (5.6)$$

The overall frequency of sliding will depend on the free parameter κ_s called **basal sliding rate**. The results of the current chapter as well as the previous chapter, will be the physical foundation behind the Zipping Graph Model, which will be discussed in the next chapter.

6 Zipping graph model of DNA hybridization

The observed kinetics of many complex physical processes can be described by kinetic transition networks, or **kinetic graphs**, where transitions represent *elementary transitions* [60] which are more easily described by standard physical chemical knowledge.

In kinetic graphs, discrete well defined chemical states correspond to nodes, whose edges encode the underlying *elementary* transitions. Moreover, for quasistatic reversible transformations, the Markov approximation holds and transitions between states are taken to be independent random processes. The quasistatic condition is an assumption of local equilibrium, meaning that the equilibration following any reaction event is much faster than any reaction time scale.

We construct here a computational model, which combines the physically-grounded description of the kinetics of transitions involved in DNA hybridization (chap. 4 & 5) with the ACTG-alphabet based combinatorial nature of DNA Hybridization.

We call such a computational model **Zipping Graph Model** of DNA Hybridization.

Free Parameters

The zipping graph model depends on **four** global free parameters:

- **Zipping Rate:** defined and calculated in section 4.3, we choose to keep it as a free parameter in order to confirm or invalidate zipping rate calculations performed in the referenced section.
- **Basal Sliding Rate:** referred to equation 5.1 of section 5.1.
- **Steric angles (θ, ϕ):** defined in section 4.2.3, in particular refer to figure 4.2 therein.

6.1 Snapshot of the model

We construct an algorithmic generator of a sequence-dependant kinetic digraph. Stochastic simulations are performed over discrete node states with transition rates governed by laws depicted in chap. 4 and 5. The observed rate constant of DNA hybridization kinetics is observed from the Mean First Passage Time of trajectories.

Given two complementary input sequences of same length L , the ZG model computes the sequence-dependant Minimum Free Energy (MFE) secondary structure λ_r associated with each r possible register¹ (for a definition of register refer to the introduction of chapter 5 and figure 5.1c therein). Each computed λ_r is kept if stable, namely $\Delta G(\lambda_r) < 0$. This is performed by embedding algorithms for nucleic acids thermodynamic analysis from NUPACK 4.0 [7] efficient implementation of the nearest neighbor model of DNA thermodynamics (chap. 2).

The MFE complex $\lambda_{r_{min}}$ with $r_{min} = \min_r(\Delta G(\lambda_r))$ will almost always² be the duplex state at register $r = 0$ and will be the target state of the generated kinetic graph. All MFE structures λ_r with $r \neq r_{min}$ will be connected to $\lambda_{r_{min}}$ with transition rates as outlined in chapter 5.

For all the identified MFE complexes λ_r (target $\lambda_{r_{min}}$ included), the ZG recursively generates the *zipping leaf* subgraph that leads from the simplex³ state to each λ_r . The algorithmically generated zipping leaf subgraph leads from the "root" simplex state to each λ_r by processively generating the subsequent formation of new basepairs on both physical directions of the forming complex, only if bases adjacent to already formed ones are correctly WCF couples, as shown in the figure below:

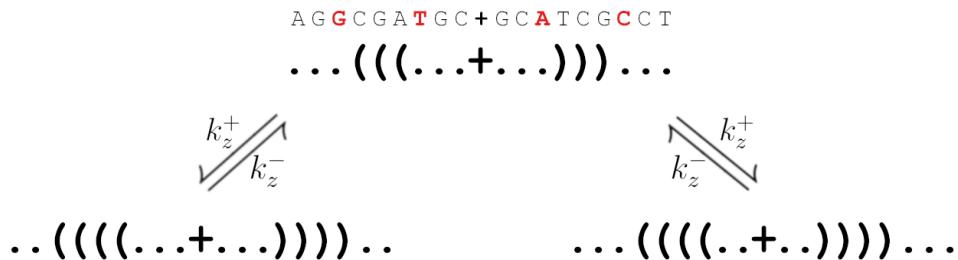


Figure 6.1: At the top, a register zero zipping state with three already formed base pairs and the two candidate base pairs (shown in red) that will form. At the bottom the corresponding structures to newly formed base pairs. Bottom left: formation of an G · C pair. Bottom right: formation of an A · T pair

Structuring the zipping connection routine on such nearest neighbor approach, the zipping leaf subgraphs naturally carry a graph structure which may be called a "square ladder graph"⁴ (fig 6.2):

¹possible registers r are bounded by total length L as $1 - L \leq r \leq L - 1$

²some very rare exceptions happen depending on sequence and stabilizing dangling ends

³short for single stranded

⁴The graph structure is similar to a Generalized Triangle Ladder Graph, but with minimum cycle length of 4.

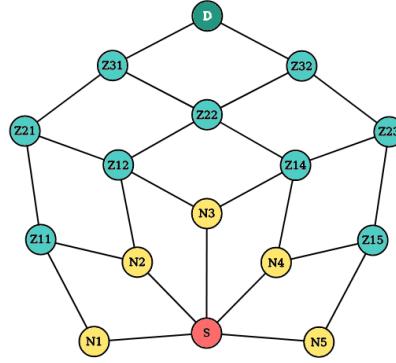


Figure 6.2: A *zipping leaf* is pictured. Originating from simplex state S (in red) with target state a structure with 5 adjacent base pairs denoted by D (in green), neighbors of S are the first 5 nucleation attempts (N_x in yellow) with only one base pair. These will subsequently zip (Z_x in cyan) one base pair at a time until reaching the target.

Connecting the simplex state to each combinatorially determined nucleation, then generating zipping leaf subgraphs for registers possessing a stable MFE λ_r and connecting each λ_r to the most stable $\lambda_{r\min}$ will generate an unique sequence-dependant zipping-based kinetic graph corresponding to the input sequences.

As a showcase example, figured below 6.3 the graph relative to the complementary sequences: 5'-AACCAACC-3' and 5'-GGTTGGTT-3'.

The resulting zipping graph is then subjected to Stochastic simulation to analyse the reaction kinetics as the system transitions from the initial simplex state to the target duplex state. We employ the Stochastic Simulation Algorithm (SSA), also known as the Gillespie Algorithm [33], allowing us to take into account the inherent randomness and probabilistic nature of the reaction pathways involved.

By performing a large number of independent simulations, we gather a statistically significant sample of reaction trajectories, which are used to calculate the Mean First Passage time (MFPT) $\langle \tau_{MFPT} \rangle$ representing the average time it takes for the system to reach the target duplex state for the first time, starting from the initial simplex state.

The observed rate of hybridization k_h is related to the MFPT $\langle \tau_{MFPT} \rangle$ as:

$$k_h = \frac{1}{\langle \tau_{MFPT} \rangle} \quad (6.1)$$

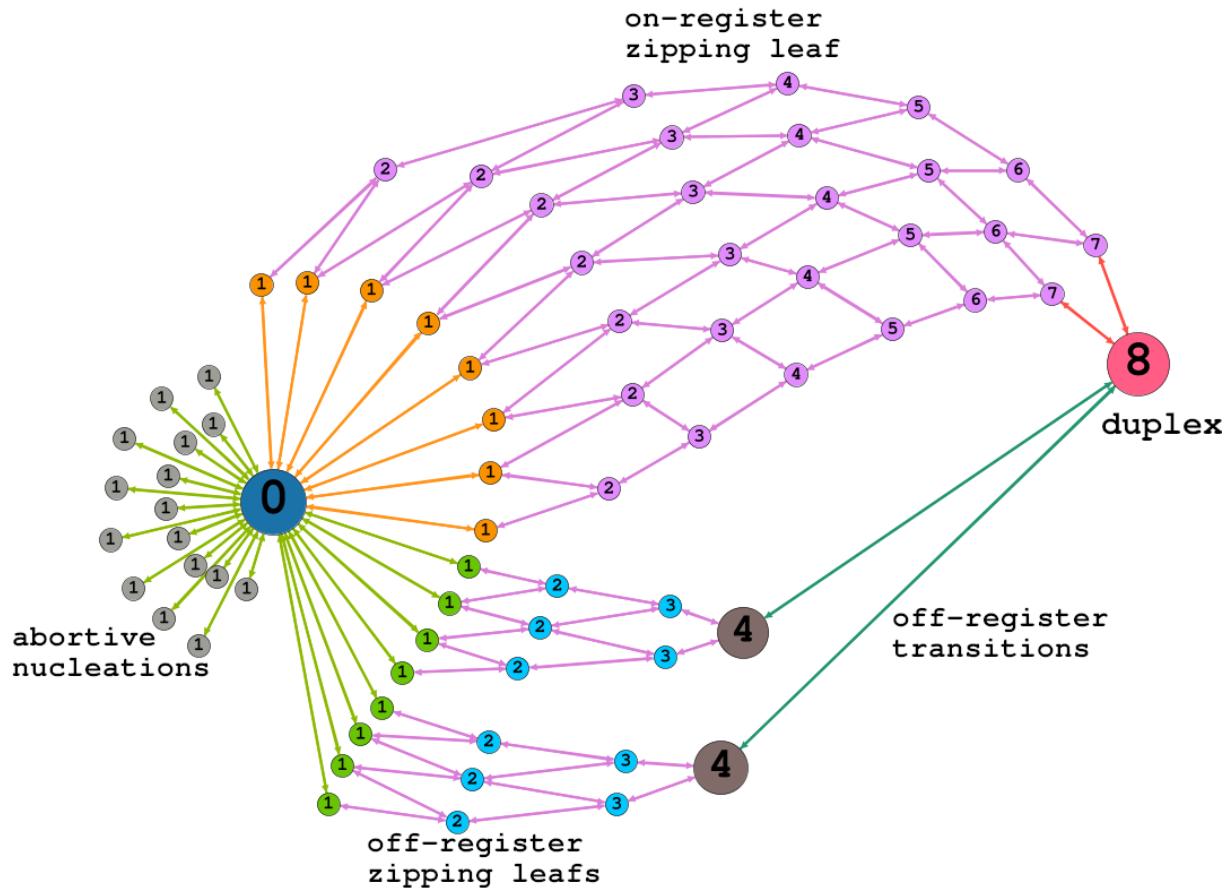


Figure 6.3: Example zipping graph generated from complementary sequences
5'-AACCAACC-3' - 5'-GGTTGGTT-3'.

Numbers inside nodes indicate the number of base pairs of the underlying secondary structure formed. Node with 0 pairs indicates the root simplex state and node with 8 pairs indicates the target duplex state.

Colours of transitions indicate the following: *green* indicates off-register nucleation events, which can either lead to subsequent zipping or be abortive, *orange* indicates on-register ($r = 0$) nucleation events, *purple* indicates zipping transitions, *dark green* indicates sliding transitions through inchworming and/or pseudoknotting pathways (chap. 5).

The computational model so described, has been subjected to optimization against experimental data, providing on-par predictive power when compared to recently developed models of DNA Hybridization kinetics [32, 61, 62] while retaining the advantages of being physically grounded and not stemming from nonphysical fitted semi-empirical parameters. This permits us to generate predictions of the two-dimensional case of DNA hybridization, not possible with any other model.

Moreover, the model helps us to gain valuable insights into the hybridization process, helping us to make informed choices on the design and optimization of DNA sequences and reaction conditions for our application of interest in the Lesion Induced DNA Amplification system 3, as well as for various applications in DNA nanotechnology.

6.2 Object-oriented implementation

The zipping graph model has been implemented as an Object-Oriented Program. Mainly in Python 3.10 with auxiliary Julia 1.9 algorithms for computationally heavy tasks. The goal is to build a flexible program that enables us to quickly and flexibly perform simulations with the bare minimum coding interface. We aim to separate the algorithmic implementation of the model to its deploy in the prediction of experimental data, moreover we aim to put the basis for the construction of an easy-to-use Python package based on the zipping graph model, to be used as a computational tool to aid the research in the fields of nucleic-acid based Artificial Life investigations and DNA nanotechnology. Below a flowchart explain our program 6.4 with the developed classes, their dependencies and their specific functions.

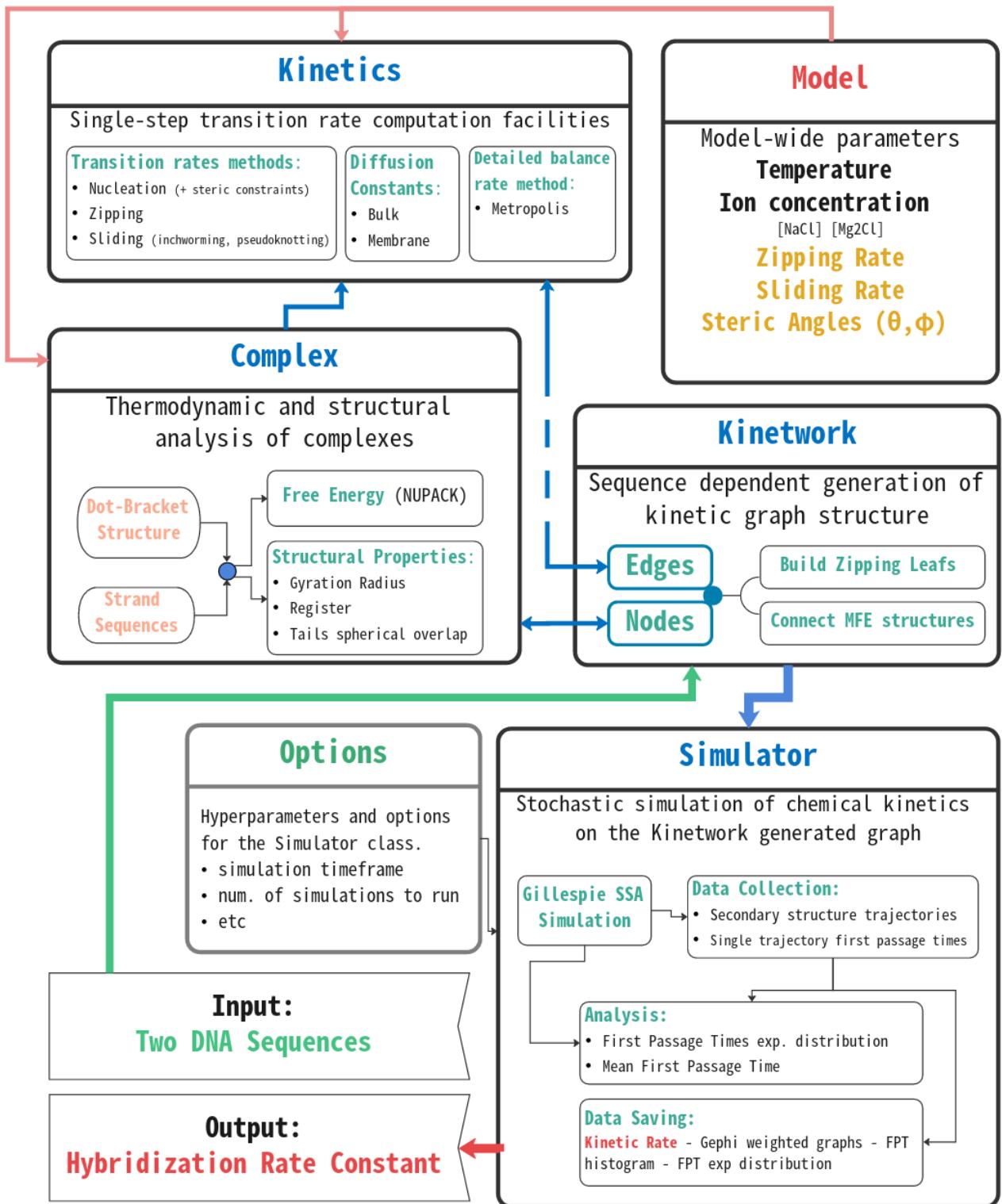


Figure 6.4: Flow chart like representation of the Object-Oriented Programming implementation of the zipping graph model. Titles of blocks refer to implemented classes, in each a description of the main tasks performed by the class. Arrows show dependencies.

6.2.1 Kinetic graph generation: `class Kinetwerk`

The construction of kinetic graphs starts from `class Kinetwerk`. It needs two inputs: *system-wide* conditions, values for the *free parameters*, and **input sequences**.

System-wide conditions are temperature and ion concentrations [$NaCl$] and [$MgCl_2$].

Free parameters are zipping rate, sliding rate and steric angles θ and ϕ .

These two are defined through the `class Model`, which is fed as input to the `Kinetwerk` class. Strand sequences are provided as simple `python strings`.

The main methods of `Kinetwerk` are `def get_leafs` and `def connect_slidings`.

If inputs are correct, graph construction starts from method `get_leafs`.

Its main core is built around three nested iterators, the first iterates a value w from 1 to L with L the length of the strand.

It is followed by two nested sliding windows that iterate strand sequences over w long windows. At each iteration sequences are compared for correct WCF base pairing and a node is added to the graph under construction if the condition is true.

Nodes added to the graph are instances of the `class Complex`, defined by two sequences and a Dot-Bracket secondary structure. This class implements the nearest neighbor model of DNA thermodynamics through the Python package NUPACK 4.0 [7] as well as polymer physics based calculations on spatial properties corresponding to the specific structure given.

During iteration, edges are then added by connecting the nodes with the appropriate neighbors at distance of one adjacent base pair difference, by memorizing nodes created at the previous iteration step. (see fig.s 6.16.2). Edges are directional and each edge is associated with its corresponding kinetic transition rate computed by calling methods from `class Kinetics`.

At this point, the graph is a collection of independent *zipping leafs* as in figure 6.2. These share only one central node corresponding to the simplex state.

Method `connect_slidings` takes care to create sliding transitions from Minimum Free Energy off-register complexes (MFEs) to the target duplex state. It first searches the MFE at minimum energy $\lambda_{r_{min}}$ to connect with all the other possible MFEs λ_r .

We note that we cannot easily separate MFEs solely on the basis of which node, for each register, corresponds to the MFE one, as doing so may leave some leafs unconnected to the duplex⁵.

⁵This may happen in the case of a register having two separated zipped structures that start zipping from nucleations at different parts of the strand.

We then implement an approach based on the topology of the network generated by `get_graph`. We disconnect the simplex state from the network, separating it in a number of subgraphs exactly corresponding to all the previously generated zipping leafs. We can now iterate over all the resulting (strongly) connected components of the digraph, finding the MFE for each of these will successfully identify all the off-register MFE complexes that can undergo sliding transition towards the duplex.

The corresponding sliding rate is calculated through the `Kinetics` class by an interplay with the `Complex` class instance of the MFEs, considering both pseudoknotting and inchworming pathways to the duplex (see chap. 5).

It needs to be made clear that, even though nucleation and zipping are well defined *elementary* kinetic transitions, the sliding of an off-register MFE to the target duplex is a *non elementary* transition that is still represented in the zipping graph as a **single step** transition, on par with elementary transitions. We are hence **approximating** a quite complex transition from an MFE to the duplex to a single step, of which rate is defined as in chapter 5.

6.3 Simulation methods in `class Simulator` and analysis of results

Stochastic Simulation is performed on the `Kinetwork` generated kinetic graph by embedding its structure in a Python-Julia algorithm with the aid of `BioSimulator.jl` [63] for efficient implementation of the SSA Gillespie algorithm. This is embedded in `class Simulator`.

The simulator is initialized by fixing a simulation time-frame T and a number of simulations N_m to be performed.

Time-frame T has been fixed throughout all of our analyses to a value of $T = 4\mu s$. This time-frame has been found to be low enough to be computationally non-expensive and to be high enough for simulations to successfully reach completion with probability ~ 0.995 . The number of simulations N_m has been chosen to minimize error with respect to data as discussed in section 6.4.

Stochastic Simulation

The simulation is initiated with the two strands A and B at initial number concentration $A_0 = 1$ and $B_0 = 1$.

We model the initial bimolecular collision of the two species separately, since simulating it inside the stochastic simulator in a volume $V = 1$ would require us to add an initial waiting time of order $\sim N_a$ Avogadro number before hybridization starts through unimolecular transitions in the secondary structure space⁶.

⁶This is due to differences between deterministic reaction kinetics on molar concentrations and stochastic reaction kinetics on number of molecules. A bimolecular molar reaction rate k_a and a bimolecular

A)	B)	C)	D)
GTTCCGGTCTA+TAGACCGAAC	ACACACACCA+TGGTGTGTTG	CCACCCAACAACAAAC+GTTGTTGTTGGTGG	CCACCCAACAACAAAC+GTTGTTGTTGGTGG

Figure 6.5: **A)** after an initial abortive nucleation at register 0, the 10-mer hybridizes on register 0. **B)** a 10-mer zips on register 2, from there it inchworms to the duplex. **C)** a 14-mer nucleates and zips one base pair at register -10 to then unzip and re-nucleate at register 0, abort, re-nucleate and then zip to the duplex. **D)** a 14-mer performs several abortive nucleations and one abortive zipping at register 3. It reaches a meta-stable off-register state at register 7 from which it transitions to the duplex. In this case both inchworming and pseudoknotting are possible transition pathways.

We model⁷ the encounter reaction between the two single strands *A* and *B* reacting into *AB* in volume *V* as: $A + B \xrightarrow{k_c} AB$

Since waiting times are exponentially distributed [33] what we do is to sample from an exponential distribution $Exp(t; \lambda)$ with scale parameter λ equal to the diffusion limited rate constant for the collision of chain-like linear polymers as in equation (4.7) and then start the kinetic graph simulation from the encountered state *AB* corresponding to the simplex state of the graph figured as the initial concentration $AB_0 = S_0 = 1$.

From the initial encounter simplex state, SSA simulation proceeds in an uni-molecular fashion and produces a trajectory through secondary structure space. These trajectories can be useful in several ways, such as finding kinetic traps or visualizing a kinetic pathway. From the trajectory, we store the time at which the first occurrence of the duplex state is registered, this will be a single First Passage Time (FPT) point.

Getting the hybridization rate constant

By collecting all the FPT points, we construct a probability distribution of First Passage Times τ for the given simulation data. A Beta probability density function is fitted on FPT distribution data.

stochastic reaction rate c_a of the same reaction will be related as $c_a = \frac{k_a}{VN_A}$ with N_A the avogadro number, resulting in waiting times $\tau = \frac{1}{c_a} \propto N_A$

⁷Here we assume that after first collision the two involved DNA single strands *A* and *B* will not separate.

We find FPTs to be Beta distributed $B(t; \alpha, \beta)$ with sequence dependant shape parameters α and β :

$$Beta(t; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1} (1-t)^{\beta-1} \quad (6.2)$$

The Mean First Passage Time, MFPT is calculated as follows:

Given the FPT distribution of simulation batch σ $f_\sigma(t)$, we generate the fitted Beta distribution on $f_\sigma(t)$, denoted $Beta(t; \alpha_\sigma, \beta_\sigma)$. We then compute the Mean First Passage Time of simulation batch σ as:

$$\langle \tau_\sigma \rangle = \int_0^1 t Beta(t; \alpha_\sigma, \beta_\sigma) dt = \frac{\alpha_\sigma}{\alpha_\sigma + \beta_\sigma} \quad (6.3)$$

Getting the hybridization kinetic rate $k_{hyb,\sigma}$ corresponding to simulation batch σ as:

$$k_{hyb,\sigma} = \frac{1}{\langle \tau_\sigma \rangle} = \frac{\alpha_\sigma + \beta_\sigma}{\alpha_\sigma} \quad (6.4)$$

Moreover we can have insights on the variance of First Passage Times of simulation batch σ either by calculating it directly from its FPT distribution or from the fitted Beta distribution $Beta(t; \alpha_\sigma, \beta_\sigma)$

$$Var(\tau_\sigma) = \frac{\alpha_\sigma \beta_\sigma}{(\alpha_\sigma + \beta_\sigma)^2 (\alpha_\sigma + \beta_\sigma + 1)} \quad (6.5)$$

By considering that in order to model the FPT distribution, the Beta distribution will have shape parameters $\beta \gg \alpha$ and $\alpha \approx 1$ we can see that the behavior of DNA hybridization deviates in fact very marginally from being exponentially distributed. By denoting $Beta_\sigma$ and Exp_σ the distributions with parameters $(\alpha_\sigma, \beta_\sigma)$ and λ_σ fitted on simulation batch σ for Beta and Exp distributions respectively, we see that for $\beta \gg \alpha$ and $\alpha \approx 1$:

$$Mean(Beta_\sigma) = \frac{\alpha_\sigma}{\alpha_\sigma + \beta_\sigma} \cong \frac{\alpha_\sigma}{\beta_\sigma} \cong \frac{1}{\beta_\sigma} \approx \frac{1}{\lambda_\sigma} = Mean(Exp_\sigma) \quad (6.6)$$

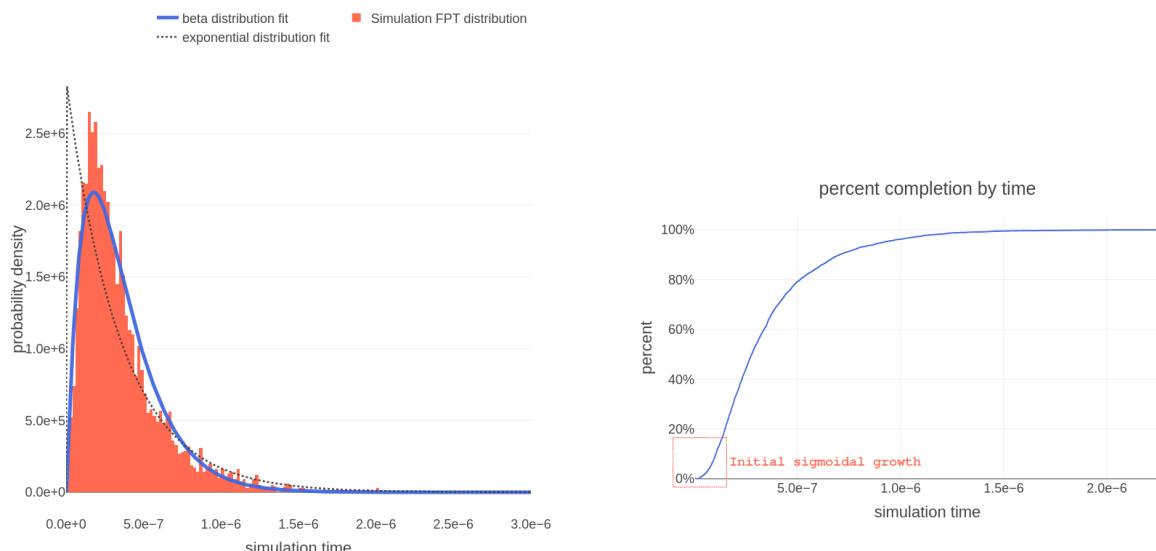
$$Var(Beta_\sigma) = \frac{\alpha_\sigma \beta_\sigma}{(\alpha_\sigma + \beta_\sigma)^2 (\alpha_\sigma + \beta_\sigma + 1)} \cong \frac{\alpha_\sigma}{\beta_\sigma^2} \cong \frac{1}{\beta_\sigma^2} \approx \frac{1}{\lambda_\sigma^2} = Var(Exp_\sigma) \quad (6.7)$$

So in general DNA Hybridization can be reduced to a process with exponentially distributed transition times. However the beta distribution of FPTs shouldn't be ignored as it may hint at a deeper understanding of the hybridization process, which may be considered for future theoretical developments stemming from this thesis work.

By plotting the percentage of completed trajectories on the y-axis as a function of simulation time on the x-axis in figure 6.6b, we see that the Beta distributed nature of the FPT distribution of DNA hybridization trajectories corresponds to initial sigmoidal growth, which may hint at *Hill-equation* governed kinetics.

We hypothesize that this may be connected to the nucleation step as a sequence-dependant cooperative binding process between sequence monomers.

We observe however that while the percent completion graph may look very similar to an experimental fluorescence curve produced by hybridization experiments, it should **not** be automatically assumed that these two are comparable, due to the fundamental differences between the stochastic simulation of two molecules in their configuration space against experimental realizations of bimolecular reactions in which the total number of molecules changes with time. Deeper analyses are then needed to investigate this finding that lay outside of the scope of the present thesis work.



(a) Distribution of First Passage Times of simulation trajectories with best beta probability density function fit. Exponential distribution fit is found non-descriptive.

(b) Percent of completed trajectories by simulation time. We see that by time $2\mu s$ all trajectories reach the duplex state. Initial sigmoidal growth hints at possible Hill-like nucleation kinetics

Figure 6.6: The initial sigmoidal growth as a consequence of the beta-distributed first passage time.

6.4 Fit against experimental data

We optimized the zipping graph model against experimental data provided in Hertel et al. [32] on 41 experimentally measured⁸ hybridization rate constants relative to sequences with no stable intra-strand secondary structure (i.e. hairpins and intra-strand pseudoknots). The lack of intra-strand secondary structure in the considered sequences renders their data-set especially tailored to be used for the zipping graph model, as the description in it developed is limited, so far, to non-secondary-structured sequences.

⁸produced through Surface Plasmon Resonance experiments

The four parameters we optimize are: κ_s **basal sliding rate**, k_z **zipping rate** and **steric angles** of nucleation θ and ϕ . These meaning of these parameters is described at the beginning of the current chapter.

For each sequence provided, an ensemble of 5000 simulations has been performed in order to reduce error in the measure of the mean first passage time and the rate. We particularize our description to match the experimental conditions underlying our data, namely:

$$T = 25^\circ C \quad (6.8)$$

$$[NaCl] = 150 mM \quad (6.9)$$

Predicted hybridization rates k_{mod} are plotted on the y-axis against experimentally determined hybridization rates k_{exp} on the x-axis in figure 6.7. The plot is constructed around the identity line, in such a way that if model predicted values are close to the experimental ones, the resulting point (k_{exp}, k_{mod}) will be close to the identity line.

For each measured rate i and the variance related to its first passage time distribution σ_i^2 we measure the standard error $se = \sigma_i / \sqrt{N_s}$ where $N_s = 5000$ refers to the number of performed simulations. The standard error informs us on the accuracy of the measured MFPT relative to the number of simulations performed. We find that $N_s = 5000$ performed simulations ensure the precision of the computed rate constants relative to their variance to be accurate.

By calculating the *Pearson* correlation coefficient between model and experimental data $\rho_{mod,exp}$ we find strong correlation between the zipping graph model predicted hybridization rates and experimental data $\rho_{mod,exp} = 0.6832$:

$$\rho_{mod,exp} = \frac{\text{cov}(k_{mod}, k_{exp})}{\sigma_{mod}\sigma_{exp}} = 0.6832 \quad (6.10)$$

Considering D the data-set and for elements $d \in D$ consider $d(seq)$ the associated sequence, denoting $k_{exp}(d(seq))$ and $k_{mod}(d(seq))$ the experimental and the computed rates from given sequence, we calculate the Root Mean Squared Error (RMSE) and find a value for the optimized model of $RMSE = 0.1398$:

$$RMSE = \sqrt{\frac{1}{|D|} \sum_{d \in D} \left(\log_{10} \left(\frac{k_{mod}(d(seq))}{k_{exp}(d(seq))} \right) \right)^2} \quad (6.11)$$

$$RMSE_{ZG} = 0.1398 \quad (6.12)$$

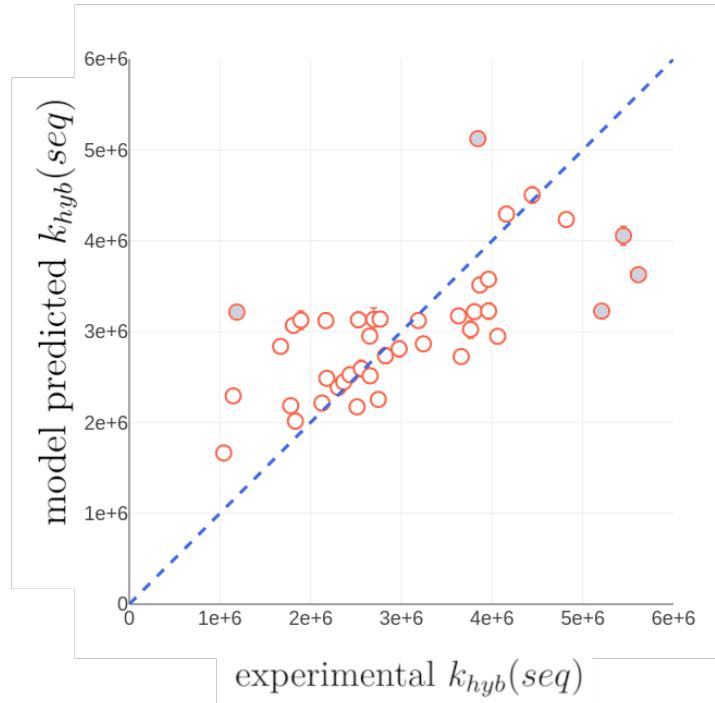


Figure 6.7: Model predicted versus experimental hybridization rates. In **blue** the identity line. This scatter plot corresponds to following error measures:

$$\rho_{mod,exp} = 0.6832 \text{ Pearson correlation}$$

$$RMSE = 0.1389 \text{ Root Mean Squared Error}$$

$$F2_{acc} = 0.9512 \text{ Distance 2 accuracy}$$

Error measures relative to experimental data are not available. Error measures relative to simulation results are too low to be visible due to optimization of number N_m of simulations to perform. **Solid dots** refer to **outliers** which are points that the zipping graph model doesn't predict well.

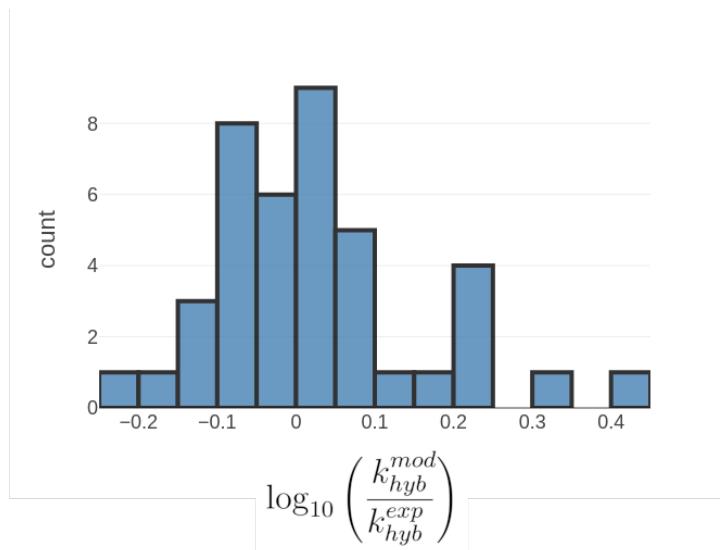


Figure 6.8: Error histogram produced by computing the difference $\log_{10}(k_{hyb}^{mod}) - \log_{10}(k_{hyb}^{exp})$. Negative values correspond to under-evaluation and positive values to over-evaluation. We see that on average the zipping graph model with the found parameters is slightly under-evaluating.

Best correlation $\rho_{mod,exp} = 0.6832$ has been obtained with the following parameters:

$$\begin{aligned}
 k_z &= 8.34 \cdot 10^7 && \text{zipping rate} \\
 k_s &= 5.00 \cdot 10^5 && \text{sliding rate} \\
 \theta &= \frac{\pi}{2} && \text{azimuthal angle} \\
 \phi &= \frac{2\pi}{3} && \text{inclination angle}
 \end{aligned}$$

6.5 Comparison with known models

We confront the performance of the zipping graph model, described in the last section with the performance of three most recent similar models.

- **Hertel et al. 2022 [32]:** Authors develop a parsimonious model based on combinatorics, thermodynamics and two parameters fitted against data. The physical significance of their parameters is quite general and is not well defined.
- **Zhang et al. 2018 [61]:** Authors develop a Machine Learning model based on a modification of the k-nearest neighbors algorithm. They provide the model with six-dimensional data related to sequences. The model cannot provide mechanistic understanding of hybridization pathways.
- **Hata et al. 2017 [62]:** Authors develop a model that explicitly incorporates secondary structure and sequence-dependant nucleation. However their model is based on 32 free parameters fitted against 47 experimental data points. Their model is prone to over-fitting.

	ZG Model	Hertel	Zhang	Hata
<i>data-set size</i>	41	41	210	47
<i>free parameters</i>	4	2	ND	32
$\rho_{mod,exp}$	0.6832	0.69	0.60	0.92
RMSE	0.1398	ND	0.280	ND
$F2_{acc}$	0.9512	ND	0.795	ND
<i>secondary struct.</i>	N/A	N/A	yes	yes

As previously seen, $\rho_{mod,exp}$ is the Pearson correlation coefficient and RMSE is the Root Mean Square Error between the experimental data-set and model predictions on such set. An auxiliary error measure called $F2_{acc}$ has been used in Zhang [61], which refers to the number of sequences predicted to be inside a range of two from the true value.

Below we confront scatter plots between the zipping graph model and the other three considered model.

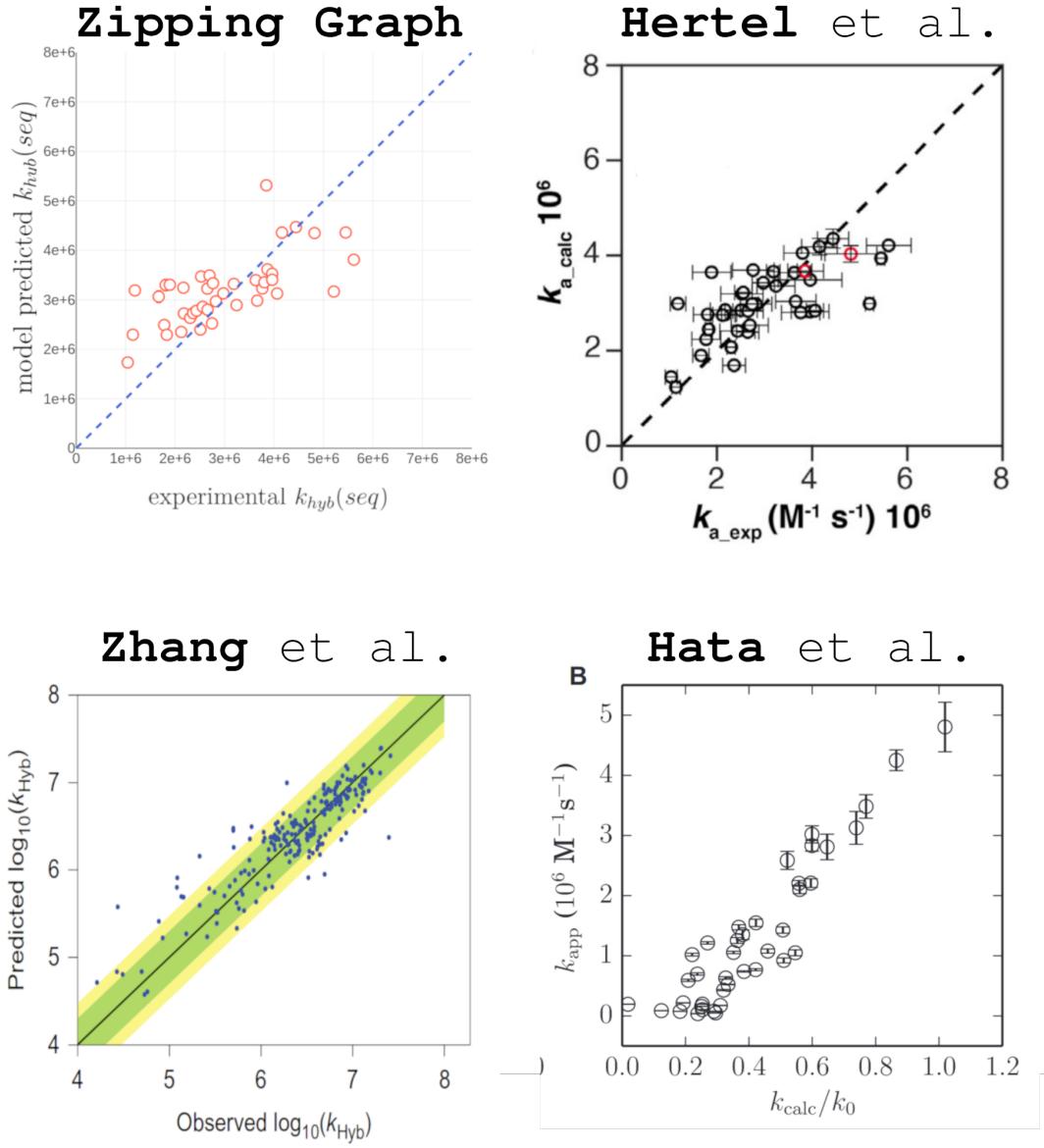


Figure 6.9: Comparison of the zipping graph scatter plot against the models [32, 61, 62].

We conclude the description and analysis of the zipping graph model, by witnessing that even with its current modeling limitations, it manages to generate results that are on-par with modern state-of-the-art attempts at the same problem.

As a deal-breaker, our model is physically-grounded in such a way that allows us to consider the membrane-anchored DNA hybridization case in a more meaningful way, by imparting appropriate mathematical modifications to rate calculation methods and graph structure generation of the zipping graph model, in order to produce the first recorder theoretical predictions on hybridization kinetics of membrane-anchored DNA-lipid conjugates.

6.6 Zipping graph model of membrane-anchored DNA

We are interested in the characterization of DNA hybridization kinetics for membrane-anchored DNA-lipid conjugates. Such theoretical predictions are of scientific interest for a wide array of applications of membrane-anchored DNA-lipid conjugates, such as the Membrane-LIDA design first outlined in section 3.4, DNA-lipid mediated binding of Vesicles and generally applications of DNA mediated membrane functionalization in synthetic biology. The model is instantiated in the 2D case by setting the `class Model` to the two-dimensional case. This will communicate to all the other classes, which behavior inherently depends on `Model`, to construct the model in a two-dimensional compliant way.

The **modifications** conferred to the zipping graph model in the two-dimensional case are:

- Negligible off-register transitions.
- Only azimuthal steric constrain.
- Top-Bottom separated nucleation.

Meanwhile the following are assumed to be **independent** on dimensionality:

- Zipping Rate.
- Thermodynamics.

We adopt now and again the acronym for Membrane-Anchored DNA, membrane anchored DNA.

We start our discussion on such assumptions by considering the two last elements, assumed to be independent on dimensionality.

Zipping Rate

We argue for the zipping rate being still a reaction happening between monomers moving in three dimensional space. Their motion is not affected by center of mass diffusion of the membrane anchored DNA, which is two orders of magnitude lower than the center of mass diffusion of DNA strands in three dimensions and three orders of magnitude lower than single monomer diffusion (reference to eq. (4.34) in section 4.4.1). For such reasons we argue that the zipping rate will not be affected by membrane anchoring as it is associated with the motion of single monomers.

Thermodynamics

The assumption on dimensionality-independence of **thermodynamics** is in fact an approximation on disguise. We argue that the thermodynamical description of DNA-lipid anchoring may be modified by considering the entropy-reduction effect resultant from membrane anchoring.

As from the nearest neighbor model of DNA thermodynamics (chapter 2), we have an initiation free energy penalty which at $T = 26^\circ$ is approximately $\Delta G^{init} \sim 1.9 \frac{Kcal}{mol}$. This free energy penalty is completely associated with the entropy reduction of the whole system related to the ordering consequential to hybridization of two initially separated DNA single stands.

We argue that the entropic effect brought by membrane-anchoring of DNA-lipid conjugates, will affect the two-dimensional initiation energy penalty ΔG_{2D}^{init} by keeping it in a range between zero and the three dimensional case, as the two-dimensional case is in fact a globally more ordered state than the three dimensional one.

Since the effect of ΔG_{2D}^{init} is destabilizing it results in our model in the slow-down of the nucleation process. We now know that it is bound to be in the range $0 < \Delta G_{2D}^{init} < \Delta G^{init}$, where ΔG^{init} refers to the three dimensional case. Then by introducing this thermodynamic approximation, we are in fact **underestimating** the resulting transition rates. We in fact expect, as argued, that the two-dimensional nucleations will be on-average more stable than the three dimensional ones, and more stable nucleations will surely lead to faster kinetics. Quantitative considerations on such arguments are left for future and deeper studies on membrane anchored DNA hybridization.

We now consider the easier to treat assumptions on the modifications that we expect two-dimensionality brings to the system.

Negligibility of off-register transitions

As in figure 6.10, we expect bottom nucleation to dominate the nucleation process. Imagine as an example an off-register nucleation happening between the top strands, which being far from the surface are conformationally free and allowed to initiate off-register nucleations. If this top-nucleation zips and has some stability, it will keep strands glued and close to each other in a very finite region, resulting in a very high chance of bottom on-register nucleation. This nucleation will most likely zip until the duplex, displacing the off-register top nucleation. If top strands nucleate on-register then the strand will just zip to the duplex. Last case is if a collision happens directly at the bottom, such case is also straightforward.

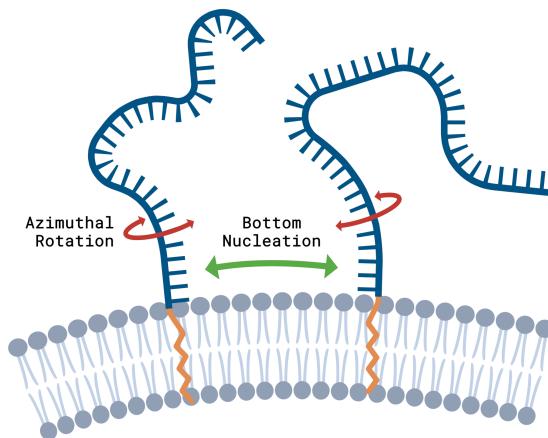


Figure 6.10: Cartoon representation of two-dimensional collision and nucleation. We expect bottom base pairs to nucleate only in-register, resulting in higher on-register probability of nucleation and consequent on-register zipping, increasing rates. We also expect for such nucleations that the only relevant steric angle is the azimuthal one.

Steric angle θ at the bottom strand

For bottom nucleations then we imagine the steric factor stemming from orientational considerations between colliding monomers to be only driven by azimuthal relative orientation, due to the orthogonal directional constraint imparted on the inclination, as explained for (4.35) in chapter 5.

Top-bottom nucleation separation

We then separate nucleation events in two categories: *top* nucleations and *bottom* nucleations. Dividing the two strands in top and bottom part is done by considering the polymeric properties of single stranded DNA outlined in 4.2.1, and fixing the bottom part at the Kuhn length $b = 2l_p$ where l_p is the persistence length of single stranded DNA.

As done in the simpler three-dimensional case (sec. 6.3), we compute nucleation by sampling from two exponential distributions, each associated with collision rates relative to top and bottom nucleations.

Top nucleation time will be sampled from an exponential distribution $Exp(t; \lambda_{top})$ with λ_{top} corresponding to a Sano-Tachiya (4.30) (sec. 4.4.1) diffusion controlled reaction calculated over the gyration radius (4.15) of the top-strand⁹, with center of mass diffusivity given by lipid diffusion (4.34) and steric hindrance given by both steric angles (θ, ϕ) as in equation (4.17).

⁹As discussed in section 4.4.1, the gyration radius of a membrane-anchored single strand of DNA is well approximated by the corresponding free-in-solution value for strands with less than 40 nucleotides [53]

Bottom nucleation time will be sampled from an exponential distribution $\text{Exp}(t; \lambda_{bot})$ with λ_{bot} corresponding to a Sano-Tachiya (4.30) (sec. 4.4.1) diffusion controlled reaction calculated over the physical radius of single stranded DNA monomers, with center of mass diffusivity given by lipid diffusion (4.34) and steric hindrance given by only the azimuthal angle θ as in equation (4.35)

The two random variables T_{bot} T_{top} produced by sampling from $\text{Exp}(t; \lambda_{bot})$ and $\text{Exp}(t; \lambda_{top})$ are compared:

$$T_{collision} = \begin{cases} T_{bot} & \text{if } T_{bot} \leq T_{top} \\ T_{top} & \text{if } T_{bot} > T_{top} \end{cases} \quad (6.13)$$

Depending on which collision happens we distinguish three cases:

$$\text{collision} \rightsquigarrow \begin{cases} \text{bottom} & \text{if } T_{bot} < T_{top} \\ \text{top} & \text{if } T_{bot} > T_{top} \\ \text{any} & \text{if } T_{bot} \approx T_{top} \end{cases} \quad (6.14)$$

We are now ready to run simulations and produce zipping graph based predictions on DNA hybridization on membranes.

6.7 Zipping graph based predictions on membrane-anchored DNA hybridization kinetics

We run the zipping graph model on the same data it has been validated with in section 6.4 and on the same model conditions. Crucially, as discussed, the zipping rate parameter k_z is kept at the same exact value as in the optimized model of section 6.4. We simulate membrane-anchored DNA hybridization for a membrane with lipids of radius $r_{lip} = 2 \text{ nm}$ and membrane viscosity $\eta_{mem} = 0.5 \text{ nPa s m}$. Polymeric properties of DNA are left unchanged.

We find membrane-anchored DNA hybridization kinetics to be ~ 10 times faster than the same in bulk. With k_{2D} and k_{3D} the zipping graph computed results on given sequences for the two and three dimensional case, we find on average:

$$\langle \frac{k_{2D}}{k_{3D}} \rangle_{\text{sequences}} = 10.83$$

We find a striking length-dependence of two-dimensional hybridization rates.

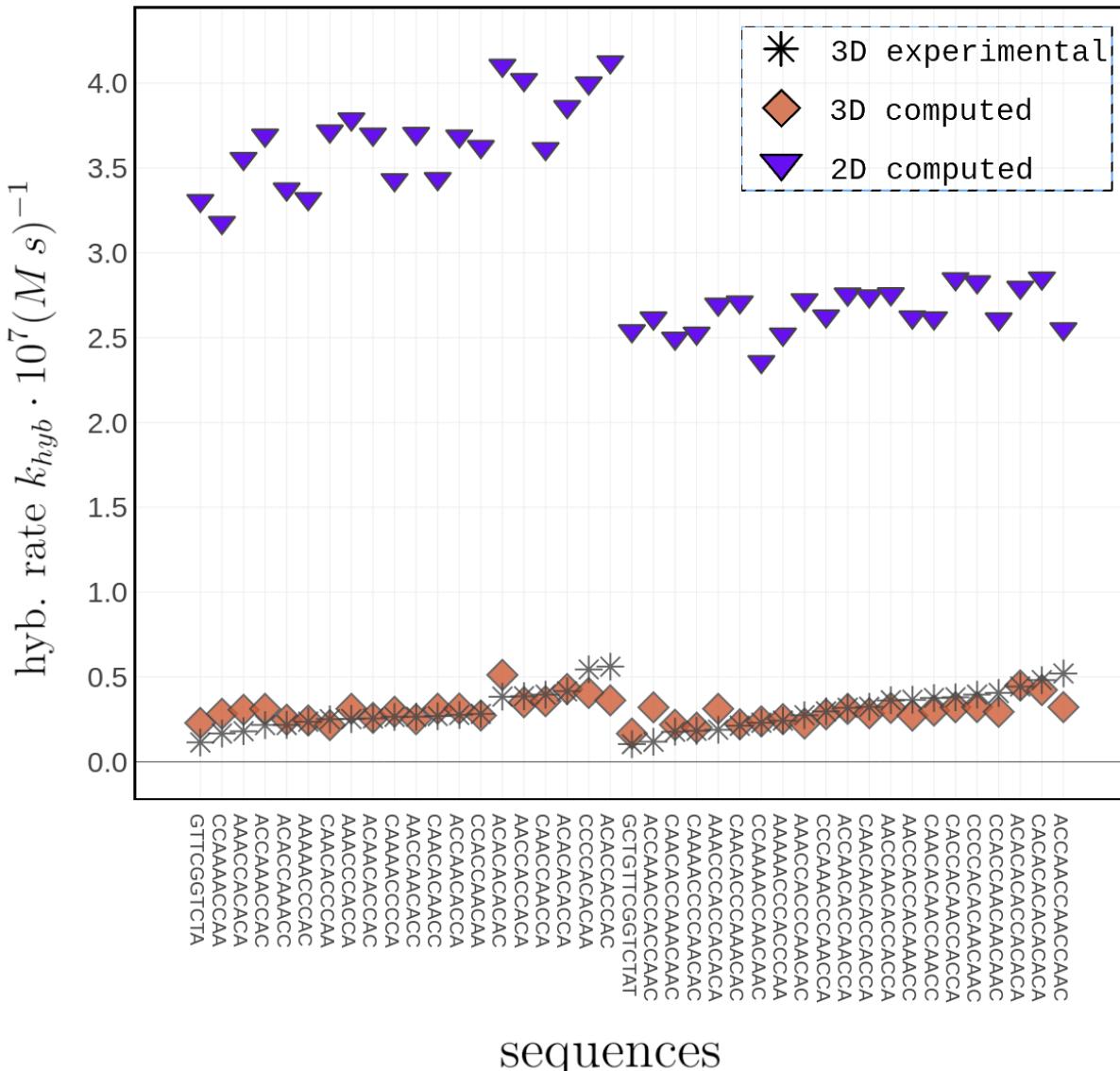


Figure 6.11: Two-dimensional zipping graph predicted kinetics are plotted together with three dimensional results and experimental data of three-dimensional kinetics. Membrane-anchoring results in an order 10 increase in the kinetics of hybridization and negative sequence dependance of hybridization rates, i.e. longer sequences hybridize slower than shorter ones.

Short sequences hybridize faster than longer sequences, contrary to the three dimensional case in which hybridization kinetics are independent from sequence.

With L_{14} we denote the set of short sequences (length 10 nucleotides) and with L_{10} we denote the set of long sequences (length 14 nucleotides). We find on average that:

$$\langle k_{2D} \rangle_{L_{14}} = 2.657 \cdot 10^7 \text{ (M s)}^{-1}$$

$$\langle k_{2D} \rangle_{L_{10}} = 3.663 \cdot 10^7 \text{ (M s)}^{-1}$$

$$\langle \frac{k_{2D}}{k_{3D}} \rangle_{L_{14}} = 9.62$$

$$\langle \frac{k_{2D}}{k_{3D}} \rangle_{L_{10}} = 12.09$$

From our preliminary results we can thus expect from future experiments aimed at measuring DNA Hybridization kinetics on membranes, to observe an acceleration in the kinetics by one order of magnitude in conditions similar to the presently described.

6.8 Zipping graph on the LIDA system

We can now apply the developed methods to generate an accurate prediction on Membrane-LIDA replication kinetics. We report here as a reminder the involved strands:

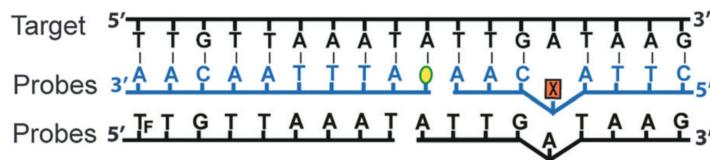


Figure 6.12: Caption

We have to neglect the presence of mismatches on hybridization kinetics, as it is not implemented yet in the zipping graph model. We suggest that their presence may hamper the zipping process and slow down hybridization kinetics. Since we are not considering this factor both for the 3D and the 2D case, we can then go on with making predictions on relative change, by keeping in mind that the absolute scale may vary.

By calculating hybridization kinetics for the involved strands in the 3D and in the 2D case, we observe a 27% reduction of in LIDA replication time τ from $\tau_{3D} = 74.55 \text{ min}$ to $\tau_{2D} = 54.41 \text{ min}$. Below the corresponding deterministic concentration time-evolution plots of the two systems as carried out in chapter 3:

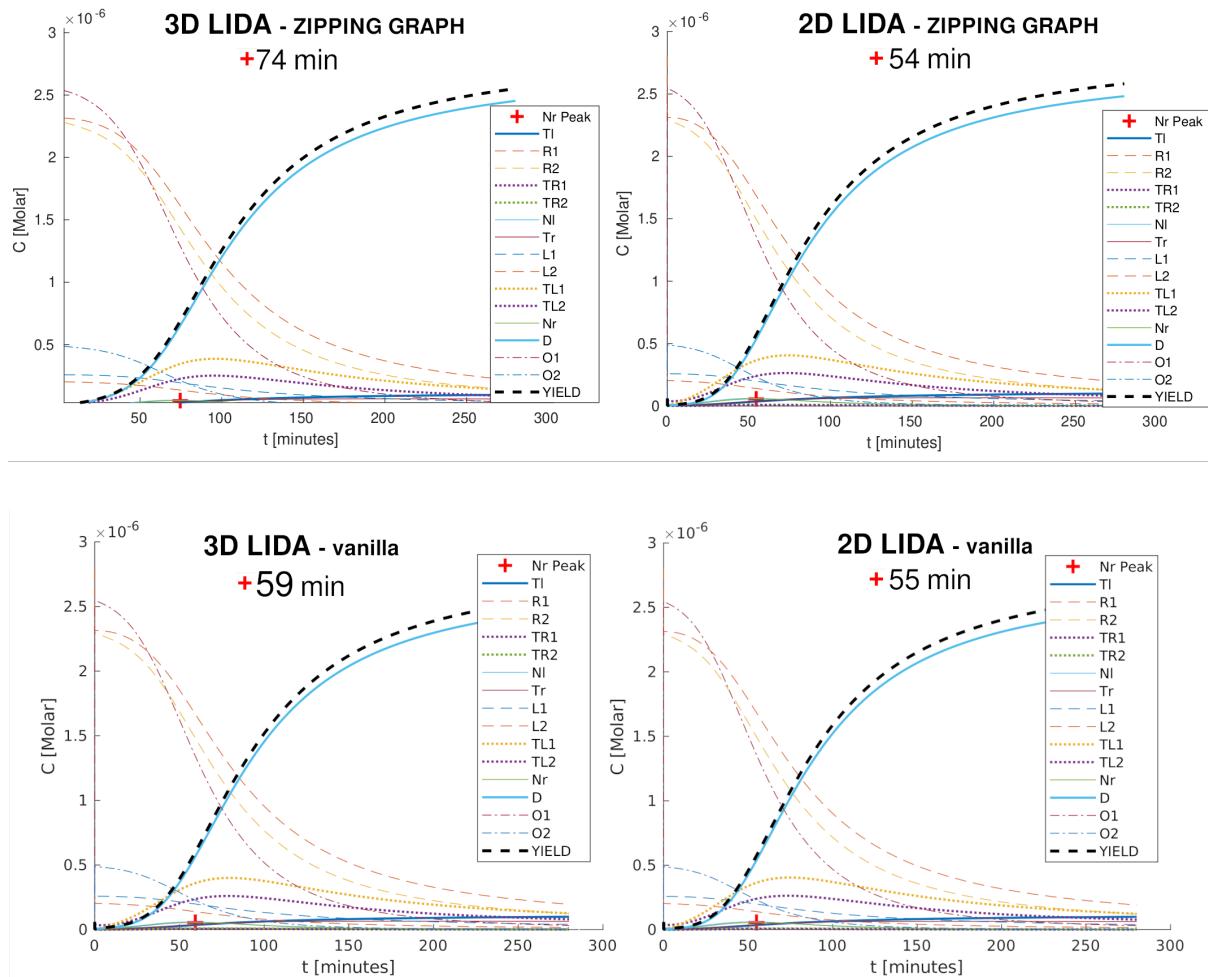


Figure 6.13: 3D LIDA kinetics derived from the zipping graph model has 74.55 minutes replication half-life. 2D LIDA kinetics on the other hand has 54.41 minutes replication half-life.

7 Discussion and prospects

The research endeavor carried out in this thesis work, and the investigative choices taken by the author and the supervisor, can be fully understood if framed under the fundamental question we want to give an answer to:

*How does membrane-anchoring of DNA affect the
Lesion Induced DNA Amplification (LIDA) system?*

The many research paths that may have been taken for this thesis work, have been constrained and directed by our aim to answer such question.

7.1 Considerations on LIDA and membrane-LIDA

In chapter 3 we examined the exponential dependency of LIDA replication half-life on duplex stability, quantitatively confirming the design intuition behind the system. We probed and observed a power-law dependency on the ligation rate in the range of parameters herein considered.

We instantiated the ODE-based model of LIDA in the Membrane-LIDA design which entails surface-surface and surface-bulk DNA hybridization in section 3.4 therein. We simplify the modeling approach the Membrane-LIDA system by considering the volume to surface ratio scaling of the two-dimensional kinetic rate into an equivalent three dimensional rate.

We then explore replication half-life surfaces of M-LIDA by probing surface-bulk hybridization against surface-surface hybridization rates. Everything else kept equal, we find a striking exponential reduction of replication half-life based on the increase in surface-surface hybridization rates, which, due to the lack of experimental data on surface-surface DNA hybridization rates, propels our research into the realm of theoretical modeling of DNA hybridization kinetics, in order to explore theoretical predictions on surface-surface hybridization.

7.2 Considerations on DNA hybridization modeling

Through extensive joint literature research of the author and the supervisor, discussions with theoretical and experimental experts on DNA hybridization kinetics such as Prof. Thomas Schäfer of the Polymat institute in San Sebastian (Spain), directly involved in the

experimental characterization of DNA hybridization kinetics, and Dr. Harold Fellerman of Newcastle University, involved in molecular modeling efforts of DNA computing systems for which DNA hybridization kinetics are of fundamental importance; we arrive to the conclusion of the lack in our current knowledge of DNA hybridization modeling efforts that can be easily adapted to the two-dimensional surface-surface DNA hybridization case.

In chapters 4 we apply the knowledge on fundamental processes involved in DNA hybridization kinetics and reaction rate theory in solution and in surfaces to produce our first analytical predictions on the effects of membrane-anchored DNA hybridization kinetics on the LIDA system, finding only a slight improvement.

The simple analytical modeling approach therein proposed and largely based on well-known facts on DNA hybridization kinetics reported in the literature, doesn't possess enough descriptive power to assess the sequence-dependant kinetics of hybridization in short DNA strands. As a fact, for such modeling efforts, analytical models are more adapted to the prediction of DNA hybridization for strands of minimum length 100 nucleotides, where analytical coarse-graining produces effective results. For a detailed review on such matters consider [41].

We then advance our research in 5 by applying knowledge from Molecular Dynamics simulations to produce new analytical coarse-grained models for enhancing our description of the kinetic pathways involved in DNA hybridization, already started in chap. 4.

We aim to instantiate the analytical models so far developed for specific hybridization pathways to create a broader and coherent computational model for DNA hybridization that encompasses such physics-based knowledge, which we can adapt in the two-dimensional case in a physically meaningful way through the adoption of reaction-rate theory for surface reactions and considerations on the geometrical modifications imparted by surface-anchoring.

7.3 Considerations on the Zipping Graph Model

We successfully built an "umbrella" mesoscale computational model of DNA hybridization that combines base-pairing combinatorics, off-register metastable states and zipping-based dynamics in a coherent kinetic graph structure representing dominant hybridization pathways. Detailed description of the zipping graph model is to be found in chapter 6.

The kinetic graphs so constructed implement reaction pathways based on reaction-rate theoretical reasoning and polymer physics based spatial properties of DNA described in chapters 4 and 5.

So far, the zipping graph model achieves state-of-the-art performance on relevant data when compared with most recently achieved levels of accuracy in the prediction of DNA hybridization kinetics [32].

Through the zipping graph model we can glimpse into DNA hybridization rates in two-dimensional motion, representing hybridization kinetics of membrane-anchored DNA single strands. We produce predictions of a 10 fold boost for the kinetics of membrane-mediated hybridization.

The so far developed zipping graph model is however limited to non-secondary-structured DNA sequences, namely sequences that lack intra-strand stable structures such as hairpins or pseudoknots.

This greatly limits the descriptive power of the currently proposed computational model. Nevertheless, its built-in modularity enables the future upgrade of the zipping graph modeling capacities to include intra-strand secondary structures, without affecting the already built description but rather by amplifying its descriptive capabilities if correct implementation is achieved.

The biggest drawback of the zipping graph implementation so far, stands in our description of sliding transitions.

These are in reality remarkably intricate transitions that involve simultaneous zipping and unzipping of base pairs at different registers with fine-tuned rates dictated by the structure-sequence dependant underlying destabilizing loops, which deliver a loop-sequence dependant destabilizing effect relative to affected local portions of the involved complex. By then approximating the sliding rate as in 5.1.3 we are removing the layer of complexity associated with sliding transition themselves.

A subsequent layer of complexity is added by considering the ongoing competition between different registers governed by the relative free-energy landscape, which may result in long-lived dynamic meta-stability between geometrically compatible and free-energy similar registers, as an example of such long-lived metastable states, we point the reader to figure 7.1. We neglected such sequence-dependant free-energy-competition landscapes of off-registers by assuming that this can be neglected due to the large free-energy domination of the duplex state at minimum free energy. By doing this we discarded the competition layer of complexity.

Moreover, the zipping graph model is so far limited in its algorithmic implementation to consider only input strands of same nucleotides length. If we need to consider hybridizing sequences of different length, where the shorter strand is complementary to a portion of the longer one, we incur in the annoying activity of manually "cutting" the input strands to consider the correct hybridizing portion and manually modify the **class Kinetics** to

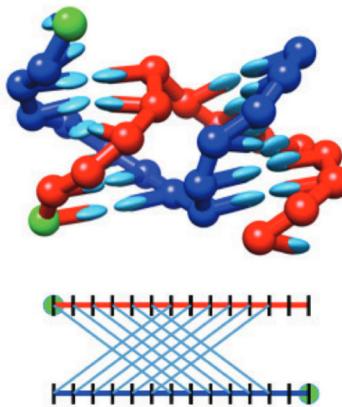


Figure 7.1: Long-lived pseudoknotted state at geometrically compatible registers 6 and -6. Both registers present the stacking of 6 base pairs, resulting in a long-lived competition between the two registers, as none of the two strongly dominates the other in terms free-energy measured stability. Figure from [58]

match polymeric geometric properties of the longer strand. Such functions are of easy implementation and will be applied soon.

The object oriented algorithmic implementation of the zipping graph model is very efficient. We report on the speed of the whole program, from input sequences and system conditions to hybridization rate output, results in computation time $T_c(L)$ for input sequences long L ranging from $T_c(10) \sim 1.5\text{ s}$ to $T_c(80) \sim 240\text{ s}$ due to the exponential growth of number of nodes in the kinetic graph relative to strand length. The reported computation time ranges are relative to whole-computation tasks, from completion of NUPACK-aided thermodynamics calculation for each involved complex, graph generation routines, execution and analysis of 5000 stochastic simulation and data-saving routines.

The zipping graph model is nonetheless capable of achieving state-of-the-art performance by correct algorithmic implementation of the most fundamental processes involved in DNA hybridization and a good balance between detailed fine-tuning and introduced approximation. Specifically on and off register nucleation, zipping and sliding (inchworming and pseudoknotting). Relative to the state-of-the-art, the zipping graph model introduces more detailed physical modeling of involved elementary transitions, allowing us to explore the effects of membrane-anchoring on DNA hybridization kinetics by considering appropriate changes in the physical modeling of the involved elementary processes.

We emphasize that no experimental data has ever been recorder on DNA Hybridization kinetics of membrane-anchored DNA-lipid conjugates. This is mainly due to the high experimental difficulty of measuring such processes happening on the outer-layer of a lipid vesicle surface. Such topics have been extensively discussed by the author with Prof.

Thomas Schäfer of the Polymat institute in San Sebastian, Spain. The elaboration of an experimental setup apt at probing membrane-mediated DNA hybridization is under discussion and construction. In-vitro experiments will be carried out in collaboration with Prof. Schäfer and Prof. Sabine Müller at Universität Greifswald in Germany.

The zipping graph model has thus been capable of producing the first ever recorded theoretical *and* quantitative predictions on membrane-mediated DNA hybridization kinetics, which we believe are ready to be tested in the laboratory.

8 Conclusion

In this thesis, we have investigated the effects of membrane-anchoring on DNA hybridization kinetics and its impact on the Lesion Induced DNA Amplification (LIDA) system, which plays a crucial role in protocellular systems. Through the development of the Zipping Graph Model, we have extended our understanding of hybridization kinetics for membrane-anchored DNA.

Our model has allowed us to make the first ever recorded theoretical predictions on membrane-mediated DNA hybridization kinetics, which we believe to be ready for experimental validation. The findings of this study suggest that under specific conditions, membrane-anchoring can lead to a 10-fold acceleration in hybridization rates and a 27% increase in self-replication efficiency of the LIDA scheme when compared to the three-dimensional case. Additionally, we have identified a strong length-dependence on membrane-anchored hybridization rates which decrease with increasing length, in stark contrast to the length-independence observed in the three-dimensional case.

While our current model has limitations, such as the exclusion of intra-strand secondary structures and the coarse-graining of off-register transitions, its modularity allows for future upgrades to enhance its descriptive capabilities, ensuring the potential for continued optimization and adaptation of the model to address new challenges in the field of DNA hybridization kinetics.

The insights gained from this research not only contribute to our fundamental understanding of DNA hybridization kinetics in membrane-anchored systems but also provide a foundation for the design and optimization of non-enzymatic self-replicating systems in protocell models. As experimental techniques advance, it will be essential to corroborate our theoretical predictions with empirical data, ultimately refining our understanding of the underlying processes and guiding the development of more efficient and robust protocellular systems.

Concluding, we remark our findings to be pointing towards the broader origins of life perspective, for which dimensionality reduction may have been a key mechanism through which early living systems increased their metabolic efficiency.

9 Acknowledgement

I'm extremely grateful to my supervisor, Prof. Steen Rasmussen, for his guidance and advice during this master thesis project. I'd like to thank also Prof. Michele Caselle for his precious help and helpful critique. Special thanks go to Prof. Thomas Schäfer for the fruitful discussions on DNA Hybridization as well as thanks go to Dr. Harold Fellerman for his advice on the same matters. I would like to express my deepest gratitude to Prof. Ricardo Díez Muiño for hosting me as a visiting student at the Donostia International Physics Centre of San Sebastian, Spain. I'd like to thank all my professors at University of Turin, for teaching me not only physics but also what's the spirit of a physicist. Thanks to all my colleagues for their help. Lastly, I'm extremely grateful to the European Center for Living Technologies, for inviting me at their yearly meeting in September in Venice, giving me the opportunity to meet in person many great scientists and people. I'm also thankful to Prof. Ralf Metzler for inviting me at the Fluctuations in Small Complex Systems VI conference in Venice, which has helped me broaden my perspectives.

Bibliography

- [1] Steen Rasmussen, Adi Constantinescu, and Carsten Svaneborg. Generating minimal living systems from non-living materials and increasing their evolutionary abilities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1701):20150440, 2016.
- [2] Michael C Wamberg, Rafał Wieczorek, Søren Bo Brier, Jan Willem de Vries, Minseok Kwak, Andreas Herrmann, and Pierre-Alain Monnard. Functionalization of fatty acid vesicles through newly synthesized bolaamphiphile–dna conjugates. *Bioconjugate Chemistry*, 25(9):1678–1688, 2014.
- [3] Michael S DeClue, Pierre-Alain Monnard, James A Bailey, Sarah E Maurer, Gavin E Collis, Hans-Joachim Ziock, Steen Rasmussen, and James M Boncella. Nucleobase mediated, photocatalytic vesicle formation from an ester precursor. *Journal of the American Chemical Society*, 131(3):931–933, 2009.
- [4] Jonathan L Cape, Joseph B Edson, Liam P Spencer, Michael S DeClue, Hans-Joachim Ziock, Sarah Maurer, Steen Rasmussen, Pierre-Alain Monnard, and James M Boncella. Phototriggered dna phosphoramidate ligation in a tandem 5'-amine deprotection/3'-imidazole activated phosphate coupling reaction. *Bioconjugate Chemistry*, 23(10):2014–2019, 2012.
- [5] Pierre Mignon, Stefan Loverix, Jan Steyaert, and Paul Geerlings. Influence of the $\pi-\pi$ interaction on the hydrogen bonding capacity of stacked dna/rna bases. *Nucleic Acids Research*, 33(6):1779–1789, 2005.
- [6] John SantaLucia Jr and Donald Hicks. The thermodynamics of dna structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, 33:415–440, 2004.
- [7] Joseph N Zadeh, Conrad D Steenberg, Justin S Bois, Brian R Wolfe, Marshall B Pierce, Asif R Khan, Robert M Dirks, and Niles A Pierce. Nupack: Analysis and design of nucleic acid systems. *Journal of computational chemistry*, 32(1):170–173, 2011.
- [8] Robert M Dirks, Justin S Bois, Joseph M Schaeffer, Erik Winfree, and Niles A Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM review*, 49(1):65–88, 2007.

- [9] Hatim T Allawi and John SantaLucia. Thermodynamics and nmr of internal g · t mismatches in dna. *Biochemistry*, 36(34):10581–10594, 1997.
- [10] Hatim T Allawi and John SantaLucia. Nearest-neighbor thermodynamics of internal a · c mismatches in dna: Sequence dependence and ph effects. *Biochemistry*, 37(26):9435–9444, 1998.
- [11] Hatim T Allawi and John SantaLucia. Nearest neighbor thermodynamic parameters for internal g · a mismatches in dna. *Biochemistry*, 37(8):2170–2179, 1998.
- [12] Hatim T Allawi and John SantaLucia Jr. Thermodynamics of internal c · t mismatches in dna. *Nucleic acids research*, 26(11):2694–2701, 1998.
- [13] Nicolas Peyret, P Ananda Seneviratne, Hatim T Allawi, and John SantaLucia. Nearest-neighbor thermodynamics and nmr of dna sequences with internal a · a, c · c, g · g, and t · t mismatches. *Biochemistry*, 38(12):3468–3477, 1999.
- [14] Addie Kingsland and Lutz Maibaum. Dna base pair mismatches induce structural changes and alter the free-energy landscape of base flip. *The Journal of Physical Chemistry B*, 122(51):12251–12259, 2018.
- [15] Craig A Gelfand, G Eric Plum, Arthur P Grollman, Francis Johnson, and Kenneth J Breslauer. Thermodynamic consequences of an abasic lesion in duplex dna are strongly dependent on base sequence. *Biochemistry*, 37(20):7321–7327, 1998.
- [16] Mark E Fornace, Nicholas J Porubsky, and Niles A Pierce. A unified dynamic programming framework for the analysis of interacting nucleic acid strands: enhanced models, scalability, and speed. *ACS Synthetic Biology*, 9(10):2665–2678, 2020.
- [17] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- [18] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of modern physics*, 62(2):251, 1990.
- [19] Abu Kausar, Rosalie D McKay, Jade Lam, Rohan S Bhogal, Alexandra Y Tang, and Julianne M Gibbs-Davis. Tuning dna stability to achieve turnover in template for an enzymatic ligation reaction. *Angewandte Chemie*, 123(38):9084–9088, 2011.
- [20] B Safeenaz Alladin-Mustan, Catherine J Mitran, and Julianne M Gibbs-Davis. Achieving room temperature dna amplification by dialling in destabilization. *Chemical communications*, 51(44):9101–9104, 2015.

- [21] Abu Kausar, Catherine J Mitran, Yimeng Li, and Julianne M Gibbs-Davis. Rapid, isothermal dna self-replication induced by a destabilizing lesion. *Angewandte Chemie (International ed. in English)*, 52(40):10577–10581, 2013.
- [22] Jan Engelhardt, Bent Andresen, and Steen Rasmussen. Thermodynamics and kinetics of lesion induced dna amplification (lida). In *Artificial Life Conference Proceedings 32*, pages 269–272. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , 2020.
- [23] Ditlev Hartmann Bornebusch, Christina Colaluca Sørensen, Peter Zingg, Gianluca Gazzola, Norman Packard, and Steen Rasmussen. Automated replication optimization for protocellular information system. In *ALIFE 2020: The 2020 Conference on Artificial Life*, pages 219–220. MIT Press, 2020.
- [24] Alexey V Cherepanov and Simon De Vries. Kinetics and thermodynamics of nick sealing by t4 dna ligase. *European Journal of Biochemistry*, 270(21):4315–4325, 2003.
- [25] Gregory JS Lohman, Stanley Tabor, and Nicole M Nichols. Dna ligases. *Current Protocols in Molecular Biology*, 94(1):3–14, 2011.
- [26] Ulla Christensen, Nana Jacobsen, Vivek K Rajawanshi, Jesper Wengel, and Troels Koch. Stopped-flow kinetics of locked nucleic acid (lna)–oligonucleotide duplex formation: studies of lna–dna and dna–dna interactions. *Biochemical Journal*, 354(3):481–484, 2001.
- [27] Albert L Lehninger, David L Nelson, Michael M Cox, et al. *Lehninger principles of biochemistry*. Macmillan, 2005.
- [28] Gheorghe Craciun and Martin Feinberg. Multiple equilibria in complex chemical reaction networks. *SIAM Journal on Applied Mathematics*, 66(4):1321–1338, 2006.
- [29] Patrick M Arnott, Himanshu Joshi, Aleksei Aksimentiev, and Stefan Howorka. Dynamic interactions between lipid-tethered dna and phospholipid membranes. *Langmuir*, 34(49):15084–15092, 2018.
- [30] Michael F Hagan, Aaron R Dinner, David Chandler, and Arup K Chakraborty. Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in dna. *Proceedings of the National Academy of Sciences*, 100(24):13922–13927, 2003.
- [31] James G Wetmur and Norman Davidson. Kinetics of renaturation of dna. *Journal of molecular biology*, 31(3):349–370, 1968.
- [32] Sophie Hertel, Richard E Spinney, Stephanie Y Xu, Thomas E Ouldridge, Richard G Morris, and Lawrence K Lee. The stability and number of nucleating interactions

- determine dna hybridization rates in the absence of secondary structure. *Nucleic Acids Research*, 50(14):7829–7841, 2022.
- [33] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.
- [34] MV Smoluchowski. Mathematical theory of the kinetics of the coagulation of colloidal solutions. *Z. Phys. Chem.*, 92:129–168, 1917.
- [35] Thomas Guérin, Olivier Bénichou, and Raphaël Voituriez. Non-markovian polymer reaction kinetics. *Nature chemistry*, 4(7):568–573, 2012.
- [36] Michael Rubinstein, Ralph H Colby, et al. *Polymer physics*, volume 23. Oxford university press New York, 2003.
- [37] Qingjia Chi, Guixue Wang, and Jiahuan Jiang. The persistence length and length per base of single-stranded dna obtained from fluorescence correlation spectroscopy measurements using mean field theory. *Physica A: Statistical Mechanics and its Applications*, 392(5):1072–1079, 2013.
- [38] Roman Shusterman, Sergey Alon, Tatyana Gavrinyov, and Oleg Krichevsky. Monomer dynamics in double-and single-stranded dna polymers. *Physical review letters*, 92(4):048303, 2004.
- [39] Bruno H Zimm. Dynamics of polymer molecules in dilute solution: viscoelasticity, flow birefringence and dielectric loss. *The journal of chemical physics*, 24(2):269–278, 1956.
- [40] Jean-Louis Sikorav, Henri Orland, and Alan Braslav. Mechanism of thermal renaturation and hybridization of nucleic acids: Kramers’ process and universality in watson-crick base pairing. *The Journal of Physical Chemistry B*, 113(12):3715–3725, 2009.
- [41] Manoel Manghi and Nicolas Destainville. Physics of base-pairing dynamics in dna. *Physics Reports*, 631:1–41, 2016.
- [42] Wenpeng Qi, Bo Song, Xiaoling Lei, Chunlei Wang, and Haiping Fang. Dna base pair hybridization and water-mediated metastable structures studied by molecular dynamics simulations. *Biochemistry*, 50(44):9628–9632, 2011.
- [43] Thomas Garel and Henri Orland. Generalized poland–scheraga model for dna hybridization. *Biopolymers: Original Research on Biomolecules*, 75(6):453–467, 2004.
- [44] Ernesto Carrillo-Nava, Yamilet Mejia-Radillo, and Hans-Jurgen Hinz. Dodecamer dna duplex formation is characterized by second-order kinetics, positive activation

- energies, and a dependence on sequence and Mg^{2+} ion concentration. *Biochemistry*, 47(50):13153–13157, 2008.
- [45] David Axelrod and MD Wang. Reduction-of-dimensionality kinetics at reaction-limited cell surface receptors. *Biophysical journal*, 66(3):588–600, 1994.
- [46] S Goldstein. Some two-dimensional diffusion problems with circular symmetry. *Proceedings of the London Mathematical Society*, 2(1):51–88, 1932.
- [47] Lloyd P Smith. Heat flow in an infinite solid bounded internally by a cylinder. *Journal of Applied Physics*, 8(6):441–448, 1937.
- [48] JC Jaeger and HS Carslaw. Heat flow in the region bounded internally by a circular cylinder. In *Proc. Royal Soc*, volume 61, page 223, 1942.
- [49] Cornelis A Emeis and PL Fehder. Microscopic mechanism for diffusion and the rates of diffusion-controlled reactions in simple liquid solvents. *Journal of the American Chemical Society*, 92(8):2246–2252, 1970.
- [50] K Razi Naqvi. Diffusion-controlled reactions in two-dimensional fluids: discussion of measurements of lateral diffusion of lipids in biological membranes. *Chemical Physics Letters*, 28(2):280–284, 1974.
- [51] David C Torney and Harden M McConnell. Diffusion-limited reaction rate theory for two-dimensional systems. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 387(1792):147–170, 1983.
- [52] Hisatake Sano and M Tachiya. Theory of diffusion-controlled reactions on spherical surfaces and its application to reactions on micellar surfaces. *The Journal of Chemical Physics*, 75(6):2870–2878, 1981.
- [53] Joaquin Ambia-Garrido, Arnold Vainrub, and B Montgomery Pettitt. A model for structure and thermodynamics of ssDNA and dsDNA near a surface: A coarse grained approach. *Computer physics communications*, 181(12):2001–2007, 2010.
- [54] PG Saffman and M Delbrück. Brownian motion in biological membranes. *Proceedings of the National Academy of Sciences*, 72(8):3111–3113, 1975.
- [55] A Sonnleitner, GJ Schütz, and Th Schmidt. Free brownian motion of individual lipid molecules in biomembranes. *Biophysical journal*, 77(5):2638–2642, 1999.
- [56] Jun Soo Kim and Arun Yethiraj. Crowding effects on association reactions at membranes. *Biophysical journal*, 98(6):951–958, 2010.

- [57] Dietmar Pörschke. Model calculations on the kinetics of oligonucleotide double helix coil transitions. evidence for a fast chain sliding reaction. *Biophysical chemistry*, 2(2):83–96, 1974.
- [58] Thomas E Ouldridge, Petr Šulc, Flavio Romano, Jonathan PK Doye, and Ard A Louis. Dna hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic acids research*, 41(19):8886–8895, 2013.
- [59] Michael S Jones, Brennan Ashwood, Andrei Tokmakoff, and Andrew L Ferguson. Determining sequence-dependent dna oligonucleotide hybridization and dehybridization mechanisms using coarse-grained molecular simulation, markov state models, and infrared spectroscopy. *Journal of the American Chemical Society*, 143(42):17395–17411, 2021.
- [60] Oleg N Temkin, Andrew V Zeigarnik, and DG Bonchev. *Chemical reaction networks: a graph-theoretical approach*. CRC Press, 1996.
- [61] Jinny X Zhang, John Z Fang, Wei Duan, Lucia R Wu, Angela W Zhang, Neil Dalchau, Boyan Yordanov, Rasmus Petersen, Andrew Phillips, and David Yu Zhang. Predicting dna hybridization kinetics from sequence. *Nature chemistry*, 10(1):91–98, 2018.
- [62] Hiroaki Hata, Tetsuro Kitajima, and Akira Suyama. Influence of thermodynamically unfavorable secondary structures on dna hybridization kinetics. *Nucleic Acids Research*, 46(2):782–791, 2018.
- [63] Alfonso Landeros, Timothy Stutz, Kevin L Keys, Alexander Alekseyenko, Janet S Sinsheimer, Kenneth Lange, and Mary E Sehl. BioSimulator.jl: Stochastic simulation in Julia. *Computer Methods and Programs in Biomedicine*, 167:23–35, 2018.

List of Figures

1.1	Organizational and functional closure in a protocell involves utilizing environmental resources and free energy for primitive metabolism and constructing building blocks for self-assembling containers and informational molecules. The container maintains proximity between metabolism and the informational system, which facilitates catalytic enhancement of metabolic rate kinetics and inheritance of catalytic capabilities across generations.	6
2.1	Chemical structure of Watson-Crick-Franklin base pairing. G · C pairs have three hydrogen and A · T pairs have two. The overall stability of hydrogen bonded pairs depends on their nearest base pairing context, due to the interplay between aromatic $\pi - \pi$ stacking and hydrogen bonding between subsequent pairs [5]. This fact underlies the Nearest Neighbor Model of DNA Thermodynamics.	9
2.2	Example secondary structure notations, figure from Nupack [7]	10
2.3	Example of possible loops composing a secondary structured complex. Here a complex of three strand. Figure from Nupack [7]	11
2.4	Intrastrand pseudoknotted tertiary structure (left) with polymer graph representation (right). A pseudoknotted structure can be recognised by checking crossover between lines in a polymer graph. Figure from [8].	12
2.5	An example of a C— · T mismatch (orange) and an abasic site flanked by C · G and A · T with opposing T base (yellow)	13
3.1	Isothermal turnover in DNA-templated ligation reactions using destabilizing templates, figure from [19].	18
3.2	LIDA system at the hybridization level of abstraction as depicted in [20]. (the figure has been modified from the original to highlight the reversibility of hybridization steps)	19
3.3	Schematic representation of the LIDA system with assigned nomenclature. T_1 refers to our target template, R_1 and R_2 refer to destabilizing oligomers, everything else follows the same pattern as in figure 3.2.	21
3.4	As shown in [24], T4 Ligase can seal nicks with a mismatched region up to 5 mismatched base pairs. These particular strands have been repaired with an observed kinetic rate of $k_{repair} = 4.7 \cdot 10^{-3} s^{-1}$. We'll use this value as a lower bound on possible ligation kinetics	24

3.5 Target template and oligomers sequences used in our description. The X label refers to a generic mismatch. We'll use a A · G mismatch for our free energy calculations. Figure from [20]	25
3.6 LIDA percent yield of target template over time against experimentally measured data from [20]	26
3.7 Realization of the LIDA system with rate constants as in section 3.2.1 and initial oligomer concentration $\rho(\text{oligo}, 0) = 0.1mM$ with shown time evolution of concentration for all involved species. Detailed explanation of the main steps involved can be found in the figures below.	26
3.8 Rapid equilibrium for oligomer duplexes. O_2 reaches a fast equilibrium with $K_E(R13) = 1.68 \cdot 10^6$ due to the presence of the mismatch. O_1 presents a higher equilibrium constant $K_E(R12) = 4.86 \cdot 10^7$, this is due to the external position of the abasic defect, acting as a dangling end.	27
3.9 Accumulation of intermediates at concentrations comparable with the initial template concentration.	27
3.10 A peak is observed in the concentration of intermediates, We found that the peak of concentration of the "right" Nicked duplex N_r always coincides with the zero for the second derivative of the curve representing the time evolution of duplex concentration. We use this point as indicative of the half-life of replication $\tau_{1/2}$	28
3.11 Local Sensitivity Analysis on Right Nicked Duplex shows strongest dependence on ligation rate k_{lig} and duplex de-hybridization rate $k_{\text{bd}_{\text{duplex}}}$	29
3.12 surface plot of LIDA replication half-life surface by varying k_{lig} and duplex destabilization k_{bd}	30
3.13 side views	31
3.14 Illustration of some of M-LIDA reactions on a surface. Bulk $R2$ hybridizes with $TR1$ making the left nicked duplex N_l , which ligates in the duplex D . After dehybridization the new strands will enter two new cycles of replication with their complementary oligomers.	32
3.15 Circled species are membrane-anchored	32
3.16 An example of a DNA-lipid conjugate structure with cholesterol tail.	33

3.17 Two perspectives of the same surface plot for target-probe hybridization rate k_3^+ versus surface hybridization rate k_2^+ . For the considered parameter range, strong exponential dependence of replication half-life is observed on two-dimensional hybridization rates. Weak exponential dependence is observed on three-dimensional ones. We explore parameters: two-dimensional hybridization rate $k_2^+ = [2 \cdot 10^7, \dots, 2 \cdot 10^{12}]$, three-dimensional hybridization rate $k_3^+ = [2 \cdot 10^6, \dots, 2 \cdot 10^8]$ Three surfaces are produced at different values of duplex stability k_{bd} , in order from top to bottom we have $k_{bd} = (0.003, 0.03, 30)$	36
4.1 Cartoon illustration of the two step model presented in this chapter. Two initially unbound strands perform first nucleation attempts and subsequently close their configuration reaching the duplex state. This process is in reality underlined by mid-way meta-stable states, which are not considered in the two-step model. Figure adapted from [32].	38
4.2 Rotational degrees of freedom of a stiff cylinder representing a persistence segment of ssDNA. In green the reactive surface area corresponding to nucleobases. θ is the azimuth angle and ϕ the angle of inclination. l_p is the persistence length of ssDNA (4.11)	43
4.3 A water molecule impedes the formation of a hydrogen bond in an A · T base pair, slowing down the process. Figure from [42]	44
4.4 (b) The x axis refers to the number of formed base pairs, the F axis refers to the complex free energy. ΔG_0 refers to the increase in free energy due to the joining of a new base pair. E_b refers to the activation energy to make a new base pair, related to solvation (see <i>Activation Energy</i> paragraph of section 4.2)	46
4.5 Geometric representation corresponding to the Sano-Tachiya diffusion-controlled rate on a spherical surface. We'll use in our description $a = R_g(n)$ gyration radius of a strand with n nucleotides. Figure from [52]	50
4.6 Confrontation of LIDA kinetics realizations with three-dimensional and two-dimensional hybridization rate constants. Crossed in red, the nicked duplex peak corresponding to replication half-life.	55
5.2 Molecular Dynamics simulation of an inchworming transition from the initial slide with register $r = 4$ to the duplex state $r = 0$. (i) Initial binding of 10 base pairs at register 4. (ii) Owing to thermal fluctuations, three base pairs form a register 0 at the expense of register 4. (iii) The bulge loop is passed down the helix. (iv) After the final base pair in register 4 breaks, register 0 is able to form all its possible base pairs. Figure from MD simulations performed by Ouldridge [58]	59

5.3	Molecular Dynamics simulation of an inchworming transition from the initial slide with register $r = 4$ to the duplex state $r = 0$. (i) A complex is initially in register 10. (ii) Additional base pairs form at register 0. (iii) Further base pairs form at register 0 at the expense of register 10. (iv) The final base pairs at register 10 are broken and the duplex forms. Figure from MD simulations performed by Ouldridge [58]	60
5.4	Two complementary strands long 40 nucleotides have a stable off-register complex at register $r = -8$. The left side tails, long 27 nucleotides for the top strand and 19 nucleotides for the bottom strand, can undergo direct on-register nucleation, initiating the transition from the off-register $r = -8$ to the on-register duplex $r = 0$. On the right side two complementary base pairs form at register $r = -2$, creating an <i>interior loop</i> (see sec. 2.1). Below two equivalent secondary structure notations referring to the pictured complex: top) <i>run-length-encoded dot-bracket notation</i> , bottom) plain <i>dot-bracket notation</i>	62
5.5	Two dimensional representation of the overlapping spherical caps representing the pseudoknotting probability. D_b is the duplexed segment length, $R_g(c_1)$ and $R_g(c_2)$ refer to gyration radiiuses of single stranded tails.	62
6.1	At the top, a register zero zipping state with three already formed base pairs and the two candidate base pairs (shown in red) that will form. At the bottom the corresponding structures to newly formed base pairs. Bottom left: formation of an G · C pair. Bottom right: formation of an A · T pair	65
6.2	A <i>zipping leaf</i> is pictured. Originating from simplex state S (in red) with target state a structure with 5 adjacent base pairs denoted by D (in green), neighbors of S are the first 5 nucleation attempts (Nx in yellow) with only one base pair. These will subsequently zip (Zx in cyan) one base pair at a time until reaching the target.	66
6.3	Example zipping graph generated from complementary sequences 5'-AACCAACC-3' - 5'-GGTTGGTT-3'. Numbers inside nodes indicate the number of base pairs of the underlying secondary structure formed. Node with 0 pairs indicates the root simplex state and node with 8 pairs indicates the target duplex state. <i>Colours of transitions</i> indicate the following: <i>green</i> indicates off-register nucleation events, which can either lead to subsequent zipping or be abortive, <i>orange</i> indicates on-register ($r = 0$) nucleation events, <i>purple</i> indicates zipping transitions, <i>dark green</i> indicates sliding transitions through inchworming and/or pseudoknotting pathways (chap. 5).	67

6.4	Flow chart like representation of the Object-Oriented Programming implementation of the zipping graph model. Titles of blocks refer to implemented classes, in each a description of the main tasks performed by the class. Arrows show dependencies.	69
6.5	A) after an initial abortive nucleation at register 0, the 10-mer hybridizes on register 0. B) a 10-mer zips on register 2, from there it inchworms to the duplex. C) a 14-mer nucleates and zips one base pair at register -10 to then unzip and re-nucleate at register 0, abort, re-nucleate and then zip to the duplex. D) a 14-mer performs several abortive nucleations and one abortive zipping at register 3. It reaches a meta-stable off-register state at register 7 from which it transitions to the duplex. In this case both inchworming and pseudoknotting are possible transition pathways.	72
6.6	The initial sigmoidal growth as a consequence of the beta-distributed first passage time.	74
6.7	Model predicted versus experimental hybridization rates. In blue the identity line. This scatter plot corresponds to following error measures: $\rho_{mod,exp} = 0.6832$ Pearson correlation $RMSE = 0.1389$ Root Mean Squared Error $F2_{acc} = 0.9512$ Distance 2 accuracy Error measures relative to experimental data are not available. Error measures relative to simulation results are too low to be visible due to optimization of number N_m of simulations to perform. Solid dots refer to outliers which are points that the zipping graph model doesn't predict well.	76
6.8	Error histogram produced by computing the difference $\log_{10}(k_{hyb}^{mod}) - \log_{10}(k_{hyb}^{exp})$. Negative values correspond to under-evaluation and positive values to over-evaluation. We see that on average the zipping graph model with the found parameters is slightly under-evaluating.	76
6.9	Comparison of the zipping graph scatter plot against the models [32, 61, 62].	78
6.10	Cartoon representation of two-dimensional collision and nucleation. We expect bottom base pairs to nucleate only in-register, resulting in higher on-register probability of nucleation and consequent on-register zipping, increasing rates. We also expect for such nucleations that the only relevant steric angle is the azimuthal one.	81
6.11	Two-dimensional zipping graph predicted kinetics are plotted together with three dimensional results and experimental data of three-dimensional kinetics. Membrane-anchoring results in an order 10 increase in the kinetics of hybridization and negative sequence dependance of hybridization rates, i.e. longer sequences hybridize slower than shorter ones.	83
6.12	Caption	84

6.13 3D LIDA kinetics derived from the zipping graph model has 74.55 minutes replication half-life. 2D LIDA kinetics on the other hand has 54.41 minutes replication half-life.	85
7.1 Long-lived pseudoknotted state at geometrically compatible registers 6 and -6. Both registers present the stacking of 6 base pairs, resulting in a long-lived competition between the two registers, as none of the two strongly dominates the other in terms free-energy measured stability. Figure from [58]	89

*In loving memory of my Father, Angelo.
I wish you could read this. We miss you.*