Coarse-grained modelling of nucleic acids

Petr Šulc
University College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2014

Dedicated to my parents

Acknowledgements

Foremost, I would like to express my gratitude to my supervisors, Ard Louis and Jonathan Doye, for their guidance and support throughout my doctoral studies at Oxford.

I would further like to offer my special thanks to my colleagues Flavio Romano, Tom Ouldridge and Lorenzo Rovigatti, with whom I had the opportunity to collaborate on multiple projects during my doctorate and whose advice and help was always greatly appreciated.

I have also greatly benefited from the interactions with collaborators and researchers in the field of DNA biophysics and nanotechnology. In particular, I want to thank Jon Bath, Filip Lankaš, Alex Lucas, Eyal Nir, Agnes Noy, Joseph Schaeffer, Niranjan Srinivas, Andrew Turberfield, and Erik Winfree.

Finally, I am grateful for stimulating discussions with my office-mates as well as other D.Phil. students and post-docs from the Doye and Louis groups: Kamal Dingle, Pedro Fonseca, Ryan Harrison, Christian Matek, Majid Mosayebi, Aleks Reinhardt, Steffen Schaper, John Schreck, Ben Snodin, Ioannis Zacharoudiou and Irwin Zaid.

I am grateful for the financial support from Scatcherd European Award and Bobby Berman Fellowship at University College, Oxford.

Last but not least, I want to thank my friends and family for their continuous support and encouragement.

Abstract

This thesis considers coarse-grained models of DNA and RNA, developed in particular to study nanotechnological applications as well as some important biophysical processes. We first introduce sequence-dependent thermodynamics into a previously developed coarse-grained rigid base-pair model of DNA. This model is then used to study sequence-dependent effects in multiple DNA systems including: the heterogeneous stacking transition of single strands, the fraying of a duplex, the effects of stacking strength in the loop on the melting temperature of hairpins, the force-extension curve of single strands, and the structure of a kissing-loop complex. We further apply the DNA model to study in detail the properties of an autonomous unidirectionally propagating DNA nanotechnological device, called the "burnt bridges motor". We then apply the coarse-graining methods developed for the DNA model to construct a new sequence-dependent coarse-grained model of RNA, which aims to capture basic thermodynamic, structural and mechanical properties of RNA molecules. We test the model by studying its thermodynamics for a variety of secondary structure motifs and also consider the force-extension properties of an RNA duplex. This RNA model allows for efficient simulations of a variety of RNA systems up to hundreds or even thousands of base-pairs. Its versatility is further demonstrated by studying the thermodynamics of a pseudoknot folding, the formation of a kissing loop complex, the structure of a hexagonal RNA nanoring, and the unzipping of a hairpin.

Contents

Chapter 1

Introduction

DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) molecules are key components of living cells [1]. The discovery of the double helical structure of DNA and its significance for information storage of genetic material by Watson and Crick is generally considered as one of the most important scientific achievements of the twentieth century [2]. Much effort was then dedicated to the study of how the genetic information is translated into proteins, a process which involves RNA molecules at several stages as well. The study of genetic code cumulated in the sequencing of the human genome in 2001 [3], and has been advancing rapidly ever since.

It was long believed (the central dogma of molecular biology) that DNA codes for messenger RNA (mRNA) that is then translated into proteins. However, only 2% of human DNA codes for proteins and the human genome has roughly the same number of protein coding regions as much simpler organism *C. Elegans* which has only about 1000 cells. It is now thought that the complexity of an organism scales with the percentage of its genome that does not code for proteins [4]. Interestingly, there is substantial evidence that there are multiple regions of the genome that are transcribed into RNA molecules which are not further involved in protein production, but are themselves the final product. Thousands of these "non-coding RNAs" (ncRNA) have been identified and their function is a very active field of research [5, 6].

Furthermore, as we will see, DNA and RNA molecules are well-suited for con-

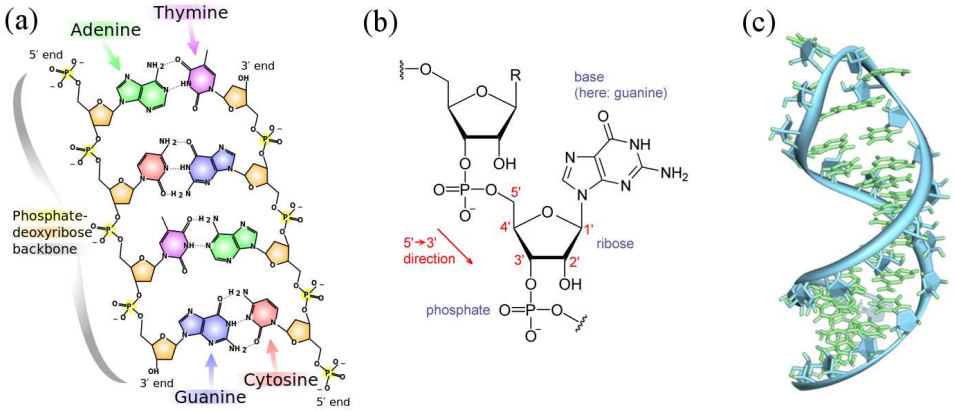

Figure 1.1: A schematic illustration of (a) the DNA duplex in a planar projection, (b) an RNA nucleotide and (c) an RNA hairpin, where the stem adopts an A-helical structure.

struction of artificial self-assembled nanostructures [7, 8]. In addition, the mechanical properties of DNA, and to a lesser extent RNA, have been intensely studied over the last decade by physicists because they present a well-defined model system with which to study the fundamental physical properties of single molecules. Hence, because of the importance of DNA and RNA molecules in nature as well as in designed nanosystems, their thermodynamic, mechanical and structural properties have been the subject of an intensive experimental and theoretical research efforts.

1.1 Structure, composition and function of DNA and RNA molecules

DNA strands are composed of a deoxyribose sugar-phosphate backbone with four different kinds of bases attached: adenine (A), thymine (T), cytosine (C) or guanine (G), as illustrated in Fig. 1.1(a). The RNA strand is similar, but instead of deoxyribose, has a ribose sugar in its backbone, which has an additional OH group on $2'$ carbon atom, as shown in Fig. 1.1(b). Four different types of bases can be attached to the RNA backbone. Three are the same as in DNA (adenine, cytosine and guanine) while one is different: uracil (U), which differs from a thymine by a single methyl group.

These bases have highly anisotropic mutual interactions that are responsible for

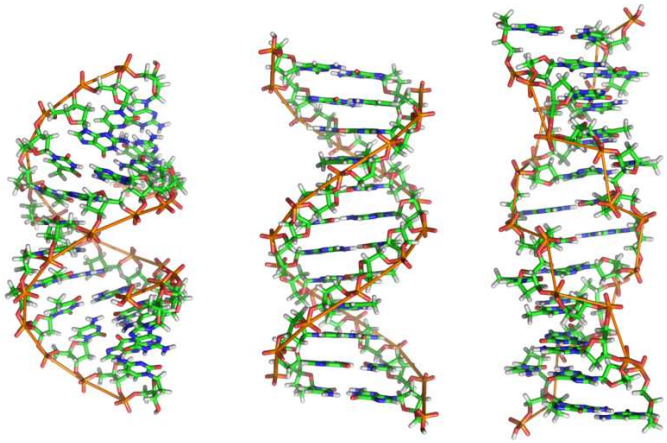

Figure 1.2: A schematic illustration of the A, B and Z-helical DNA duplex (from left to right) respectively.

the formation of non-trivial structures primarily through hydrogen bonding and stacking interactions. Both RNA and DNA can form double-helical molecules, stabilized by hydrogen bonds between complementary Watson-Crick base pairs: AT (AU in the case of RNA) and GC. For RNA, wobble base pairs (GU) can also stabilize the duplex form.

Despite the similarity in building blocks, RNA and DNA properties are different. DNA forms different types of helices, with the biologically active ones being right-handed A or B-helix and left-handed Z-helix [9], which are schematically shown in Fig. 1.2. In nature, it is most often found in the B-helical conformation, which has a rise per base pair of 3.4 Å and a pitch estimated around 10.5 base pairs per turn [9, 10]. The presence of the extra OH group on the sugar of RNA nucleotides causes the RNA duplexes to form A-helical (or similar A′-helical) structures [9, 11, 12] (as shown for example for the stem of a hairpin in Fig. 1.1(c)). In contrast with the B-helical structure, the A-helix has a smaller rise per base pair (2.8 Å) and a smaller helical twist per base pair, leading to a pitch of approximately 11 base pairs per turn. Furthermore, in the A-helix the centers of the base pairs do not lie on the helical axis and they are more inclined with respect to the axis than in the case of B-helix. The structure of RNA duplexes will be further discussed in detail in Chapter 5.

The RNA duplexes with Watson-Crick base pairs are on average more stable than the DNA duplexes and have a larger persistence length of about 53 nm at 0.5 M salt as opposed to 44 nm for DNA [13, 14]. RNA strands can furthermore have interactions that involve the OH group on the backbone [15] as well as interactions involving wobble base-pairs and thus exhibit more complexity than DNA in terms of the number of possible structures that they can form.

DNA's role in biological systems is to store genetic information and it is most often found in a double-helical form. RNA is more versatile than DNA. It is essential to gene transcription, where it stores information as messenger RNA (mRNA), which is used as a template for protein synthesis. Respective amino acids are delivered by transfer RNAs (tRNAs) onto ribosomes, where ribosomal RNA is essential in their synthesis into the resulting proteins. Some RNA molecules can also affect genetic regulation. The mechanisms of how RNA can regulate gene expression are just beginning to emerge. One identified mechanism is via an siRNA complex (an RNA duplex with short single-stranded overhangs). It forms an RNA-protein complex that degrades mRNA corresponding to a specific gene. Other possible regulation mechanism involves a short single-stranded microRNA (miRNA) which binds directly to mRNA strands and affects their translation [15]. Moreover, as famously discovered by Cech, RNA can also act as a catalyst [16]. Since it can accomplish both storage of genetic material like DNA (for example, some viruses store their information as RNA strands) as well as metabolism (like proteins), it has been postulated that earlier in the evolution life was based on RNA before DNA-based organisms appeared (The RNA World hypothesis). Furthermore, RNA is used for molecular recognition, where RNA specifically binds to other molecules. Some RNA molecules have a scaffolding function, allowing for the assembly of ribonucleoprotein complexes, such as ribosome or telomerase. In telomerase, RNA furthermore acts as a template from which new DNA bases are synthesized [15, 1].

One reason that RNA can achieve so many different functions is that, in contrast to DNA, most naturally occurring RNA molecules are single-stranded and fold into complex structures that contain double-helical segments as well as loops, bulges, junctions and often various tertiary structure interactions that stabilize the folded molecule.

1.2 Nucleic acid nanotechnology

While RNA is used by nature for a wide variety of different tasks, DNA was, for a long time, seen merely as a passive information carrier. This picture began to change when it was realized that mechanical properties of DNA can play a role in regulation, as was shown for example by DNA loop formation in lactose operons [17]. But a bigger paradigm shift came in 1982, when Seeman suggested that the specificity of DNA hybridization could be harnessed to form artificial structures, including DNA crystals [7]. This marked the beginning of the field of DNA nanotechnology. Multiple artificial DNA structures and devices have been designed over the past three decades and the field is now well established.

One might have expected RNA nanotechnology to grow as rapidly as DNA nanotechnology, given the versatility of RNA roles in the cell. This has not been the case, however, as RNA is more difficult to handle in experimental conditions. Due to its single-stranded nature, it is more susceptible to degradation by hydrolysis as well as by ubiquitous RNA catalyzing enzymes (RNases). Artificial nucleic acid synthesis is also more expensive for RNA than for DNA. However, nanotechnology devices and structures that use RNA molecules as building blocks are now being developed as well and the field of RNA nanotechnology is becoming increasingly popular [8]. Moreover, nanostructures and nanodevices based on DNA/RNA hybrid duplexes have also been realized [18, 19, 20].

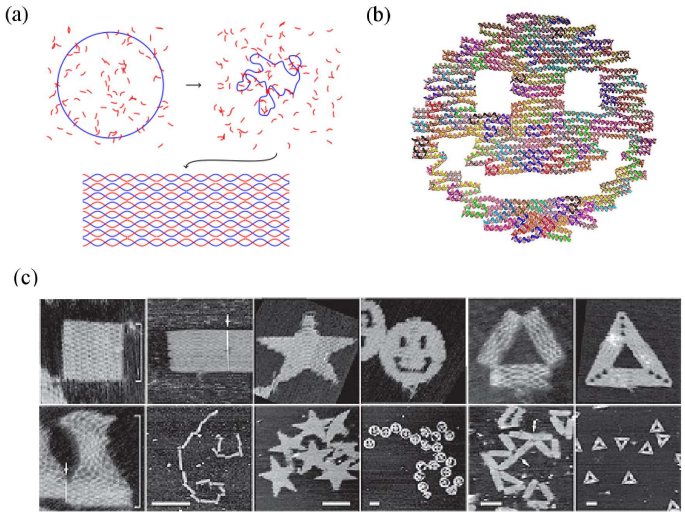

Figure 1.3: (a) A schematic illustration of DNA origami assembly. Parts of short staple strands (shown in red) bind to their complementary sections on the long substrate stand (shown in blue), thus forming the final structure (rectangle in this case). (b) A picture of DNA origami structure "Smiley face", which consists of the substrate strand bound to the staple strands. Different strands are colored with different colors. (c) AFM images of various DNA origami structures, taken from the original reference where the technique was introduced [21] and reproduced with the permission of Nature Publishing Group.

In the section below, we give a brief overview of some of the structures and devices that have been assembled out of DNA and RNA.

1.2.1 DNA nanotechnology

The design space of DNA strands is enormous. The number of possible sequences for a strand of length L grows as 4^L, so that, for example, a short oligomer of $L = 15$ has approximately 10^9 different sequences, which grows to about 10^{12} by adding just 5 extra bases to $L = 20$. One consequence is that it is possible to design a set of single strands so that a very particular configuration is the global free-energy minimum of the system. One can also make sure that other metastable structures are far enough in free-energy that they do not significantly compete. Using relatively straightforward design principles, an enormous range of nanostructures have been realized simply by cooling solutions of single-stranded DNA (ssDNA) based on this principle. Finite-size

structures have been designed that assemble from a small number of short oligonucleotides [22]. A technique known as DNA "origami" (illustrated in Fig. 1.3) uses a single long strand and many short staple strands and has been used to assemble a range of structures, whose applications range from targeted drug delivery to tools for single-molecule experiments [21, 23, 24, 25, 26]. Structures assembled from a large number of short stands (referred to as single stranded tiles or DNA "Lego") have also been recently realized [27, 28]. Three dimensional "DNA cages" were also obtained by assembling multiple strands [22, 29].

Additionally, hierarchical self-assembly has been also successfully realized, where first multiple strands assemble into DNA tiles, which have several single-stranded regions that can bound to the complementary regions on other tiles. Via the self-assembly of such tiles, large one-dimensional ribbons of lengths up to 5 μm [30], 2-dimensional arrays [31, 32] and 3-dimensional crystals [33] have been realized. The tiles can be designed to allow for carrying out an algorithmic self-assembly, thus realizing a cellular automaton with DNA molecules [34].

DNA nanotechnology is not confined to static structures. Dynamic systems have also been developed. Duplex formation and toehold-mediated strand displacement [35] (a process in which a strand is removed from a duplex by a competing strand that can form more base pairs with the complement) allow a DNA system to respond to its environment. In particular, these processes can couple chemical change to mechanical operations and have the potential to process signals. DNA nanotweezers [36], a switch that can be cycled through its closed and open states by the sequential addition of two types of strand, demonstrated the principle. "Clocked" addition of strands [37, 38, 39, 40, 41, 42] or permutation of external conditions [43, 44] have since been used in the design of a number of active systems, including some in which the mechanical change has been harnessed to induce unidirectional motion along a track [41, 42, 44]. Recently, autonomous devices and walkers that function without

external forcing (by catalyzing the equilibration of an out-of-equilibrium system) have also been created [45, 46, 47, 48, 49, 50, 51, 52, 53, 54]. Such devices are sometimes referred to as DNA motors or DNA walkers and we will present one particular example of such a system, the DNA burnt-bridges motor, in detail in Chapter 4 where we use the coarse-grained DNA model to study it.

The high degree of parallelization and the ability to interface directly with biological and molecular systems make DNA-based computation promising. In 1994, Adleman showed that DNA strands could be used to encode a Hamiltonian path problem, which was then solved upon mixing of the strands [55]. Since then, much work has gone into developing DNA-based logic circuits [56], with a DNA neural network that can recognize simple patterns having recently been developed [57]. DNA logic has also been combined with walking devices to produce systems that can select from distinct pathways at a junction depending on solution conditions, or properties of the walker itself [54, 58].

The versatility of making artificial DNA structures is also starting to be exploited to study novel properties of bulk materials and solutions [59, 60, 61], and shedding new light on fundamental physical processes like gel formation [62, 63].

1.2.2 RNA nanotechnology

RNA nanotechnology aims to construct nanoscale structures and devices by using RNA strands [8]. The number of possible structures that can be realized from a particular RNA sequence is larger than for a corresponding DNA sequence, as RNA can also form GU wobble base pairs and have numerous tertiary structure interactions. The resulting most probable structure of RNA strands is hence difficult to predict: secondary structure (i.e. the list of base pairs in the folded state) prediction methods reach about 73 % accuracy and predicting three dimensional structure remains even more challenging task [8, 64, 65]. Therefore functional motifs from known biologically

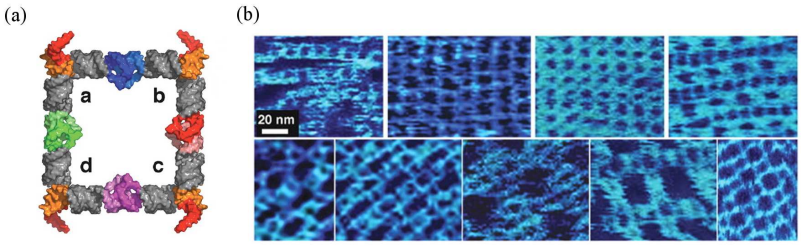

Figure 1.4: (a) A schematic illustration of an RNA square tile made of 4 RNA molecules. (b) AFM visualization of lattices of RNA square tiles. The images are from Ref. [67] and are reproduced with the permission of The American Association for the Advancement of Science.

occurring structures are often exploited rather than designing *de novo* sequences that would fold into a particular functional three dimensional structure [66].

One class of nanotechnology systems realized in RNA nanotechnology follows successful approaches previously used in DNA nanotechnology. In particular, a hierarchical self-assembly with RNA was demonstrated in [67], where RNA tiles were designed to self-assemble into designed patterns (as shown in Fig. 1.4). RNA bundles assembled from RNA monomers were also realized [68]. RNA nanocubes were also successfully assembled [69], following the example of DNA cages. The strand displacement mechanism, an essential mechanism for most active DNA nanodevices, has been also used in RNA. In particular, it was recently shown *in vitro* that cascades of RNA strand displacement reactions, triggered by the presence of an mRNA strand with a particular sequence, can be used to produce an siRNA complex [70]. The promising application *in vivo* is a conditional knockout of a gene by the RNA silencing mechanism in the presence of mRNA created by a transcription of the triggering gene.

Due to the versatile role of RNA in the cell, there has been great interest in designing RNA nanomachines that exploit RNA sequences and structures which are known to interfere with proteins or other RNA that naturally occur inside the cell. Future potential RNA nanotechnology applications *in vivo* are in diagnosis, targeted drug delivery and intracellular computation and regulation. A step in this directions is

the the aforementioned gene silencing strand displacement cascade. Another example is a designed RNA ring consisting of six RNA monomers, with each of the monomers having the possibility of being functionalized (i.e. by attaching an image-reporter molecule or by adding an RNA sequence that interacts with proteins or other RNA) [71]. It is also possible to incorporate the desired RNA sequences into the DNA that then gets transcribed into RNA [8], thus effectively delivering designed RNA nanodevices into the cell.

1.2.3 RNA/DNA hybrid nanotechnology

RNA and DNA strands can form hybrid duplexes, which are less stable than the corresponding RNA duplex, but can be actually more stable than corresponding DNA duplex for some sequences [72]. The resulting hybrid duplex has typically an A-helical structure [18, 73].

Following the development of RNA nanotechnology devices, hybrid RNA/DNA nanostructures have also been created [18, 19, 20]. As it is cheaper to synthesize DNA strands than RNA strands, it is possible that in future *in vivo* applications a particular functional RNA sequence will be delivered as a part of a hybrid duplex. Furthermore, the hybrid DNA and RNA duplexes are not as easily degraded by enzymes that break up RNA duplexes in the cell [20]. An example of a successful DNA/RNA nanotechnology design is a hybrid DNA/RNA origami structure, with 1071 nucleotides long scaffold RNA strand which binds with multiple short DNA staple strands that were designed so that the resulting structure is a rectangle, a triangle or a ribbon [18].

Hybrid RNA/DNA tiles were used to assemble two-dimensional arrays as well as dodecahedrons [19]. Finally, two different hybrid duplexes with complementary toeholds were also designed. After being mixed together they produce a DNA duplex and an siRNA complex that can interfere with the RNA silencing pathway. The functionality of this designed system was demonstrated *in vivo* [20].

1.3 Modelling of DNA and RNA

Although DNA nanotechnology shows great promise as a field, it is far from being fully mature. In particular, optimization of nanostructure assembly and nanodevice operation will be vital if they are to prove generally useful. DNA nanostructures and nanodevices are typically designed using well-established thermodynamic models of DNA duplex stability, such as the unified nearest-neighbor model of SantaLucia [74]. However, nanodevices and nanostructures can involve non-trivial multi-stranded complexes with pseudoknots [75] or complex internal loops whose stabilities have not yet been incorporated into thermodynamic models. Moreover, non-equilibrium processes can be important in these systems. Furthermore, the three-dimensional structure of a DNA complex may result in tension or compression forces [76] that cannot be described without an explicit three-dimensional representation of the system.

RNA nanostructures and devices are designed either by trying to estimate the most stable secondary structure formed by a particular sequence or by exploiting previously experimentally determined structures of RNA strands. If the designed RNA molecule is transcribed *in vivo*, the folding into its functional structure can be further complicated by long-lived metastable structures that form before the strand is fully synthesized. It is difficult to predict the resulting structure as well as the metastable intermediates, making the *de novo* design of the systems challenging.

Computer simulations provide controllable access to time and spatial resolutions that are not accessible in experiments. Simulations of DNA/RNA nanotechnology systems therefore have the potential to offer a valuable insight into aspects of their operation and design, provided the computational model accurately describes the relevant properties of the system.

Many theoretical and computational approaches have been developed to study nucleic acid and we give a general overview below. We then focus on coarse-grained models in more detail.

1.3.1 Theoretical and computational approaches to the study of nucleic acids

In an important series of works the thermodynamics of DNA [77, 74] and RNA [78, 79, 80, 81, 82, 83, 84] secondary structure was characterized in terms of nearest-neighbor models. The nearest-neighbor model is a two-state model, where the free-energy difference between single-stranded and bound (folded) state is calculated by summing the contributions from each nearest-neighbor set of two base pairs together with terms for helix initiation and various structural features such as loops and bulges. These nearest-neighbor models are the basis of various tools for the prediction of the most stable secondary structures and duplex or hairpin melting temperatures [85, 86, 87, 88, 89, 90, 91, 92]. Such tools typically use dynamic programming approaches to find the secondary structure with minimal free energy. Furthermore, some tools have been extended by adding simple kinetic descriptions to the nearest-neighbor thermodynamics, allowing folding transitions to be modeled [93, 94]. Although these methods are typically very fast, the fundamentally discrete nature of the description and the lack of structural and mechanical detail places a limit on what they can treat.

At the most fine-grained level, quantum chemistry calculations can be used to study the interactions between nucleotides [95, 96, 97, 98, 99, 100]. While they provide valuable information about the ground state energies at a high level of detail, they are computationally demanding and do not allow for the study of dynamical processes involving breaking and forming of base pairs.

Molecular simulation packages such as AMBER [101] or CHARMM [102], which retain an all-atom representation of the nucleic acids, the water solvent and explicit ions, but use empirical classical force fields to model their interactions, are extensively used for computational studies of both DNA and RNA as well as their interactions with proteins [103]. Although faster than quantum chemistry methods, they still are computationally very demanding and the time scales they can currently access

are of the order of μs, while many biologically and technologically relevant processes happen at the ms timescale or longer. At the moment, simulations of rare events, such as the breaking of a single base pair, remain at the limit of what is possible. Moreover, while the forcefields are improving, they are still under development so that different versions can generate different behavior [104, 105, 106]. A recently developed approach [107] combines fully atomistic representation with hierarchical Monte Carlo sampling, where different series of moves are used to move whole sections of a molecule (such as all atoms contained in one stem) at once. Such methods have for instance been used to study the effects of mutations in a sequence on the conformational freedom of a tRNA molecule and of a nanosquare composed of four tRNAs [108].

In order to access longer timescales relevant to rare events, such as the breaking of base pairs or the formation of large structures, one needs to use a more coarse-grained description. In this approach, atoms are incorporated into a reduced set of degrees of freedom that experience effective interactions. Solvent molecules are often integrated out. Such models always present a compromise between accuracy, efficiency and the level of detail, which determines their scope. Coarse-grained models have been developed both for DNA [109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124] and RNA [125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139]. While these models cannot describe the system at the same level of detail as atomistic simulations, they allow one to study much larger systems and address rare events. Both oxDNA and oxRNA models, described in this thesis, fall within this category. A detailed comparison of oxDNA to other coarse-grained DNA models was provided in [140] and we will give an overview of the existing coarse-grained RNA models in Section 1.3.3.

Finally, continuous models, which have been developed for DNA [141, 142, 143] and RNA [144], completely neglect the detailed chemical structure but allow for analytical treatment in the thermodynamic limit, and have been used to study macro-

scopic properties such as melting temperatures or properties of DNA under stress.

1.3.2 Parametrization of coarse-grained models

We introduced in the previous section different levels of detail that are used for the modelling of nucleic acids. The level of detail varies for coarse-grained models as well. The groups of atoms are replaced by their coarse-grained representation, ranging from one particle representing a single nucleotide [125, 112, 131, 137, 127] to one nucleotide being replaced by a group of pseudoatoms. The number of pseudoatoms varies from three-site-per nucleotide models [128, 134, 131, 136, 139, 124, 113, 115], where the particles in the simulation represent the sugar, phosphate and base of the nucleotide respectively, to more finer models where up to ten beads per nucleotide can be used [126, 133, 145].

Once the representation is selected, one needs to parametrize the interactions between the particles in the model. One possible way, which we call the "bottom-up" approach [146], attempts to parametrize the interactions to reproduce values obtained from finer-grained representations, for example from fully atomistic simulations with Amber or GROMACS or from an ensemble of experimentally determined atomic structures. The quantities that such a parametrization aims to reproduce can be for example the distribution of distances and angles between specified groups of atoms. In contrast, "top-down" models are designed to reproduce specific measured properties of the system, for example the persistence length or the melting transition from a duplex (or a hairpin) to a single-stranded state. This type of parametrization is employed by the oxDNA model, as discussed in Section 2.1, and it will be also used to develop the oxRNA model in Chapter 5.

While the bottom-up approach provides a more direct link between the coarse-grained representation and fully atomistic structures, one needs to bear in mind that the fine-grained representations to which bottom-up models are parametrized do not necessarily correctly reproduce other properties of the system. For instance, it is not

known how well the available atomic force fields can reproduce the free-energy difference between the dissociated single strands and the duplex state, because such a transition is too computationally expensive to be sampled with a fully atomistic representation. Furthermore, such a bottom-up model can be biased towards reproducing the structures that were picked for their fitting ensemble.

It is imporant to stress that coarse-grained models will always present a compromise and cannot capture all properties of the system accurately, a general phenomenon which has been called a "representability problem" [147, 148]. Each model hence has its specific domain of applicability for which it was designed and care needs to be taken in interpreting the results obtained from coarse-grained simulations and relating them to experiment.

1.3.3 Coarse-grained models of RNA

We review here some previously developed coarse-grained models of RNA that cover a wide spectra of level of detail as well as of areas of applicability. While a large number of models focus on the structure prediction for RNA, there are also models that take into account the thermodynamic properties in the parametrization of the interactions.

Knowledge-based coarse-graining uses the information extracted from experimentally determined crystal structures to develop potentials, usually with the goal to predict the folded structure for an RNA sequence, either *de novo* or with some additional input of data from the user. An example of such an approach is the NAST model [125] which represents each nucleotide as a single pseudoatom in the simulation and uses a statistical potential, inferred from known structures of RNA molecules, that depends on the distances and angles between the nucleotides. This model requires the secondary structure and tertiary contacts of the final folded RNA structure as an input for the folding simulation. It has been used to study RNA structures of

up to 158 nucleotides. While it was able to reproduce the structures of folded RNA, it was not parametrized to reproduce their thermodynamic properties.

Similarly, the model of Xia *et al.* [126], which uses 6 beads to represent a nucleotide, has interactions parametrized to reproduce known RNA structures. Xia *et al.* were able to predict the tertiary structure of several RNA molecules of lengths up to 122 nucleotides by using simulated annealing to attempt to find the global potential energy minimum of the structures in their coarse-grained model. The resulting structures were then refined by a simulation with a fully atomic representation. The thermodynamic properties of the coarse-grained model were not reported.

The recently developed model of Taxilaga-Zetina *et al.* [127] represents each nucleotide as a single bead interacting with knowledge-based interaction potentials extracted from a distribution of distances and angles observed in a set of crystallographic structures of RNA. The model was used to simulate folding of strands of up to 34 nucleotides into hairpin and pseudoknot structures.

The TOPRNA model by Mustoe and collaborators [128], developed with three-sites-per-nucleotide representation, has knowledge-based potentials parametrized to a set of experimentally known RNA structures. It was used to sample conformations of RNA structures that contain bulges of different sizes.

Finally, some knowledge-based methods combine together various structural motifs from database of experimentally determined RNA structures to predict folded RNA structure for a given sequence. The algorithms match RNA residues with known structure with a particular section of the RNA sequence. These residues are then combined to form the final structure. For example, Parisien and Major [129] used such an approach to predict the secondary and tertiary structure of RNA strands with up to 50 nucleotides. The FARFAR method [130] further uses sampling with a fully atomistic representation of the respective RNA residues in order to obtain the final structure. It successfully predicted *de novo* folded structures for RNA sequences of

size up to 20 nucleotides.

An alternative approach for model development is to use fully atomistic simulations to parametrize the effective interactions between coarse-grained representations of groups of atoms. Such an approach was adopted, for example, by Paliy *et al.* [131], who presented different levels of coarse-graining, using either one or three beads per nucleotide. The interactions between beads were fitted to reproduce the probability distribution of their mutual orientations and distances, calculated from from a simulation with a fully atomistic representation. The authors were then able to simulate the conformations of an RNA nanoring structure which consisted of 330 nucleotides.

The HiRe-RNA model [132, 133] represents each nucleotide as 6 or 7 beads with empirically chosen interactions based on a combination of atomistic simulations and known structures. It reproduces the structure of RNA duplexes and was used to simulate the association and dissociation of small oligonucleotides (16 base pairs). The model further allows the reconstruction of a fully atomistic representation of an RNA molecule from its coarse-grained representation. The model was also used to study some transitions in RNA, although a direct link between the parameters and experimental melting temperatures has not yet been made.

The above mentioned models were parametrized to structure, either through comparison to experiment, or to atomistic simulations from which thermodynamic quantities are hard to extract. While that is useful for the structure of folded RNA complexes, it makes it hard to compare with available experimental data on RNA thermodynamics, or to simulate reactions involving multiple RNA strands. The next set of models do include explicit thermodynamic information in their parametrization.

The coarse-grained model of Ding *et al.* [134] uses three beads (sugar, phosphate and base) to represent each nucleotide and has been used to study the folding of various RNA structures, including tRNA and pseudoknots, of sizes up to 100 nucleotides. The parametrization of the interactions combines a knowledge-based approach with

a parametrization of the interaction strengths to the free energies of base pairs taken from the nearest-neighbor model. Their simulation algorithm furthermore takes into account explicitly the free-energy cost for closing a loop as predicted by the nearest-neighbor model. This added free-energy contribution does not come from the model's interactions and hence ties the use of the model to this particular simulation algorithm.

The model of Hyeon and Thirumalai [135] also uses three beads per nucleotide. Its interaction strengths are based on nearest-neighbor model parameters. The model was used to study mechanical unfolding of hairpins. Recently, the model was extended by Denesyuk and Thirumalai [136] with interactions parametrized using thermodynamic data from pseudoknot and hairpin melting experiments combined with the free energies in the nearest-neighbor model for RNA thermodynamics [81]. The new model has been used to study the thermodynamics of folding of a 34-nucleotide pseudoknot. The model also includes explicit electrostatic interactions and can also represent tertiary structure contacts such as hydrogen bonds in non-canonical base pairs. We note that Hyeon *et al.* also developed the SOP model for RNA [137], which only uses one site per nucleotide, to study larger systems. The interactions in the model were set to a given energy scale and were not compared with RNA thermodynamics. The SOP model was used to study the mechanical unfolding of a 421-base ribozyme [137].

The nearest-neighbor model was also used to parametrize the lattice-based model of Cao and Chen [138] which represents the conformations of RNA as a self-avoiding walk on a 3D-lattice. This model was used to compute the heat capacities from partition functions for different mutants of the so-called 72 RNA structure, which were found to be in good agreement with experimental measurements. It was further used to study the free-energy landscape at different temperatures for a 76-nucleotide P5abc RNA structure.

Finally, the lattice model of Jost and Everaers [139] is parametrized to reproduce

the nearest-neighbor model thermodynamics. The parametrization was verified by studying thermodynamics of ten RNA hairpins and an ensemble of structures with varying internal loop sizes. The model was then used to study folding pathways of a 76 nucleotide long tRNA and a pseudoknot. While lattice based models allow for an efficient sampling of the possible conformations, the structural description of the RNA is necessarily limited by the requirement that it is placed on a lattice.

Most of the existing coarse-grained RNA models are aimed at the correct prediction of the most probable folded structure for a given RNA sequence. In these cases, the thermodynamics of RNA duplex or hairpins formation was either not explicitly considered, or was used to guide parameter choice which was then tested on a few selected systems. Of the described models, the most detailed verification of the thermodynamics was done for the model of Jost and Everaers. We further note that mechanical properties have not been reported for any of these RNA models.

1.4 Thesis outline

All the coarse-grained models discussed above have their strengths and weaknesses, depending on the systems for which they were designed. To describe processes commonly occurring in nucleic acid nanotechnology, a model needs to be able to accurately reproduce duplex association from two single strands as well as formation of hairpins and other commonly found structural motifs. For most purposes in nanotechnology, it is not necessary for the model to reproduce particular atomic-level details of DNA or RNA structures. It is more important to correctly capture the thermodynamic properties of the system, as most nanotechnological systems are designed to assemble into a structure that corresponds to the free-energy minimum at given temperature. Active nanodevices involve forming and breaking of base pairs, which the model also needs to be able to reproduce. Finally, the mechanical properties of the strands as represented by the model need to correspond to the known behavior of RNA and

DNA. In particular, while RNA and DNA duplexes are fairly rigid polymers, single-stranded DNA and RNA are very flexible, permitting the creation of structures with sharply bent regions (such as a hairpin loop). We note that some nanotechnological systems consist of up to several thousands of base pairs. To simulate such large systems, as well as rare events of bond breaking and creation, the model needs to be computationally efficient.

This thesis aims to develop coarse-grained models of DNA and RNA that meet the above listed requirements and study their biophysical properties as well as their applications to nanotechnology. We start by extending the previously developed DNA coarse-grained model by Ouldridge, Doye and Louis [149, 150, 140], called oxDNA, which was designed for DNA nanotechnological applications. In Chapter 2, we first give a brief overview of the properties of this model, which we refer to as the average-base oxDNA model, as it does not distinguish between different nucleotides in terms of the interaction strengths. We then describe an automated parametrization technique that allows us to introduce sequence-dependent interactions into the oxDNA model. We next test the new sequence-dependent parameterization by comparing melting simulation results with available thermodynamic data for a large set of different DNA sequences.

In Chapter 3, we use the sequence-dependent oxDNA model to study different sequence-dependent phenomena in DNA. In particular, we explore the heterogeneous stacking transition of single strands, the tendency of a duplex to fray at its melting point, the effect of stacking strength in the loop on the melting temperature of hairpins, the force-extension properties of single strands and the structure of a kissing-loop complex.

We use the average-base oxDNA model in Chapter 4 to study an active DNA nanotechnology device, the "burnt-bridges DNA motor".

Finally, we introduce a novel RNA coarse-grained model, oxRNA, in Chapter 5.

This coarse-grained RNA model follows the approach adopted in the development of the oxDNA model and it aims to capture basic thermodynamic, mechanical and structural properties of RNA with a minimalistic representation and pairwise interactions. The model is designed to study RNA nanotechnology systems as well as processes that involve RNA strands in a biological setting. We test the versatility of the oxRNA model in Chapter 6, where we use the model to study the thermodynamics of a pseudoknot folding, the formation of a kissing complex, the structure of an RNA nanoring, and the unzipping of a hairpin.

The simulation code that implements oxDNA and oxRNA is released for public use at dna.physics.ox.ac.uk. The results obtained in Chapters 2, 3 and 4 have been accepted for publication [151, 152] and the material from Chapters 5 and 6 is currently under review [153].

Chapter 2

Introducing sequence-dependent thermodynamics into a coarse-grained DNA model

In this chapter, we introduce a sequence-dependent parametrization of the oxDNA model.

We first present the original coarse-grained DNA model of Ouldridge *et al.* [149, 150, 140]. We then motivate the need for including sequence-dependent interactions in the model by comparing thermodynamics of DNA duplexes with different base contents. We describe in detail the fitting procedure that we developed for the sequence-dependent interactions and test the parametrization on melting of duplexes and hairpins, the latter being a case to which the model was not fitted. The examples of systems studied with the sequence-dependent version of the oxDNA model are then provided in Chapter 3.

2.1 Average-base coarse-grained DNA model

The coarse-grained DNA model oxDNA is described in detail in [150, 140]. It was designed to capture the structural, thermodynamic and mechanical properties of DNA in both the single- and double-stranded forms. It represents DNA as a string of nucleotides, where each nucleotide (sugar, phosphate and base group) is a rigid body with interaction sites for backbone, stacking and hydrogen-bonding interactions. The

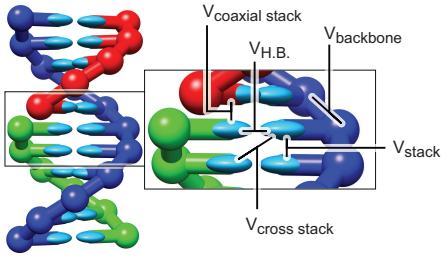

Figure 2.1: The figure shows schematically the interactions between nucleotides in the coarse-grained DNA model for two strands in a duplex. All nucleotides also interact with a repulsive excluded volume interactions.

potential energy of the system is

$$\begin{aligned} V_{\text{oxDNA}} &= \sum_{\langle ij \rangle} \left(V_{\text{backbone}} + V_{\text{stack}} + V'_{\text{exc}} \right) + \\ &+ \sum_{i,j \notin \langle ij \rangle} \left(V_{\text{H.B.}} + V_{\text{cross stack}} + V_{\text{exc}} + V_{\text{coaxial stack}} \right), \end{aligned} \tag{2.1}$$

where the first sum is taken over all nucleotides that are nearest neighbors on the same strand and the second sum comprises all remaining pairs. The interactions between nucleotides are schematically shown in the Fig. 2.1, and the explicit forms can be found in [150, 140]. The hydrogen bonding ($V_{\text{H.B.}}$), cross stacking ($V_{\text{cross stack}}$), coaxial stacking ($V_{\text{coaxial stack}}$) and stacking interactions (V_{stack}) explicitly depend on the relative orientations of the nucleotides as well as on the distance between interaction sites. The backbone potential V_{backbone} is an isotropic spring that imposes a finite maximum distance between neighbors, mimicking the covalent bonds along the strand. The coaxial stacking term is designed to capture stacking interactions between non-neighboring bases, usually on different strands. All interaction sites also have isotropic excluded volume interactions V_{exc} or V'_{exc}.

The coarse-grained oxDNA model was derived in a "top-down" fashion, as discussed in Chapter 1, i.e. by choosing a physically motivated functional form, and then focusing on correctly reproducing the free-energy differences between different states of the system, as opposed to a "bottom-up" approach that starts from a more

detailed representation of DNA and typically focuses on accurate representation of local structural details.

The model was fitted to reproduce DNA behavior at a salt concentration ($[Na^+] = 0.5\,M$) where the electrostatic properties are strongly screened, and it may be reasonable to incorporate them into a short-ranged excluded volume. Such high salt concentrations are typically used in DNA nanotechnology applications, hence motivating this approach.

The model allows for the formation of a helical duplex from two single strands. The helical structure of the duplex is an emergent property of the model. The equilibrium distance between backbone sites (6.4 Å), interacting with the V_{backbone} potential, is larger then the equilibrium distance between stacking sites (3.4 Å) that interact with V_{stack}. In order to satisfy both interaction potentials, the model adopts a helical structure with a rise per base pair and a twist per base pair that are determined by the two equilibrium distances of the backbone and stacking potentials. The competition between the equilibrium distances of the covalent bonds between the atoms forming the backbone and the stacking interaction is the reason why in nature DNA adopts helical structure [154], so it is important to have the coarse-grained model reproduce this behavior for the same reasons. We note that many bottom-up models that use B-helical DNA as the starting structure end up with angular dependent potentials to enforce the helical structure. They thus reproduce the correct structure, but for the wrong reasons. The equilibrium distances for oxDNA model were selected to reproduce the parameters of a B-helical DNA structure, with a rise per base pair 3.4 Å and a pitch of 10.34 base pairs per turn. The right-handedness of the helix is enforced in the model by a chirality term in the V_{stack} interaction, which is 1 if the bases stack in a right-handed fashion and 0 otherwise [140].

The model allows for base pairing only between Watson-Crick complementary bases, but otherwise does not distinguish between bases in terms of interaction

strengths. We refer to this parameterization of the model as the "average-base" model, as it is suited to study processes for which sequence heterogeneity is of secondary importance.

The average-base oxDNA model reproduces the melting temperature of duplexes of lengths ranging from four to twenty base pairs within 1 °C precision when compared to the prediction of the averaged SantaLucia's model (the nearest-neighbor model of SantaLucia [77], where all the free-energy terms were averaged over all possible sequences). It also captures the width of the transition from a single stranded state to a duplex within few Kelvin, with the model's width of the transition being slightly $(1 - 2\,^{\circ}\mathrm{C})$ narrower than the widths of the yields curve predicted by the averaged SantaLucia's model.

Even though the model was fitted to the duplex melting transition, it also reproduces the thermodynamics of secondary structure motifs with high accuracy. The melting temperatures of hairpins are underestimated by approximately 3 °C but trends with loop length or stem length are very well approximated. The model reproduces the destabilization caused by internal mismatches, terminal mismatches and bulges in the duplex. The change in melting temperature of duplexes with these motifs is captured with within 6 °C precision. The stabilization of a duplex by a dangling end is captured within 1 °C precision. We note that with respect to the absolute temperature 300 K, the differences of the melting temperatures predicted by the averaged SantaLucia's model and the oxDNA model are at most 2%.

In addition, the model was fitted to reproduce the structural and mechanical properties of double- and single-stranded DNA such as the persistence length and the twist-modulus. The experimentally measured persistence length of DNA is reported to be around $45 - 50$ nm at high salt concentrations, corresponding to about 130 to 150 base pairs (with rise 3.4 Å per base pair) [155, 14]. The persistence length of a DNA duplex in oxDNA is about 123 base pairs, close to the reported experimental

values. The force-extension curve of the duplex in the oxDNA model reproduces the extensible worm-like chain behavior [156], which was observed for DNA duplexes in experiments as well [157, 158]. The model also reproduces DNA overstretching transition at 74 pN at 23 °C, about $6 - 7$ pN higher than the experimentally reported overstretching force [159] for similar temperature and salt concentration.

It should be noted that the model neglects several features of the DNA structure and interactions due to the high level of coarse-graining. Specifically, the double helix in the model is symmetrical rather than the grooves between the backbone sites along the helix having different sizes, and all four nucleotides have the same structure.

The main purpose of this chapter is to go beyond the average-base parametrization of oxDNA by introducing sequence-dependent interaction strengths into the model.

2.2 Sequence-dependent effects in duplex thermodynamics

Many biological processes and technological applications of nucleic acids rely on sequence heterogeneity. It is well-known that AT and GC pairs have different relative binding strength [9], with the latter being stronger because of the presence of three rather than two interbase hydrogen bonds. Moreover, the stacking interactions that drive the coplanar alignment of neighboring bases are known to show significantly different behavior depending on sequence [9]. Furthermore, a strand of DNA possesses directionality, e.g. the phosphates of the backbone connect to the 3′ and 5′ carbon atoms in the sugars. Interactions within a strand are therefore distinct when the bases are permuted: for example, the interaction of neighboring GT bases depends on whether the G is in the 5′ direction with respect to the T or *vice versa*. Besides thermodynamic properties, it has been observed that mechanical and structural properties such as flexibility, helical twist and even helix type are also influenced by the sequence [154, 160, 161, 162, 163].

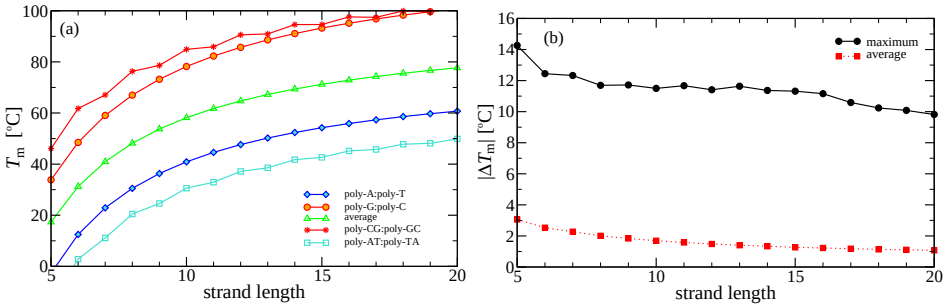

Figure 2.2: (a) Melting temperatures versus duplex length as predicted by SantaLucia's nearest-neighbor model [77] for a duplex consisting of poly(A), poly(AT), poly(C) or poly(CG) and an average sequence. (b) Maximum (circles) and average (squares) difference in melting temperature for strands with nucleotide positions randomly permuted. The terminal base pairs are kept the same, thus neutralizing different end effects. Data were generated by selecting 50000 random sequences at each length and permuting each 5000 times. The differences show the importance of the order of the nucleotides in the sequence.

To highlight the effects of sequence on the thermodynamics of DNA, we point out that the melting temperature of two oligomers with the same length but different sequences can vary by more than 50 °C, as shown in Fig. 2.2(a) where we compare the melting temperatures of poly(A), poly(G), poly(CG) and poly(AT) sequences of various lengths at an equal strand concentration of 3.36×10^{-4} M. These melting temperature differences are only marginally diminished with increasing length and are exploited *in vivo*, where, for example, it has been observed that initiation sites of transcription are often composed of a higher than average number of AT pairs [1].

Note that besides the number of AT and GC base pairs, the actual order of nucleotides in the sequence is also important: two sequences of the same length and the same number of AT and GC base pairs can still have melting temperatures that differ by more than 10 °C, as shown in Fig. 2.2(b).

Given these large variations, it is important to have a model that captures at least the thermodynamic effects of sequence. We note that some of the other coarse-grained models of DNA that have been developed do include sequence effects in various level of detail, including sequence-dependent base-pairing interactions [114] and

also sequence-dependent stacking [116] and cross-stacking interactions [115]. An extension [164] of the model in Ref. [114] also has base pair deformability parametrized to the values determined by analysis of DNA-protein crystal complexes [160]. In contrast to these models, the oxDNA model was specifically developed for applications in DNA nanotechnology and was primarily designed to represent the single- to double-stranded transition in a sufficiently physical manner. We will now introduce a parametrization of the oxDNA model that captures the sequence-dependence of DNA thermodynamics.

2.3 Parametrization of sequence-dependent interactions

We choose to perform a thermodynamic parametrization of the sequence-dependent interactions, aiming to reproduce melting temperatures of short DNA duplexes. We seek the parameters that best reproduce the melting temperatures as predicted by SantaLucia's model [77], described in the next section, which we treat as an accurate fit to experimental data on the melting of duplexes of different length and sequence. We restrict sequence dependence to the strength of the base pairing ($V_{\mathrm{H.B.}}$) and stacking (V_{stack}) interaction terms, keeping all other parameters fixed to the values of the original average-base parametrization fit.

2.3.1 SantaLucia's nearest-neighbor model

In an important series of papers, SantaLucia [77, 74] summarized the results of multiple melting temperatures of DNA oligomers, and also presented a nearest-neighbor model that reproduces the results of melting experiments (hereafter referred to as the SL model). The model assumes that DNA can exist in two states, either single-stranded or in duplex form, and gives a standard free-energy change of formation $\Delta G(T)$ of the duplex with respect to the single strands as a function of temperature.

The expected yields of duplexes can then be calculated as a function of temperature through the relation:

$$\frac{[AB]}{[A][B]} = \exp(-\Delta G^{\ominus}(T)/RT), \tag{2.2}$$

where $[A]$ and $[B]$ are molar concentrations of single strands, $[AB]$ is the molar concentration of the duplex, $\Delta G^{\ominus}$ is the standard Gibbs free-energy change and R is the molar gas constant. This result assumes the system is dilute enough to behave ideally apart from associations, a condition fulfilled in the vast majority of experiments.

The SL model assumes that $\Delta G^{\ominus}(T)$ is a sum of contributions, one for each base-pair step formed in a duplex with respect to the single-stranded state, along with corrections for end effects. A base-pair step consists of four bases; for example, the base-pair step GT/AC stands for a section of duplex that has GT bases on one strand and AC on the complementary strand. The SL model has 10 unique base-pair nucleotide steps: AA/TT, AT/AT, TA/TA, GC/GC, CG/CG, GG/CC, GA/TC, AG/CT, TG/CA, GT/AC, where pairs are given in 3′-5′ order along the strands.

The contribution to $\Delta G^{\ominus}(T)$ of each term is divided into a temperature-independent enthalpy and entropy, so that the overall form of $\Delta G^{\ominus}(T)$ is given by

$$\Delta G^{\ominus}(T) = \Delta H^{\ominus} - T\Delta S^{\ominus}, \tag{2.3}$$

with $\Delta H^{\ominus}$ and $\Delta S^{\ominus}$ being the (temperature-independent) sum of the individual contributions to the enthalpy and entropy respectively. The SL model is a *two-state* model, in that it considers two regions of state space (the duplex and single-stranded states) and assumes that there is a constant enthalpy and entropy difference between the two. In other words, it neglects the variation in enthalpy within the bound and unbound sub-ensembles.

The melting temperature T_{m} for a given sequence is defined in the SL model as the temperature at which half of the strands in the system are in the duplex state

and the other half are in the denatured state. Using this definition, the SL model has an average absolute deviation of 1.6 °C when compared to known experimental melting temperatures of 246 duplexes with lengths between 4 and 16 base pairs [74]. We fit to the T_{m} as predicted by the SL model, rather than having to re-analyze the original experimental data. This choice allows us to fit to a large ensemble of different sequences whose melting temperatures we estimate using the SL model.

We emphasize that, in contrast to the SL model, our model itself does not exhibit ideal two-state behavior. Although we observe a large difference in the typical energies of single-stranded and duplex states, allowing us to clearly differentiate the two, we also observe significant variation within these sub-ensembles. Both single-stranded and duplex states have multiple microscopic degrees of freedom, which respond differently to changes in temperature. For instance, we observe fraying of duplexes (Sec. 3.2) and that the single strands undergo a stacking transition (Sec. 3.1). The net effect is that the ΔH and ΔS of transitions that would be inferred from our model are not temperature independent, unlike in the SL model.

We note that other models for the prediction of DNA melting temperatures exist, such as the recently developed nearest-neighbor model of [165], which uses the mechanical unzipping of DNA hairpins to infer the individual base pair step free energies. Our parametrization procedure only requires estimates of the melting temperature for a large set of DNA sequences and could be also used to fit our model to the melting temperature predictions of [165]. We could also fit directly to the melting experiments. However, the differences in predicted melting temperature with the SL model and the experiments are so small that it is not worth in this case.

2.3.2 Fitting of the parameters

Our model was originally parametrized to reproduce the melting temperatures of average sequences as predicted by the SL model. Since the SL model is constructed on the level of base-pair steps, it cannot be used to differentiate between intrastrand

interactions within a step: for example, AA and TT or AG and CT. We therefore set the stacking interaction strengths of bases that belong to the same base-pair step to be equal in our parametrization procedure.

To parametrize our coarse-grained DNA model's potential V_0 (Eq. 2.1), we scale the V_{stack} and $V_{\text{H.B.}}$ interaction terms by the factors α_{ij} and η_{ij} respectively, i.e.

$$V_{\text{H.B.}} \rightarrow \alpha_{ij} V_{\text{H.B.}} \tag{2.4}$$

$$V_{\text{stack}} \rightarrow \eta_{ij} V_{\text{stack}}, \tag{2.5}$$

where α_{ij} and η_{ij} are constants for a given nucleotide pair ij. There are therefore 10 parameters η_{ij} (as shown in Table 2.1) and two parameters α_{CG} and α_{AT} to fit. Making the cross-stacking interaction sequence-dependent would also influence melting temperatures, but as we will discuss later, sequence-dependent stacking and base-pairing interactions provide enough parameters to obtain results in almost complete agreement with the predictions given by the SL model. To fit the 12 coefficients η_{ij} and α_{ij}, we used a set $\mathcal{S}$ of oligonucleotides of lengths 6, 8, 10, 12 and 18 for which we calculated the (salt-adjusted) melting temperatures using the SL model. The set contained 2000 randomly generated sequences for each of lengths 8, 10, 12, 18 and all 4160 sequences of length 6. The set was then reduced to contain only heterodimers, leaving 12 022 sequences in total. We chose to remove homodimers (self-complementary sequences) for convenience, because the inference of the bulk melting temperatures from simulations of the formation of a single duplex is different from that for heterodimers, as discussed in [166].

We select the parameter set that minimizes the function:

$$f(\alpha_{ij}, \eta_{ij}) = \sum_{s \in \mathcal{S}} |T_m^s(\text{SL}) - T_m^s(\alpha_{ij}, \eta_{ij})| \tag{2.6}$$

where $T_m^s(\text{SL})$ is the melting temperature of the oligonucleotide s in the set $\mathcal{S}$ as predicted by the SL model and $T_m^s(\alpha_{ij}, \eta_{ij})$ is the melting temperature predicted by our model with sequence-dependent base pairing and stacking potentials $\alpha_{ij} V_{\text{H.B.}}$

and $\eta_{ij}V_{\text{stack}}$. To accurately fit α_{ij} and η_{ij}, we hence need estimates of the melting temperatures of many different sequences for many different values of the interaction parameters.

If one simulates a system consisting of two complementary strands in the simulation box at exactly the melting temperature then the ratio of observed duplex states to single-stranded states

$$\Phi = \frac{N_{\text{duplex}}}{N_{\text{single}}}, \tag{2.7}$$

should be equal to 2 for heterodimers and 1 for homodimers. The value of 2 for heterodimers is a correction for finite size effects that arise when one simulates only two strands instead of a bulk ensemble at the same average concentration [166, 167, 168]. The correction assumes that the density of strands is low enough that they behave ideally apart from association.

To calculate melting temperatures for the large set of sequences $\mathcal{S}$ we employed a histogram reweighting method [169, 170]. We generated once, for each duplex length considered, a set of 5000 single-stranded and 10 000 duplex configurations $\mathcal{C}_{\text{single}}$ and $\mathcal{C}_{\text{duplex}}$. The configurations in $\mathcal{C}_{\text{single}}$ and $\mathcal{C}_{\text{duplex}}$ were sampled from the Boltzmann distribution of strands of sequence s_0 at the melting temperature T_0 using the average parametrization (i.e., $\alpha_{ij} = 1$ and $\eta_{ij} = 1$). Simulations were performed in a cell that gave a concentration of 3.36×10^{-4} M for each strand. Twice as many duplex as single-stranded states were sampled because they appear in exactly this ratio in a simulation of two strands at the melting temperature of a given sequence in the average model (T_0). Sampling was done at sufficiently large intervals that the configurations in $\mathcal{C}_{\text{single}}$ and $\mathcal{C}_{\text{duplex}}$ were uncorrelated.

In order to find the ratio $\Phi_s(T, \alpha_{ij}, \eta_{ij})$ for a sequence s at temperature T with a parameter set α_{ij} and η_{ij} that corresponds to a potential $V(\alpha_{ij}, \eta_{ij}, T)$, states in $\mathcal{C}_{\text{single}}$ and $\mathcal{C}_{\text{duplex}}$ were reweighted by the factor

$$w_{l,s}\left(T, \alpha_{ij}, \eta_{ij}\right) = \exp\left(\frac{V_0^{l,s_0}(T_0)}{k_{\text{B}}T_0} - \frac{V^{l,s}(\alpha_{ij}, \eta_{ij}, T)}{k_{\text{B}}T}\right), \tag{2.8}$$

where $V_0^{l,s_0}(T_0)$ is the energy of the l-th state generated at temperature T_0 using the sequence s_0 in the average model, and $V^{l,s}(\alpha, \eta, T)$ is the sequence-dependent potential evaluated on the same l-th state for the sequence s. Note that both interaction potentials are a function of temperature because the stacking interaction term in the model is temperature dependent [150, 140]. The configurations used in Eq. 2.8 are generated at T_0 with V_0 and s_0, but each is counted with a weight that corresponds to the desired set of new parameters.

The ratio of the duplex to single-stranded states for a given temperature T and parameters α_{ij}, η_{ij} becomes

$$\Phi_s(T, \alpha_{ij}, \eta_{ij}) = \frac{\sum_{l \in \mathcal{C}_{\text{duplex}}} w_{l,s}\left(T, \alpha_{ij}, \eta_{ij}\right)}{\sum_{k \in \mathcal{C}_{\text{single}}} w_{k,s}\left(T, \alpha_{ij}, \eta_{ij}\right)} \tag{2.9}$$

where the index l runs through all generated duplex states while k runs through all generated single stranded states. Using this method, $\Phi_s(T, \alpha_{ij}, \eta_{ij})$ can be generated for a set of temperatures and interpolated in order to find T such that $\Phi_s(T, \alpha_{ij}, \eta_{ij}) = 2$, which is by definition the melting temperature T_m of a given duplex.

The histogram reweighting method assumes that the ensemble of configurations generated at temperature T_0 with potential V_0 for sequence s_0 is also representative of the state space of the system at temperature T and potential $V(\alpha, \eta, T)$ for sequence s. To check whether we included enough states, we compared the estimation of the melting temperature by histogram reweighting of 15 000 states to an estimation which only used 6000 different states. For a test case of 71 000 sequences of oligonucleotide lengths 8, 12 and 18, the mean absolute deviation of the difference between the predicted T_m was smaller than 0.1 °C, suggesting that the choice of 15 000 states provides a large enough ensemble for estimating the melting temperatures, at least on average.

To find a set of parameters that minimize function f defined in Eq. 2.6, we used an adaptive simulated annealing algorithm [170] which consists of the following steps:

1. Randomly perturb one of the parameters (α_{ij}, η_{ij}) to obtain $\left(\alpha'_{ij}, \eta'_{ij}\right)$.

2. Accept the perturbation if $r < \exp\left(-\beta\left(f(\alpha'_{ij}, \eta'_{ij}) - f(\alpha_{ij}, \eta_{ij})\right)\right)$ where r is a random number uniformly distributed in $[0, 1]$.

3. After certain number of steps, adapt β.

The parameter β of the simulation is decreased after several steps if the acceptance ratio of randomly generated parameters is smaller than a given bound, or increased if acceptance ratio is higher than a given bound. This prevents one from being stuck in a local minimum of the function f and helps sample through the space of all available parameters.

We first fitted the base-pairing strengths α_{CG} and α_{AT} while holding the stacking parameters constant. Then we fitted the 10 stacking parameters η_{ij} in a second step. The separate fitting of the two sets of parameters simplifies the fitting procedure, as the converged values for α_{ij} provide an initial point for the stacking parameters fitting. It also allows us to compare the performance of a model where only the base-pair interaction strengths are sequence-dependent to the one where both base-pairing and stacking interactions are sequence-dependent.

We note that our fitting procedure requires the ability to efficiently estimate melting temperatures. The histogram reweighting method, using the generated states, takes only about 1 s to calculate the melting temperature of a given sequence. This is a huge reduction in computer time as compared to umbrella sampling simulations [171], which were used in the parametrization of the original average-base model [150]. The umbrella sampling simulation samples multiple single- to double-stranded transitions for a given oligomer and requires around two weeks of CPU time to calculate the melting temperature to within 0.3°C accuracy for the sequence lengths that we considered for our parametrization. Thus our histogram re-weighting methodology provides the crucial speed-up that made the parametrization possible.

Base pairing	α_{ij}
AT	0.8292
GC	1.1541
Stacking	η_{ij}
GC	1.027
CG	1.059
AT	0.947
TA	0.996
GG, CC	0.978
GA, TC	0.970
AG, CT	0.982
TG, CA	1.009
GT, AC	1.019
AA, TT	1.042

Table 2.1: Summary of the final parameters that were fitted to reproduce melting temperatures of randomly chosen oligonucleotides as predicted by the SL model. Base steps are in $3'$-$5'$ direction.

2.3.3 Parametrization results

While the parameters α_{CG} and α_{AT} were fairly robust to details of the optimization procedure, the parameters η_{ij} were more sensitive. In order to uniquely determine these parameters we selected the set with the smallest average error on an additional test set of 95 958 sequences that included all sequences of lengths 5, 6, 7 and 8 for which the SL model predicts a T_{m} greater than 0 °C for the concentration 3.36×10^{-4}M, plus a set of randomly generated sequences of lengths 10, 12 and 18. The final set of parameters η_{ij} and α_{ij}, as introduced in Equations (2.4) and (2.5), is shown in Table 2.1.

Figure 2.3 compares a histogram of the difference

$$\Delta T_{\mathrm{m}} = T_m(\alpha_{ij}, \eta_{ij}) - T_{\mathrm{m}}(\mathrm{SL}) \tag{2.10}$$

between the melting temperatures $T_m(\alpha_{ij}, \eta_{ij})$, calculated by our coarse-grained model (using histogram reweighting) and the T_{m}(SL) of the SL model, determined for each of the 95 958 sequences in our test set. The blue dashed curve shows our model's performance when only the base pairing interactions are sequence-dependent (parameters

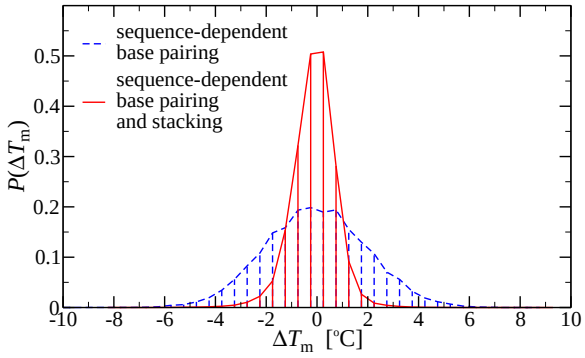

Figure 2.3: The histogram shows the performance of the fitted DNA coarse-grained model for the set of 95 958 test sequences. ΔT_{m} is the difference in the melting temperature predicted by the coarse-grained model and by the SL model. The blue dashed curve corresponds to a model where only hydrogen-bonding interactions were parametrized and the red curve corresponds to the model where the stacking interactions are also sequence-dependent (using values from Table 2.1).

α_{CG} and α_{AT} from Table 2.1) and the stacking parameters η_{ij} are all set to unity. The red solid curve shows the histogram when the melting temperatures are calculated by our model with both hydrogen bonding and stacking sequence-dependent parameters. The standard deviation of the distribution of ΔT_m with sequence-dependent base-pairing and average stacking is 2 °C, while the standard deviation for the case where stacking is also sequence-dependent is 0.85 °C. This compares to a standard deviation of 8.6 °C for the original average-base model when compared to full sequence-dependent data. We note that although the average deviation is very small, there are a number of melting temperatures in our set that differ significantly more than one would expect from a Gaussian distribution with this standard deviation. These outliers are typically highly repetitive sequences.

Since the SL model has an average absolute deviation of 1.6 °C when compared to experimental melting temperatures of 246 duplexes of lengths between 4 and 16, there is little point in trying to further improve our predictions with respect to it. That it is possible to reproduce the predictions of the SL model with our set of 12 parameters also implies that it would not be appropriate to introduce sequence dependence for other terms in the interaction potential by fitting only to T_{m}(SL). Instead, other

physical input would be needed.

It is also important to point out that, as discussed previously, by fitting to a model which considers only base pair steps it is not possible to distinguish between, for example, AA or TT stacking strengths, which are known to be different [172]. Even though we treat stacking within base pair steps equally, our method in principle allows the stacking interaction for each individual stacked pair to be parametrized differently. But in order to do this fitting, new experimental data is needed. We further discuss the parametrization of stacking interactions in Sec. 3.3.

2.4 Tests of the parametrization

We test the performance of our sequence-dependent parametrization by comparing the melting temperatures of selected duplexes, as well as for hairpins, to which the model was not directly fitted.

We have also tested the structural and mechanical properties of double-stranded DNA (away from thermodynamic transitions) on a randomly generated sequence with around 50% GC-content and confirmed that they are not changed with respect to those of the original average-base parametrization. So our double-stranded persistence length remains approximately 125 base pairs, and the B-DNA structure produced by the model is the same as in the average-base model [150, 140]. On the other hand, the structural and mechanical properties of single-stranded DNA do differ from those of the average model, and are studied in Sec. 3.1 and 3.4 in the next chapter.

2.4.1 Duplex melting

To further test our histogram reweighting method, we calculated several oligomer melting temperatures using umbrella sampling Virtual Move Monte Carlo simulations [171] (as outlined in Appendix B). While the histogram reweighting method

Sequence	$\boldsymbol{T}_{\mathbf{m}}$(US)	$\boldsymbol{T}_{\mathbf{m}}$(HR)	$\boldsymbol{T}_{\mathbf{m}}$(SL)	$\boldsymbol{T}_{\mathbf{m}}$(SL-avg)
AAGCGT	38.0	38.2	39.6	31.2
GAGATC	24.4	24.0	22.0	31.2
TCTCCATG	44.7	44.6	44.6	48.2
CCCGCCGC	71.1	70.6	71.1	48.2
ATTTATTA	21.2	21.3	23.9	48.2
ATATAGCTATAT	47.0	49.3	48.1	64.7
ATGCAGCTGCCG	74.0	74.3	72.6	64.7
GCGCAGCTGCCG	79.8	79.6	79.0	64.7

Table 2.2: Duplex melting temperatures (shown in °C) as predicted by our coarse-grained DNA model using umbrella sampling Monte Carlo simulations (T_m(US)) and histogram reweighting (T_m(HR)) compared to that for the SL model (T_m(SL)). T_m(SL-avg) is the melting temperature as predicted by the averaged SL model, which depends only on the length of the sequence. Sequences are specified in 3′-5′ direction.

estimates the melting temperature using the same 15 000 generated states for each duplex length considered and extrapolates from the average-base to the sequence-dependent potential, umbrella sampling simulations are run separately for each sequence considered. The umbrella sampling uses the sequence-dependent potential and is done close (within 3 °C) to the melting temperature of given sequence, hence providing a more accurate estimation of the melting temperatures in our model.

The comparison between the different methods is shown in Table 2.2 for a series of sequences. On average the histogram reweighting and the umbrella sampling agree to within 0.3 °C, which is very satisfactory. However, there is one significant outlier, ATATAGCTATAT, for which a difference of 2.3 °C was obtained. One reason for the difference may be that the melting temperature is about 16.6 °C lower than the melting temperature of an average strand of the same length from which the configurations were taken for the histogram reweighting. This difference is larger than the typical width of the melting transition (around 10 °C for sequences of length 12). Moreover, the sequence has a relatively high AT content and may adopt structures with significant fraying at the ends that contribute to the ensemble of configurations for the actual strand. However, such frayed states might have been poorly sampled when the

ensemble was generated using the average-base model. For these reasons, the sampled configurations may not provide a good representation of the true state-space of the system. Nevertheless, a number of other sequences tested here also have melting temperatures that differ significantly from the average sequence, without exhibiting such a large difference in the predicted melting temperatures between the two methods. Although it may be true that including a significantly larger set of states in the histogram reweighting method could reduce the errors in these outliers, we decided not to pursue this route further, given that the accuracy of the underlying SL model is not much different than our parametrization errors. Should a significantly more accurate model of the experimental data become available, however, then it may be that this point needs to be revisited.

2.4.2 Hairpin melting temperatures

We also tested our model's predictions for hairpin melting temperatures. This provides a distinct test of the parametrized model, since the sequence-dependent parameters were fitted to duplex melting temperatures only. Importantly, this test also probes the quality of the model's description of the single-stranded state, a feature often neglected in DNA models. We test melting temperatures of 4 different hairpin-forming sequences with different stem and loop lengths. We used strong and weak stem sequences to highlight sequence effects.

The simulations were performed with umbrella sampling using the number of correct base pairs in the stems as a reaction coordinate. The melting temperature T_{m} is defined as the temperature at which the system spends half of the time in the hairpin state, which is in turn defined as the ensemble of configurations with one or more correct base pairs. In Table 2.3, we compare our predictions for T_{m} with those obtained from the SL model. The average-base parametrization was previously found to consistently underestimate T_{m} for hairpins by approximately 3 °C, but to show the correct variation with loop and stem length [150]. The sequence-dependent

Sequence	T_{m}	T_{m}(SL)
AGCGTCACGC-$(T)_6$-GCGTGACGCT	86.5	86.7
AGTATCAATC-$(T)_6$-GATTGATACT	62.2	64.4
AGCGTC-$(T)_{10}$-GACGCT	64.5	67.0
AGTATC-$(T)_{10}$-GATACT	44.0	47.3

Table 2.3: Hairpin melting temperatures (shown in °C) as predicted by our coarse-grained DNA model (T_{m}) compared to the prediction by the SL model T_m(SL). Sequences are specified in 3′-5′ direction.

parametrization presented here also tends to underestimate T_{m} by roughly the same amount, but the sequence effects are well captured.

We further examine the effect of stacking on the melting temperature of hairpins with longer loops in Section 3.3, where we compare our model with the experimentally measured influence of sequence content of the loop on the hairpin melting temperature, an observation which is beyond the SL model.

2.5 Sequence-dependent parametrization summary

In this chapter, we have extended the oxDNA model of Ouldridge *et al.* [150] (which distinguishes between AT and CG base-pairing but otherwise treats these interactions at the average base level) to include sequence-dependent stacking and hydrogen-bonding interactions. To derive the new parameters, we developed a histogram reweighting procedure that allowed us to fit to thousands of melting temperatures of oligomers ranging in length from 6 to 18 base pairs. Melting temperatures were extracted from SantaLucia's nearest-neighbor model [77] which we treat here as a good fit to experiment.

Sequence can have an important effect on melting temperatures. For oligomers with the same length, but different sequences, melting temperatures can differ by as much as 50 °C. Even for the same sequence content, but different base-pair ordering, variations in stacking energies mean that melting temperatures can vary by up to 10 °C. Our new parametrization reproduces these differences and on average agrees

to within a standard deviation of 0.85 °C with the SL nearest-neighbor model. In contrast to the model's ability to capture thermodynamic properties, our coarse-grained model does not attempt to include the effects of sequence on structural or mechanical properties of double-stranded DNA [161, 173, 174, 160, 175, 163]. Instead, these remain as previously reported in [150, 140] for the average-base model.

Our new thermodynamic parametrization opens up the possibility of investigating sequence-dependent DNA phenomena, which we will consider in Chapter 3.

Finally, we note that the applicability of our new parametrization has some limitations. Firstly, it is only fit to a single salt concentration of $[Na^+] = 0.5$M, where the electrostatic properties are strongly screened. A new kind of parametrization may be necessary to reach significantly lower salt concentrations. Secondly, the model lacks certain detailed local structural information, such as major and minor grooving, or sequence-dependent elastic parameters [173, 163, 175, 176]. Furthermore, our model was fit to data that only includes the effects of base-pair steps. Additional experimental data on single-stranded stacking is needed to separate out the stacking strength of individual base combinations. Applications where the effects we neglect are crucial may therefore be best studied by other models.

Chapter 3

Sequence-dependent phenomena studied with a coarse-grained DNA model

To demonstrate some of the strengths and weaknesses of the sequence-dependent parametrization of the oxDNA model, introduced in Chapter 2, we present a series of studies of DNA systems for which sequence plays a non-trivial role. The results were obtained from either Virtual Move Monte Carlo or dynamical simulations of the model, both of which are described in Appendix B.

We explore the flexibility of the sequence-dependent oxDNA model by studying: the heterogeneous stacking transition of single strands in Section 3.1, the tendency of a duplex to fray at its melting point in Section 3.2, the effect of stacking strength in the loop on the melting temperature of hairpins in Section 3.3, the force-extension properties of single strands in Section 3.4, and the structure of a kissing-loop complex in Section 3.5.

3.1 Heterogeneous stacking transition of single strands

Single strands in oxDNA undergo a broad stacking transition as a function of temperature, i.e., a transition from a state with all or the majority of neighboring bases coplanarly aligned to a state with disrupted alignment [150, 140]. Such a transition

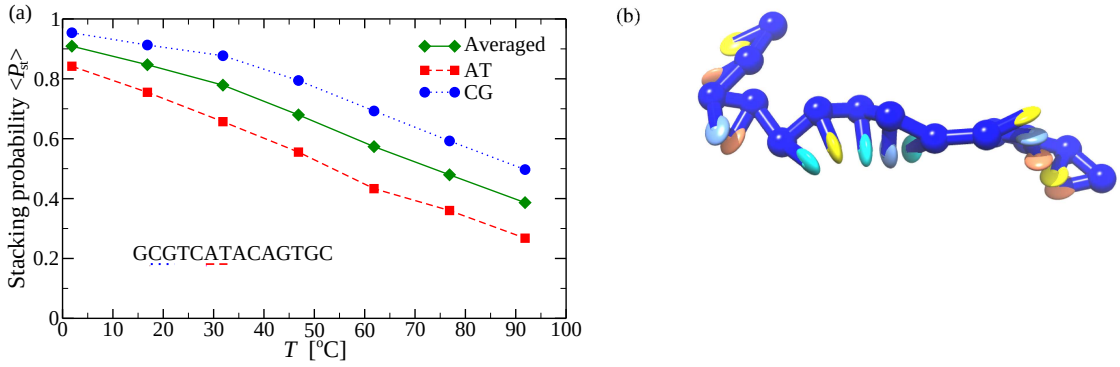

Figure 3.1: (a) The stacking probability, calculated as the fraction of time in the stacked state, varies with temperature and is heterogeneous along the sequence. Circles correspond to the strongest stacking term, CG (underscored with dotted line in sequence), while squares correspond to the weakest stacking step, AT (underscored with a dashed line in the sequence). Diamonds correspond to the average of all the stacking along the sequence. (b) A typical single stranded configuration at 45 °C. The first two bases on the left are unstacked. The strand has three stacked regions which adopt a helical geometry.

is also generally accepted to occur for DNA, although there is not a clear consensus in the literature about many aspects of this transition [9].

To investigate the sequence dependence of stacking in our model, we ran molecular dynamics simulations for a 14-base single strand with sequence 3′-GCGTCATACAGTGC-5′ (the same sequence as studied in [177]) at a range of temperatures. We measured the probability that a neighbor pair stacks. Two bases are considered to be stacked if the magnitude of their stacking interaction energy is at least 6% of its maximal value. The choice of a cutoff is one of convenience; we have checked that doubling it does not measurably change the results. Even though the different stacking strengths do not vary from the average by more than 7%, the effects on the stacking probabilities are still quite significant. For example, as shown in Fig. 3.1(a), the difference between the strongest (CG) and the weakest (AT) stacking pairs is large enough that the midpoints of the transitions are separated by about 40 °C.

The structure of the single strands is also heterogeneous, consisting of unstacked and stacked regions of various lengths, as illustrated in Fig. 3.1(b). The stacked

regions adopt a helical geometry, whereas the unstacked regions are more disordered.

The strands are also dynamically heterogeneous: over time the stacked and unstacked regions grow and shrink, while the average probability that a given neighboring pair of bases stack varies with temperature and position is measured in Fig. 3.1(a). Mechanical and structural properties of the single strands are therefore heterogeneous both in space and in time.

While we are confident that the existence of significant temporal and spatial heterogeneity in single strands is a robust qualitative prediction of our model, given the paucity of experimental and theoretical data on the detailed stacking interactions between individual bases, many questions about the nature and time scales of these heterogeneities remain open.

3.2 Hybridization free-energy profiles of duplexes

For the average-base parametrization of oxDNA, it has previously been seen that duplexes at their melting point typically have a terminal pair of bases that are unbound [140]. This behavior is called fraying, and it is generally thought that the ease of fraying is sequence-dependent with AT ends fraying more readily [178]. To explore the fraying behavior in our model, we study the free-energy profiles of the sequences ATATAGCTATAT, ATGCAGCTGCCG and GCGCAGCTGCCG (specified in $3'$-$5'$ order). Note that all three sequences have the same four central bases but different ends.

In Fig. 3.2 the free-energy profiles are shown as a function of the number of the native base pairs formed between the complementary strands. The free energies were set to be equal to 0 in the state with 0 native base pairs, i.e. when the duplex is melted.

Of most interest is how the most stable duplex state depends on sequence. For the strand with two GC ends, the free-energy minimum is a state with all 12 bonds

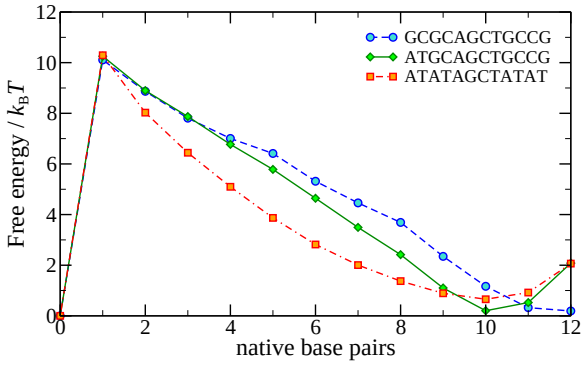

Figure 3.2: Free energy profiles for three different duplexes of length 12 as a function of the number of complementary (native) base pairs of the two strands. The simulations for each duplex were run at their respective melting temperatures, namely 48 °C, 73 °C and 80 °C.

formed, although the free-energy cost of opening up 1 base-pair is minimal. By contrast, for the case of either one or two AT ends, the duplex has the lowest free energy in a state with 10 bonds formed. Although the system pays an energetic cost for having 2 bonds unformed, it gains entropy from this opening up of the end base pairs. Thus, our model strands exhibit fraying, with the expected stronger tendency to fray for duplexes with weaker AT ends. Note that the sequence with two AT ends frays despite being at a significantly lower temperature than the GC rich sequence. Fraying has many consequences for DNA behavior. For instance, it exposes the end bases, allowing them to take part in reactions with other strands, which is important, for example, in a toehold-free displacement process [179].

Other features of note that are apparent from the free energy profiles in Fig. 3.2 are the nature of the first free energy jump and the shape of the minimum corresponding to the bound state. The fact that the first jump is almost the same for all three sequences reflects that it is dominated by the loss of center of mass entropy on association, which is the same (in units of $k_B T$) for the three systems. The shape of the free energy minimum corresponding to the duplex highlights differences in the ensemble of duplex states for different sequences. For the weakest sequence, at the melting point, the duplex can have as little as 7 base pairs for a significant fraction

of the time, and roughly with the same probability as for it being fully closed. The most GC rich sequence, on the other hand, shows little tendency to fray even at its melting point and it rarely breaks more than 3 base pairs.

3.3 Loop sequence effect on hairpin melting temperatures

In Section 2.4.2, we tested our model on melting temperatures of hairpins with short loops of lengths 6 and 10. In the SL model, the loop contribution to the free-energy difference for closing a hairpin is considered to be of purely entropic origin and sequence independent. However, it was observed experimentally [180] that hairpins with the same loop lengths but different sequences have different melting temperatures. In particular, the experiment in Ref. [180] considers sequences with the same stem sequence and loops consisting of either poly(A) or poly(T). The observed difference in melting temperature of the two different loop sequences was 4 °C for loop length 12 and increased to 12 °C for loop length 30, with the poly(A) loop always having lower melting temperature. It was proposed that the strand with a poly(A) loop region has a higher rigidity in the single-stranded case due to the base stacking and thus pays a larger penalty for closing.

Although the experiments in Ref. [180] were done at a salt concentration of 0.1 M, lower than the 0.5 M to which our model was fitted, it is instructive to see in general how stacking in the loop influences the stability of hairpins. We calculated the melting temperature for the sequences with the same stem sequence as in the experiment and a range of stacking strengths in the loop. Since our model does not distinguish between AA and TT stacking, we use an artificial base type X that is taken to stack as A with other bases and distinctly (with stacking strength η_{XX}) with other bases of the same type X.

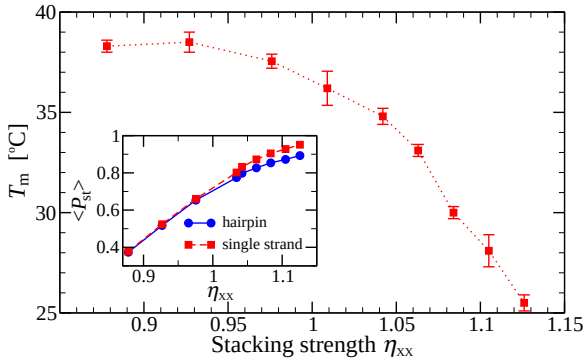

Figure 3.3: Hairpin melting temperatures as predicted by oxDNA as a function of stacking strength within the loop. We use a sequence 3′-GGGTT-$(X)_{25}$-AACCC-5′, where X is taken to stack as A with other bases, and with stacking strength η_{XX} with itself. The sequence is specified in 3′-5′ direction. The predicted melting temperature for the SL model is 37.8°C. The inset shows stacking probability $\langle P_{\mathrm{st}} \rangle$ within the loop region in the hairpin state (circles) and single-stranded case (squares) as a function of stacking strength η_{XX}.

The results, summarized in Fig. 3.3, show that for $\eta_{\mathrm{XX}} < 1$, the melting temperatures are fairly insensitive to stacking strength whereas for $\eta_{\mathrm{XX}} \gtrsim 1$, the melting temperature starts to drop significantly with increasing stacking strength. In the inset of Fig. 3.3 we show the average stacking probability in the loop, compared to that of the competing single-stranded state at the same temperature. In general, as the stacking strength increases, the probability that a piece of single-stranded DNA has long stacked regions also increases. The geometric constraints of the loop on stacking therefore become more pronounced with increasing strength, destabilizing the hairpin and leading to a drop in the melting temperature. On the other hand, for $\eta_{xx} \lesssim 1$, the stacked regions have an average length $\langle l \rangle \lesssim 3$, which is short enough that the hairpin geometry does not significantly affect the stacking.

If the data of Ref. [180] are to be interpreted using a model of stacking such as ours, we would infer that poly(A) has a very high stacking probability at these temperatures, while poly(T) has a significantly lower one. But, as the inset of Fig. 3.3 shows, we would not conclude that poly(T) is necessarily largely unstacked.

It is interesting to note that the stacking strength where destabilization becomes

noticeable coincides with the top end of our fitted strengths, and that if we were to separate poly(T) and poly(A) stacking strengths, it would not require an unreasonable change to give a signal of comparable size to that reported in Ref. [180]. In particular, if one sets η_{AA} to 1.105 and accordingly adjusts η_{TT} to 0.979 in order to keep the average of the two coefficients the same as for our base-pair step parametrization, the obtained difference in melting temperature of the hairpins with poly(A) and poly(T) loop is about 9 °C. For these values of η_{TT} and η_{AA} the standard deviation of melting predictions for the set of duplexes used in testing our parametrization increases by only 0.1 °C. Thus, if one wants to investigate a system where the difference in AA and TT stacking strengths plays an important role, these coefficients can be used. However, in the absence of a systematic study of the effects of loop sequence on hairpin melting temperature at high salt, we do not include differences between pairs that cannot be distinguished by the SL model in our parametrization in Table 2.1.

3.4 Force-extension curves of single strands

The mechanical properties of single strands have been experimentally measured for both DNA and RNA [181, 182, 183, 184, 185, 172, 165] to characterize their average as well as base-specific properties. In particular, qualitatively different behavior has been observed for single-stranded poly(T) (poly(U) in the case of RNA) compared to poly(A) (poly(C) or poly(G) in the case of RNA); the latter exhibit significant deviations from standard polymer models such as freely-jointed and wormlike chains, whereas the former do not. These deviations — concave regions with negative curvature in the force-extension curves — are described as "plateaus" [182, 183, 172].

To investigate the effects of sequence on the mechanical properties of single strands in our model, we simulate mechanical pulling and obtain force-extension curves for 50-base strands at room temperature (25 °C). We consider DNA single strands corresponding to our weakest and strongest stacking sequences, poly(GA) and poly(A),

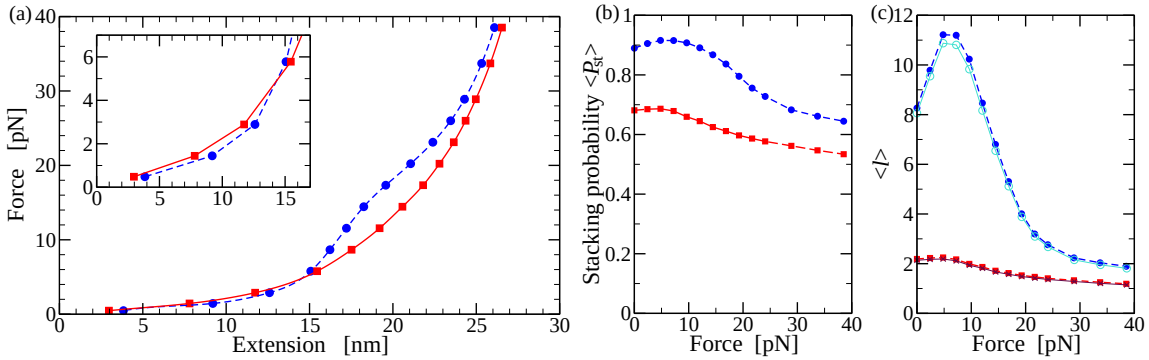

Figure 3.4: (a) Extension of 50-nucleotide single-stranded DNA at 25 °C as a function of applied force. In all panels, blue circles correspond to a poly(A) sequence (strongest stacking in our model), while red squares correspond to a poly(GA) sequence (weakest stacking). The inset in (a) shows a magnified section of the force-extension curve for low forces. (b) Stacking probability of a neighbor pair as a function of the applied force F. (c) Average length of a stacked domain $\langle l \rangle$ as a function of applied force F. The open circles and crosses show $\langle l \rangle_{\mathrm{uncoop}}$ as predicted by the uncooperative stacking model (Eq. 3.1) using $\langle P_{\mathrm{st}} \rangle$ as measured for poly(A) and poly(GA) respectively.

which differ in η_{ij} by about 7%. We note that in Sec. 3.3, we used hairpin melting to distinguish AA and TT stacking strength, but the obtained values are open to enough uncertainty that in this section we return to our original parametrization. Our focus here is on the qualitative effect of stacking differences, rather than their quantitative values.

Fig. 3.4(a) shows force-extension curves for our strongest and weakest stacking sequences. The concave section for strongly-stacked poly(A) between 15 and 25 pN is qualitatively similar to the plateau-like features observed in experiment [183, 182, 172]. The relatively weakly-stacked strand, poly(GA), follows a convex force-extension curve which is fairly typical of a classical homo-polymer model.

The poly(GA) curve is similar to the one found for the average-base model, which in turn is in reasonable quantitative agreement with experimental results for typical sequences. Although quantitative comparison with experimental data for non-homopolymeric sequences, such as λ-phage ssDNA [181], is hampered by the presence of metastable secondary structure [185, 186, 187], at tensions above about 15 pN,

where hairpins are disrupted, the extension per base at given force in the average model is within 10% agreement with Ref. [181]. A detailed discussion of the agreement between the average-base model and experiment is given in Ref. [140].

To understand the difference between the two single strands in our simulations, it is instructive to first recall that the strands consist of dynamically changing stacked and unstacked regions, as discussed in Section 3.1. When no force is applied, an unstacked region typically has a shorter end-to-end distance than a stacked region because it is more flexible and hence behaves more like a random coil. On the other hand, unstacked regions also have a greater maximum extension because the backbone is not restricted to a helical geometry as in the case of stacked regions.

To explore the effect of pulling on the structure of the single strands, we measured the stacking probability $\langle P_{\rm st} \rangle$ and the average length $\langle l \rangle$ of contiguously stacked sections for both strands, where a section of length l consists of $l+1$ bases. The results, as a function of applied force, are plotted in Figs. 3.4(b) and (c). When no force is applied, the stonger-stacking strand poly(A) has $\langle l \rangle \cong 8$ while the weaker-stacking strand poly(GA) consists mostly of short stacked regions with average length $\langle l \rangle \cong 2$.

As shown in the inset of Fig. 3.4(a), at low forces the stronger-stacking poly(A) strand is more extensible than the weaker stacking one, by as much as 20% at 1 pN force. The reason for this difference is that long stacked sections have a smaller entropic cost for aligning with the applied force than unstacked regions do. However, as the force increases further and the strands align more with the force, the curves cross (at ≈ 5 pN), and poly(A) becomes less extensible because of its shorter effective contour length.

Increasing the force also leads to significant changes in the average length of stacked regions in poly(A). Interestingly, at low force, the lower entropic cost for aligning of longer stacked strands leads to an initial increase in $\langle P_{\rm st} \rangle$ and $\langle l \rangle$ with force (up to around 5 pN). However, as the force increases further, both $\langle P_{\rm st} \rangle$ and $\langle l \rangle$

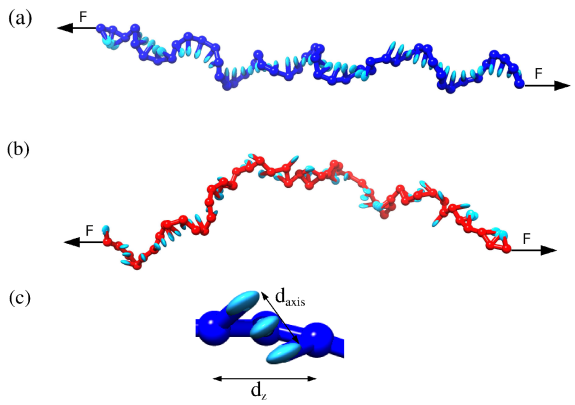

Figure 3.5: (a) Visualization of a 50-base long poly(A) ssDNA under a tension $F = 15\,\mathrm{pN}$, showing multiple stacked regions with helical geometry. The arrows indicate the applied force on the first and the last base. (b) Poly(GA) strand under a tension of 15 pN, consisting of short stacked regions as well as unstacked ones. (c) Magnified section of ssDNA illustrates that three stacked bases can align with the applied force without disrupting the stacking interaction. The contour length d_z, aligned with the force, is larger than the axial rise d_{axis}.

start to decrease because it becomes favorable for the strand to disrupt stacking to allow for greater extension. The reduction in stacking is particularly significant for the poly(A) strand over the range 15 to 25 pN, the location of the concave region in the force-extension curve. The long stacked regions are broken down into shorter ones which facilitates an increase in the overall length of the polymer. However, a short region of 3 bases can still align its backbone with the force while remaining stacked, as illustrated in Fig. 3.5(c). Therefore, even though it is progressively reduced with force for both poly(A) and poly(GA), a significant degree of stacking is preserved even at high forces.

The changes in stacking hence explain the physical cause of the concave "plateau" region in the force-extension curve for the stronger-stacking strand, poly(A). It corresponds to the structural transition as the increasing force disrupts the long stacked regions and $\langle l \rangle$ decreases. The concave segment of the force-extension curve is not present for poly(GA) because the latter already consists of mostly short stacked regions at zero force.

The differences in the structure of the poly(A) and poly(GA) strands described

above are further illustrated in Figs. 3.5(a) and (b), where snapshots of the sequences are shown for a force of 15 pN. The poly(A) strands are clearly much more stacked than the poly(GA) strands are, and also more strongly aligned with the force. From this picture one can also see why the derivative of the force-extension curve begins to rise steeply for the poly(A) curve around 15 pN: The highly stacked strand is nearing its maximum extension, whereas the unstacked strand is not.

It is interesting to note that a mere 20% difference in stacking probability between poly(A) and poly(GA) at zero force causes a significant difference in the average length of stacked regions: $\langle l \rangle \cong 8$ versus $\langle l \rangle \cong 2$. This effect can be understood by considering a simple, uncooperative model for stacking along the strand. Let p be the probability that two neighbors are stacked and $P(l)$ the probability that a stacked cluster has length l. Assuming an infinitely long polymer chain, the probability of having a continuously stacked region of length l is $P(l) = (1-p)\,p^l$, which is the probability of having l subsequent base pairs stacked (each with probability p) and the $(l+1)$-th base not stacked with the next base along the chain (which is with probability $1-p$). The average length $\langle l \rangle_{\text{uncoop}}$ of a stacked region in this uncooperative model can thus be obtained by summing over l:

$$\langle l \rangle_{\text{uncoop}} = \sum_{l=0}^{\infty} l P(l) = \frac{p}{1-p}\,. \tag{3.1}$$

Since our model has low stacking cooperativity [150], we can make the approximation $p \approx \langle P_{\text{st}} \rangle$. Fig. 3.4(c) shows that this simple model compares remarkably well with the measured values of $\langle l \rangle$. The fact that $\langle l \rangle$ diverges as $\langle P_{\text{st}} \rangle$ approaches 1 explains the sensitivity of the model strands to relatively small changes in stacking propensity at large $\langle P_{\text{st}} \rangle$ and also explains the large differences in $\langle l \rangle$ observed at zero force.

It is illuminating to compare our results to the theoretical model used by Seol *et al.* in Ref. [183] to explain the observed force-extension curves of RNA. Their model makes similar physical assumptions to the behavior of our coarse-grained model: the single strand is split into rigid helical regions and flexible random coil regions. Thus

the basic explanation for the plateau region is the same as in our model. However, there are also some differences. For example, our model suggests that absence of a plateau in the force extension curve does not necessarily mean the absence of stacking. In fact, we have observed that short stacked regions persist even while pulling the strand at a high force, because our model allows for three bases to remain stacked while aligning the backbone with the applied force, a feature that is not present in the model used in Ref. [183]. Moreover, the concave region in the force-extension curve interpreted with our model would indicate the presence of a much stronger stacking propensity than the one derived in Ref. [183]. Although our description of single strands is fairly simple, it incorporates the underlying physics of the model of Ref. [183] and in addition provides an explicit 3-dimensional representation of single-stranded nucleic acids. In summary, we believe that the presence of concave region in the force extension curve suggests that long stacked regions are present in the relaxed strand. This would either indicate strong uncooperative stacking, as in our model, or large cooperativity in stacking.

3.5 Structure of a kissing complex

A recent publication [188] used the average-base oxDNA model to investigate DNA kissing complexes, a system where topological and geometrical frustration have important effects. In this section, we show how the sequence dependence of interactions can introduce non-trivial changes to the structure of a kissing complex, with potential importance for the operation of nanotechnological systems [189, 190, 49, 50, 191, 54].

A kissing complex is a system in which two hairpins have loop regions that are complementary and can thus at least partially hybridize (see Fig. 3.6(a)). They are a common motif in RNA and are expected to form in DNA nanotechnology systems where complementary hairpins are used as fuel for DNA nanomachines [189, 192]. In the experimental system realized in Ref. [189], two strands of 40 nucleotides were

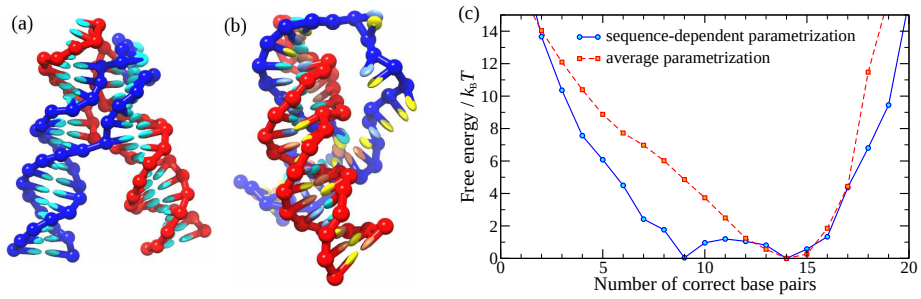

Figure 3.6: Effects of sequence dependence on the structure of kissing hairpins. (a) Typical structure found in both the average-base and sequence-dependent parametrization, with 14 intramolecular base pairs. (b) Second free-energy minimum found only in the sequence-dependent parametrization, with 9 intramolecular base pairs. Please note the exposed bases—not present in (a)—that can be used as a toehold by the catalyst strand to initiate displacement. (c) Free-energy profile for binding with the two parametrizations, with the sequence-dependent one exhibiting a second minimum corresponding to the structure depicted in (b).

designed to be both complementary and also able to form a hairpin with a stem of 10 base pairs. As the remaining 20-base loops are complementary to each other, the two hairpins can form a kissing complex. The sequences are:

3'-CGCAACGACG-GCTCCCCTCTTCTCATTTTA-CGTCGTTGCG-5'

and

3'-CGCAACGACG-TAAAATGAGAAGAGGGGAGC-CGTCGTTGCG-5'

where the hyphens separate stem and loop regions. A dilute solution of such strands tends to form hairpins much more quickly than full duplexes, due to a lower kinetic barrier for the former process. The hairpins in turn form kissing complexes, an intermediate metastable state with respect to full hybridization that requires a significant amount of rearrangement to transform into the full duplex. The kinetic barrier, due to the topological frustration of the complex, is so high that full hybridization is almost impossible. However, this barrier can be reliably resolved by the introduction of a DNA catalyst strand, designed to open one of the hairpins by displacement and trigger full hybridization, thus releasing the stored free energy [189].

Following Ref. [189], the work in [188] studied the structure of the resulting kiss-

ing complex with the average-base oxDNA parametrization and found that the system typically assumed a structure with two symmetric parallel helices, as shown in Fig. 3.6(a). However, as the loop sequences used in Ref. [189] are very asymmetric in GC content, we expect that the average-base model should overestimate the stability of the weakly bound region and conversely underestimate that of the strongly bound, GC-rich region.

When we repeated the structural study with the sequence-dependent potential, we obtained a qualitatively different result. Computing the binding free-energy profile of the system, using the number of native base pairs (i.e. base pairs that would be present in the final full duplex) as an order parameter (Fig. 3.6(c)), we found a second minimum at around nine interstrand base pairs that was not observed for the average-base model. A typical configuration associated with this minimum is shown in Fig. 3.6(b). It is evident that as well as being able to form the structure with two symmetric helices, the system is also able to adopt an alternative structure with a single intermolecular helix that both contains the GC-rich section and is slightly larger than either individual helix in the two-helix form.

This competing minimum has potentially important consequences for the nanotechnological applications of kissing hairpins. In Ref. [189], a catalyst strand was introduced to the system in order to facilitate full hybridization of the complex: the strand was designed to bind to the weaker half of one of the loops, and then to open up the hairpin by displacement. The fact that a competing minimum exists in which the whole weaker half of the loop is available for binding will favor this process, as it provides a long, easily accessible toehold for displacement. Such toeholds are known [179] to accelerate displacement reactions by several orders of magnitude. Our model suggests that if the strand was instead designed to bind to the stronger half of the loop, its effectiveness would be hindered rather than helped by the presence of the alternative minimum. We would therefore expect such a catalyst to be less effective

than the one used in Ref. [189].

The qualitative difference between the results of the two parametrizations in this case highlights that if one is interested in the detailed properties of a system like this one, where short binding regions with asymmetric GC content are present, it is important to have a model with sequence-dependent binding strengths to be able to make more accurate predictions. Were the GC pairs in the loop more evenly distributed, we would expect the results of the average-base model free-energy profile to accurately describe the kissing complex.

3.6 Summary

In this chapter, we applied the DNA model with sequence-dependent parametrization of stacking and hydrogen-bonding interactions to a variety of phenomena where sequence plays an important role. In particular, we studied the following systems:

(a) Heterogeneous stacking transition in single-stranded DNA: Even though our stacking parameters do not vary by more than 7%, they can induce significant spatial and temporal heterogeneity in the stacking of single strands. For example the difference in stacking probability between the strongest and the weakest stacking pairs in the oligomer we studied is large enough that the midpoints of the stacking transition of two separate pairs in a single strand can be separated by as much as 40 °C. These results suggest that structural and mechanical properties of single-stranded DNA should be highly heterogeneous as well.

(b) The hybridization free-energy profiles of duplexes: We studied three different 12-mer sequences at their respective melting temperatures, finding that sequence heterogeneity also has significant effects on the probability that the ends of a duplex are open, i.e. that they fray. We found that AT ends are typically frayed, while sequences with GC ends exhibit a free-energy minimum for a completely closed duplex.

(c) The effect of stacking strength in the loop on hairpin stability: The SL model only distinguishes base-pair steps. Given that we used this model to generate the melting temperatures to which we fit, we were unable to uniquely isolate the stacking strength of individual base combinations. Additional experimental data on single-stranded stacking is needed to separate these interactions. One potential source of data that goes beyond the SL model is given by experiments on melting of hairpins with poly(A) and poly(T) loops [180]. By calculating how increasing the stacking strength in the loop lowers the melting temperatures, we showed that parameters could be derived that reproduce the expected stronger AA compared to TT stacking, without significantly changing the quality of our fit to the overall melting temperatures of duplexes. Nevertheless, we do not yet include this difference in our sequence-dependent parametrization, because to be consistent we would need similar data to distinguish between other base-pair steps.

(d) The force-extension properties of single strands: Another experimental situation where differences in single-stranded stacking have been measured experimentally is in the force extension of ssDNA. We show that more strongly stacked sequences should be more extensible for small forces up to about 5 pN. For certain sequences, experiments have observed a concave "plateau" region in the force-extension curves. We are able to qualitatively reproduce this feature and, in agreement with previous explanations [183], attribute the plateau region to the different force response of stiffer stacked and more flexible unstacked regions. Furthermore, we show that the onset of the plateau region is correlated with a sharp decrease in the average length of stacked regions with increasing force. Because the average length of stacked regions drops rapidly with a relatively small decrease in the average stacking, we argue that a very large propensity to stack ($> 90\%$) is necessary to give a similar results to those observed in experiment. We therefore conclude that if these phenomena are to be explained through largely uncooperative stacking of bases to form helical ssDNA, as

in our model, a high stacking propensity is required. Furthermore, failure to observe a force plateau for a sequence does not imply an absence of stacking.

(e) The structure of a kissing-loop complex: Finally, we applied the sequence-dependent oxDNA model to study the effect of sequence on the structure of a kissing complex formed by two hairpins. When the sequences used in the experiments of Ref. [189] are studied, the average-base model exhibits one minimum free-energy structure [188], while the sequence-dependent model also generates a second, qualitatively distinct, stable structure. The new structure completely exposes a toehold which may significantly accelerate the DNA catalyst mediated release of free energy stored in the kissing complex.

The examples described above suggest that the oxDNA model with sequence-dependent parametrization can be used for many other DNA applications in nanotechnology and biology where sequence plays a significant role. Our model should work particularly well for situations where single-to-double stranded transitions are important.

Our work also highlights the need for new systematic experiments, in particular to elucidate the basic physics of single-stranded stacking interactions.

Chapter 4

Simulating a burnt-bridges DNA motor

In this chapter we use the average-base oxDNA model to study the operation of the "burnt-bridges" DNA motor created by the Turberfield group [47, 53, 58]. We use the average-base version of the oxDNA model [150, 140] rather than the sequence-dependent parametrization introduced in Chapter 2 because the generic features of the system are easier to resolve without sequence-dependent complications. For a direct comparison with experimental results, the sequence-dependent version of the model can be used to compare experiments with different sequences.

We first give a brief overview of the design of the DNA burnt-bridges motor and then present results of simulations of one step of this nanodevice. The motor consists of a single DNA strand (cargo) which moves from one complementary strand (stator) to the next through toehold-mediated strand displacement. We study the process for different distances between stators, different strengths of the attachments of the stators to a surface and for different lengths of the toehold, particularly focusing on the consequences of inter-stator distance.

4.1 Burnt-bridges DNA motor

The burnt-bridges DNA motor studied in this work is a system that produces autonomous, unidirectional motion of a single-stranded cargo along a track of single-

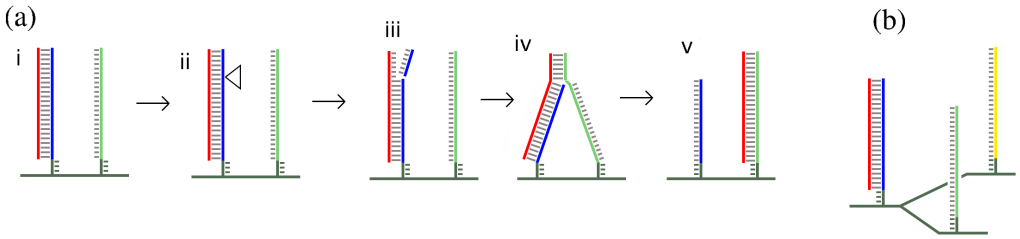

Figure 4.1: (a) A schematic illustration of the stepping process of a burnt-bridges motor: (i) The cargo (red) is attached to the first stator (blue), which is a complementary DNA strand. (ii) The nicking enzyme binds to its recognition sequence, and catalyzes the hydrolysis of the backbone. (iii) The nicking enzyme and stator fragment have dissociated. (iv) The exposed toehold of the cargo binds to the second stator (green). (v) The second stator fully displaces the first, and the motor completes a step. (b) Decision making at a junction. The cargo could step to either the yellow or green stator. By altering or blocking the toeholds, one direction or the other can be made preferable.

stranded stators. The motor was experimentally realized Refs. [47, 53, 58], and the stepping process of the motor is schematically illustrated in Fig. 4.1(a). The motor consists of a single DNA strand (called the cargo strand), which moves along a track consisting of strands (referred to as stators) that can be attached to a DNA duplex [47], or a DNA origami surface [53, 58].

Initially, the cargo strand is attached to the first stator. A nicking enzyme (N.B bvC 1b) is present in the solution [193]. These enzymes can bind specifically to a certain sequence of double stranded DNA present in the stator/cargo duplex, and cut the backbone of the stator strand (but not the connected cargo strand) a short distance from the $5'$ end of the stator (6 bases in Refs. [53, 58], and 8 in Ref. [47]). The binding of the shorter stator fragment is unstable under experimental conditions, and it tends to detach, revealing a short toehold on the cargo. The next available stator is positioned close to the first stator (around 7 nm in Ref. [47] and 6 nm in Refs. [53, 58]) and the exposed toehold can bind to the next stator. Strand displacement can then occur, allowing the cargo to replace bonds to the first stator with bonds to the second stator.

Once the displacement process is complete, the cargo is totally detached from

the first stator and fully bound to the second one. The stepping process can now be repeated, with the stator to which the cargo is attached being cut again by the enzyme and the cargo making a step to the next stator in line. The backward step is now highly improbable, as the preceding stator has been nicked and has fewer complementary bases with the cargo strand. Used stators therefore get disabled as the cargo travels along the track, leading to the description "burnt-bridges". Directional motion is possible because the walker's motion catalyzes the hydrolysis of the stator's backbone, a free-energetically favorable process.

In the original experiment [47], three stators were attached to a double-stranded DNA track. Fluorescence was used to demonstrate that the cargo stepped along the track, visiting the stators in order. Stators were later attached to a DNA origami surface [53], and the cargo was observed to move along a 17-stator track by atomic force microscopy. Motion along an 8-site track was also observed via fluorescence, and was strongly suppressed by the removal of a stator. Recently, motors have been designed that can choose a pathway at a junction based on information either carried by the motor itself or provided externally [58].

The suggested future applications of DNA walkers such as the burnt-bridges motor include programmable chemical synthesis and a molecular realization of a Turing machine [194].

4.2 Simulating the burnt-bridges motor

4.2.1 System

Our simulated system consists of three DNA strands: the first stator, the cargo strand and the second stator. The representation of these strands by the model is illustrated in Fig. 4.2. We simulate the process by which the motor steps to the next stator, stages (iii)–(v) in Fig. 4.1(a). The first stator has six bases fewer than the second stator, emulating the state shown in Fig. 4.1 (a)-(iii) after the small fragment of the stator

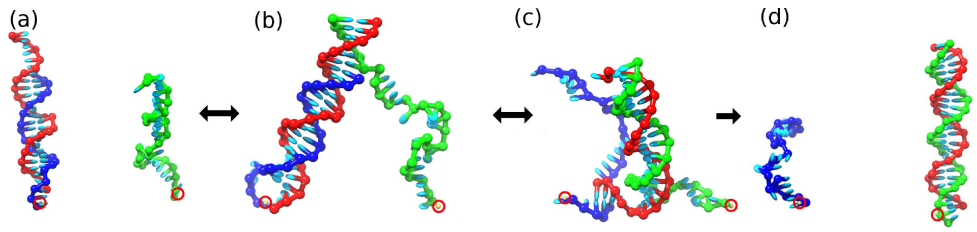

Figure 4.2: Typical configurations that are sampled in simulations of the stepping process, with the distance between the stators 7.1 nm and attachment spring constant 131 pN nm^{-1}. The yellow spheres inside red circles indicate the position of the points to which the 3′ ends of the stators are attached by spring potentials. Double-headed arrows indicate transitions which are expected to be reversible under the system conditions. (a) The cargo (red) strand is attached to the first stator (blue), with a 6-base toehold exposed after the nicking of the first stator. (b) The cargo's toehold makes 6 bonds with the second stator (green). (c) Displacement is nearly complete – the cargo has two bonds with the first stator and nineteen with the second stator. (d) Displacement is complete: the cargo is fully bound to the second stator.

has dissociated after nicking. We model the attachment of the stators to a surface via a spring potential, as modelling explicitly the whole DNA origami substrate to which the stators are attached in the experiment would be too computationally demanding. The spring potentials in our simulations act on the first 3′ base of the stators:

$$V_{\text{spring}}(k, \mathbf{r}_1, \mathbf{r}_2) = \frac{k}{2} \left\|\mathbf{r}_1 - \mathbf{r}_1^0\right\|^2 + \frac{k}{2} \left\|\mathbf{r}_2 - \mathbf{r}_2^0\right\|^2, \tag{4.1}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the positions of the center of mass of the first 3′ nucleotide of the first and second stator respectively and $\mathbf{r}_1^0$ and $\mathbf{r}_2^0$ are the positions to which they are attached by the springs. This potential energy is included in the total potential energy of the system. We define a set of coordinates such that the stator attachment points lie in the $z = 0$ plane, and are separated by a distance d in the x direction. We will compare three different values of d: 3.3 nm, 7.1 nm and 9.4 nm. 7.1 nm approximately corresponds to the distance used in Ref. [47] and we chose 3.3 nm and 9.4 nm to test shorter and longer distances respectively.

In most simulations, we use the spring constant $k = 131\,\text{pN}\,\text{nm}^{-1}$, chosen so that the variance of the distance between the attached nucleotides of the first and second stators, $\langle \|\mathbf{r}_1 - \mathbf{r}_2\|^2 \rangle - \langle \|\mathbf{r}_1 - \mathbf{r}_2\| \rangle^2$, is approximately equal to the variance

Strand	Sequence
First stator	TCAGCCCAACTAACATTTTA
Second stator	GGAACCTCAGCCCAACTAACATTTTA
Cargo	CGATGTTAGTTGGGCTGAGGTTCC

Table 4.1: The sequences of stator and cargo strands used in the simulations. The sequences are given in 5′-3′ order. The complementary segments of cargo and stator strands are shown in red for the toehold region and in blue for the region that is displaced by the second stator during the stepping process.

of the distance between two nucleotides that are 11 base pairs away on a strand in DNA duplex as simulated with our model at temperature 37 °C. Choosing the spring constant in this manner means that its magnitude is physically sensible for a DNA-based system. We will, however, consider the consequences of varying it.

To further mimic the presence of the DNA origami substrate, we forbid the cargo and stator strands from crossing the $z = 0$ plane. To achieve this we introduce an additional potential

$$V^i_{\text{repulsion}}(\mathbf{r}_i) = \begin{cases} \frac{k_r}{2} z_i^2, & \text{if } z_i < 0 \\ 0 & \text{if } z_i \geq 0 \end{cases} \tag{4.2}$$

which acts on the positions of center of mass $\mathbf{r}_i = (x_i, y_i, z_i)$ of all nucleotides i in the simulation. We set k_r to $1142\,\text{pN}\,\text{nm}^{-1}$, which is sufficiently large to effectively prevent nucleotides from crossing the $z = 0$ plane. The total potential energy of the simulated system is then

$$V(k, d) = V_{\text{oxDNA}} + V_{\text{spring}}(k, d) + \sum_{i=1}^{N} V^i_{\text{repulsion}}. \tag{4.3}$$

All our simulations are done at temperature 37 °C, the temperature used in experiment. The sequences used in our simulations are shown in Table 4.1. They correspond to the sequences used in Ref. [53]. In the experiments, there is an additional 20-base segment at the 5′ end of the stator that was used to bind a blocking strand that was displaced before the beginning of the stepping measurements. We do not include this segment in our simulation. The last 4 bases at the 3′ end of the stators are not complementary to the cargo strand, acting as a flexible linker with the surface. Note that

the two bases at the $5'$ end of the cargo strand are not complementary to the stator strands: in the experiment [53] this dangling end was used to attach a fluorophore to the system for tracking.

4.2.2 Simulation setup

We use the VMMC algorithm combined with umbrella sampling, which is outlined in more detail in Appendix B.2. In our simulations in this chapter, we compute the equilibrium free energy of the system as a function of the number of base pairs between the cargo and the first stator (b_{1c}) and the cargo and the second stator (b_{2c}), which are the order parameters for the umbrella sampling simulations.

It is convenient in our case to split the umbrella sampling protocol into two windows. In the first window we study attachment of the cargo's toehold to the second stator, and in the second we consider the displacement process. To do this, we restrict the system to $b_{1c} \geq 14$ and $b_{2c} \leq 8$ in the first case and $14 \leq b_{1c} + b_{2c} \leq 23$, $b_{1c} \geq 1$ and $b_{2c} \geq 5$ in the second. The windows are then combined using the overlap between the two with the weighted histogram analysis method [195].

We note that we do not sample all possible values of b_{1c} and b_{2c} using these two windows. To do so would have been computationally costly without being likely to provide much insight. In particular, we do not consider the breaking of many base pairs of the first stator/cargo duplex unless the cargo is attached to the second stator. The high free-energy cost of spontaneous breaking of base pairs at this temperature make substantial melting extremely unlikely. For the same reason, we do not consider displacement intermediates with fewer than 14 base pairs involving the cargo. Finally, we do not sample the transition to states with $b_{1c} = 0$, when the cargo is detached from the first stator. To do so would be especially difficult as it would require simulating the reattachment of the cargo to the first stator in the absence of a toehold. Once detached from the first stator, however, the system is in a configuration very similar to the initial one, the only difference being the extra six base pairs available with

the second stator. The free-energy change associated with adding a base pair to an isolated duplex in our model is known to be around $2.3\,k_{\mathrm{B}}T$ at 37°C (this can be seen from the slope of the free-energy profile in Fig. 4.3 (b) for the initial stages of toehold binding), and so we can roughly estimate the free energies of the $b_{1c} = 0$ states from those of the $b_{2c} = 0$ states. Most importantly, however, only states with both $b_{1c} > 0$ and $b_{2c} > 0$ are expected to be sensitive to the separation of the stators, as only in these states is the cargo stretched between both.

The umbrella sampling simulations of the DNA motor system reported in this chapter involve at least 3×10^{11} attempted VMMC moves. We note that 3×10^{11} VMMC moves with our simulation code corresponds to approximately 200 days of computer time on a single 2.3 GHz CPU. However, multiple simulations of the same system can be run independently on multiple cores and the sampling data from all of them can then be combined.

We note that the obtained free-energy landscapes do not completely determine the kinetic properties of our system, but they are indicative of how the system responds to certain changes of parameters. For example, the rate at which a process occurs is often limited by the need to pass through a high free-energy (improbable) state. Raising (or lowering) this barrier by perturbing the system generally results in an exponential decrease (or increase) in the rate of the process. In order to explicitly extract the rates of the stepping process, molecular dynamics simulations would have to be carried out.

4.3 Simulation results

We first present the free-energy landscape, and a profile along a one-dimensional pathway, for two stators 7.1 nm apart attached by springs with stiffness $k = 131\,\mathrm{pN/nm}$, to illustrate the basic features of the attachment of the cargo to the second stator and of the branch-migration process. We then compare the free-energy profiles for

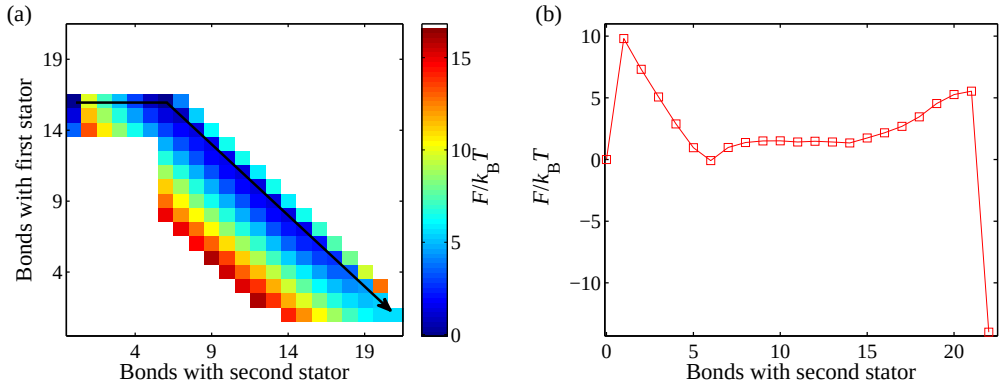

Figure 4.3: (a) The free-energy landscape of motor stepping as a function of bonds between the cargo and the first and second stator. The arrow indicates a pathway through the landscape that is used to plot (b). (b) The free-energy profile of displacement, plotted along the one-dimensional path shown in (a). A stage on this path is uniquely specified by the number of bonds with the second stator. The final point (23 bonds with the second stator) is estimated as discussed in the text, not measured.

different attachment spring constants, different distances between the stators, and different lengths of the toehold on the first stator.

4.3.1 Stators separated by 7.1 nm

The free-energy landscape as a function of b_{1c} and b_{2c} is shown in Fig. 4.3 (a), with major features highlighted by inspecting the one-dimensional pathway shown in Fig. 4.3 (b). The free energy is normalized to be equal to 0 for the case when the cargo has no bonds with the second stator. The basic features of the landscape are the following.

- There is a rise in free energy associated with the formation of the initial base pair with the second stator, due to the loss of configurational entropy when the first contact is formed.

- The free energy decreases as successive base pairs are formed in the toehold, due to the cooperative nature of duplex formation (once the first contact is formed, successive base pairs are much more likely).

- As displacement begins (once the seventh base pair is formed with the second stator), there is an initial rise in free energy, followed by a plateau. This initial rise is a generic feature of displacement resulting from steric interference at the displacement interface, and is investigated in detail in Ref. [196].

- Later stages of displacement, after around 15 base pairs have formed between the cargo and the second stator, involve an increase in free energy of around $4\,k_{\mathrm{B}}T$.

An unusual feature of the free-energy profile shown in Fig. 4.3 (b) is the increase in free energy towards the end of displacement. We attribute this to an increase in tension within the system. When the first bond between the cargo and the second stator is formed, the contact point is far away from the nucleotides that are attached to the surface (the 3′ end of the stators). It is therefore not difficult for the strands to reach each other at the contact point. By contrast, when more bonds have formed between the cargo and the second stator, the contact point is closer to the 3′ end of the stators. Eventually, the length of DNA between the contact point and the surface attachments gets so short that maintaining the structure causes considerable tension, which is free-energetically unfavorable and results in the observed rise in the profile.

The role of the attachment of stators to a surface can be tested by changing the spring constant of the attachment to the surface. The results of otherwise identical simulations with different spring constants are shown in Fig. 4.4. Increasing or decreasing the spring constant from $k = 131\,\mathrm{pN/nm}$ by a factor of about 4 has a very small effect on the free-energy profile. The reason is that in this range the attachment spring is fairly stiff, and the less costly way for the system to come close together is to stretch the single-stranded sections. By decreasing the spring constant by nearly two orders of magnitude, down to $k = 3\,\mathrm{pN/nm}$, we are able to access a regime where the attachment springs are sufficiently weak that the strands can relax the tension effectively by moving the bases at the 3′ end of the stators closer together rather than

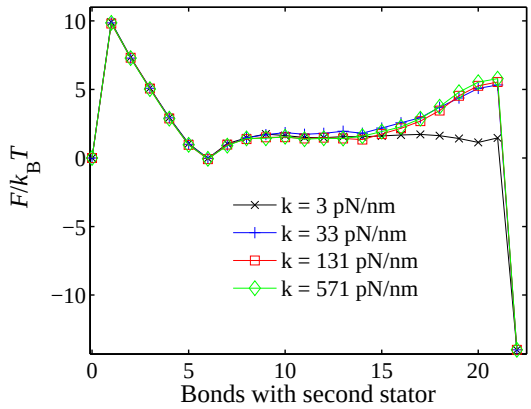

Figure 4.4: The free-energy profile of motor stepping for various strengths of attachment to the substrate, all using a stator separation of 7.1 nm. The profile is taken along the path illustrated in Fig 4.3 (a).

stretching the single-stranded sections. We stress that the physically relevant regime for stators attached to a DNA duplex or to a DNA origami is the high-k one.

4.3.2 Varying the distance between stators

We now compare the stepping of the motor for three different distances between the stators: $d = 3.3\,\text{nm}$, $7.1\,\text{nm}$, and $9.4\,\text{nm}$. The free-energy profile along the one-dimensional pathway indicated in Fig. 4.3 (a) is shown in Fig. 4.5 (b) for all three distances. Fig. 4.5 (a) shows the full two-dimensional free-energy landscape for the case of $d = 9.4\,\text{nm}$. All the profiles were produced with the same spring constant, namely $k = 131\,\text{pN}\,\text{nm}^{-1}$. The free energies have been normalized to zero when there are no bonds between the cargo and the second stator. One of the typical configurations sampled by our simulations for distance 9.4 nm between stators is illustrated in Fig. 4.6(c).

The initial part of the free-energy profile, from 0 bonds to 6 bonds with the second stator, is nearly identical for all three distances considered. As the second stator makes more bonds with the cargo strand, we see that the increase in free energy is bigger for larger distances d between the stators. This effect is consistent with our understanding that the rise in free energy is associated with increasing tension within

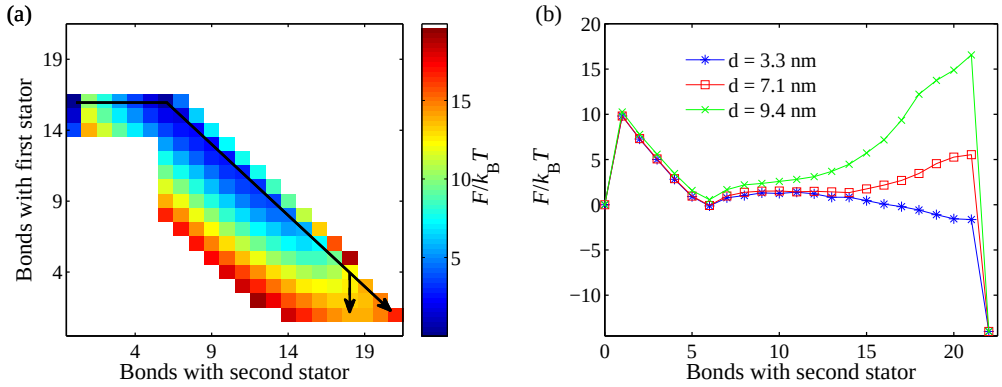

Figure 4.5: (a) Two-dimensional free-energy landscape of stepping for $d = 9.4\,\text{nm}$. (b) The free-energy profiles of motor stepping along the pathway illustrated in Fig 4.3 (a) for three different distances d between the attachment points of the stators. The illustrated alternative pathway in (a) shows that the profile in (b) for $d = 9.4\,\text{nm}$ probably overstates the difficulty of displacement, although it is still far more difficult than for $d = 7.1\,\text{nm}$.

the complex due to the need to stretch DNA between the surface attachment points and the junction. The snapshot in Fig. 4.6(c) clearly illustrates the tension in the system at later stages of displacement for $d = 9.4\,\text{nm}$.

In fact, inspection of Fig. 4.5 (a) shows that the tension for the 9.4 nm case is so great that when the cargo has only one bond with the first stator, it is thermodynamically more favorable for the cargo to be bound by only 18 or 19 base pairs to the second stator, rather than by 21 base pairs as for the other values of d. It is therefore highly probable that a typical displacement pathway would involve the first stator detaching when the second stator has significantly fewer than 21 base pairs with the cargo, along an alternative pathway such as that shown in Fig. 4.5 (a). As such, the profile along the pathway shown in Fig. 4.5 (b) tends to overstate the difficulty in stepping between the two stators for $d = 9.4\,\text{nm}$, although it is still a much more difficult process than for the shorter values of d.

Interestingly, for the smallest separation of $d = 3.3\,\text{nm}$ there is actually a decrease in free energy with increased binding of the cargo to the second stator. We attribute this to the fact that the number of bases under constraint actually decreases

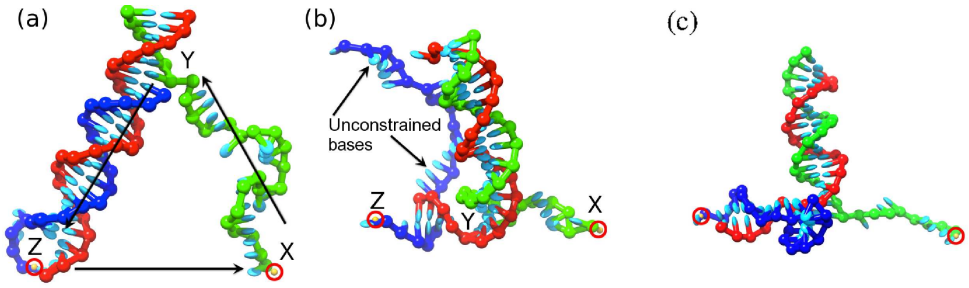

Figure 4.6: (a,b) Illustration of the reduction in constrained bases as simulation progresses. The cargo is colored in red, first stator in blue and second stator in green. (a) Toehold-binding only: the DNA along the path $X \to Y \to Z$ is constrained by the need to for the system to from a closed loop $X \to Y \to Z \to X$ (with the vector between attachment sites as part of the loop). (b) Later on during displacement, fewer bases are constrained within the loop. Those that remain, however, are under tension due to the need to stretch between attachment sites. (c) One of the typical configurations sampled in the simulation for distance between stators $d = 9.4\,\mathrm{nm}$. The cargo (colored in red) has three bonds with the first stator (colored in blue) and eighteen bonds with the second stator. The tension acting on the DNA between the attachment points and the displacement interface can be clearly seen from the stretched arrangement of the second stator (colored in green) close to its attachment point.

as displacement proceeds. As can be seen from Fig. 4.6 (a), when the second stator is bound only to the toehold of the cargo, most of the bases in the system are effectively held within a closed loop ($X \to Y \to Z \to X$ in the diagram – this loop includes the vector between attachment sites) that reduces their conformational freedom. As displacement proceeds, a single-stranded section of the first stator that is not constrained by looping is generated, and the number of bases in the loop is reduced, as depicted in Fig. 4.6 (b). All else being equal, transferring bases from within the constrained loop to an unconstrained section should be a favorable process because of the increase in a configurational entropy. However, the extra tension felt by the short loop overwhelms this effect for large stator separations.

4.3.3 Different toehold lengths

Finally, we compare the free-energy profile obtained at $d = 7.1\,\mathrm{nm}$ and $k = 131\,\mathrm{pN\,nm^{-1}}$ with that for an identical system, except with a longer toehold (9 bases rather than

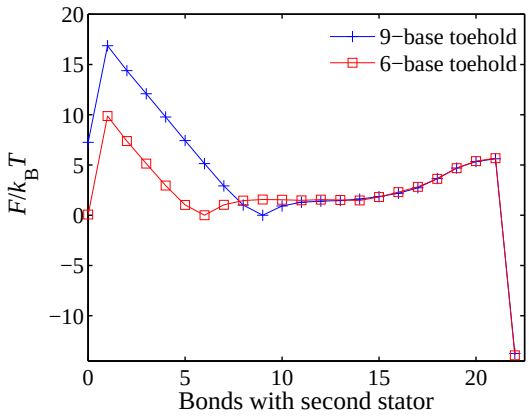

Figure 4.7: The free-energy profile of motor stepping for two different values of cargo toehold length. In both cases, the stators are at the same distance of 7.1 nm. For the 6-base toehold, the profile is obtained along the path shown in Fig. 4.3 (a). For the 9-base toehold, an analogous pathway in which the cargo first binds to the second stator by 9 bases and then the second stator displaces the first base-by-base is used. Once again, the final state is estimated as discussed in the text.

6) exposed by the nicking of the first stator. The free-energy profiles (taken along the same pathway as indicated in Fig. 4.3 (a) for the 6-base toehold and along an analogous path for the 9-base case) are shown in Fig. 4.7. The profiles have been normalized so that the free energy is equal to 0 when the cargo strand is bonded fully by its toehold to the second stator, i.e. by 6 and 9 bases respectively. In contrast with setting the free energy to 0 when the cargo is detached, as used in previous figures, this choice of normalization makes it easier to compare the profiles for the case of two different toehold lengths. As expected, in the case of the 9-base toehold, we observe a larger barrier for detachment of the cargo strand from the second stator once it is bound by the toehold, because is needs to break 9 base pairs to detach, as opposed to only 6 base pairs in the case of the shorter toehold. By contrast, once displacement is well underway the free-energy profiles do not differ by more than $0.3\,k_BT$, which is comparable with the estimated errors. Thus, once displacement has been initiated, the free-energy changes are only dependent on the distance between the strands and not the toehold length.

4.3.4 Consequences of free-energy profiles for motor operation and track design

Our data suggest that the tension generated within the motor as displacement proceeds has a potentially significant effect on motor operation. In particular, the resultant rise in free energy will suppress the speed with which the second stator can fully displace the first after binding to the toehold of the cargo. If this suppression is strong enough, the probability that the second stator detaches from the toehold rather than completing the displacement increases. The slope of the free-energy profile changes rapidly with d, being fairly gentle at $d = 7.1\,\mathrm{nm}$ and very significant at $d = 9.4\,\mathrm{nm}$. We would therefore expect its effects on stepping success to become noticeable for $d \sim 8$ or $9\,\mathrm{nm}$ for a 6-base toehold. By contrast, we note that the initial toehold contacts have almost equal free energies for the three distances studied. So, even though the distances vary, this result suggests that initial binding rates are relatively similar in all cases.

If stepping to the next stator after the current one has been cut is the limiting stage in motor operation, then the reduction of this rate would be manifested in the overall stepping speed of the motor. However, the binding, action and unbinding of the nicking enzyme all contribute to the time required to take a step, so (at least at low enzyme concentrations and moderate values of d) this change may have a negligible effect on the overall stepping speed. At sufficiently large d, however, increasing d further should have a noticeable effect on overall motor speed. This limit was clearly reached in Ref. [53], when removal of a single stator (estimated gap $\sim 12\,\mathrm{nm}$) dramatically reduced the overall rate at which the cargo reached the end of the track. Our results suggest this reduction was primarily due to the increased difficulty in completing displacement, rather than a reduced rate in making contact between the second stator and the toehold of the cargo.

Interestingly, the potential reduction in the success probability of displacement

may be advantageous at decision-making junctions such as those illustrated in Fig. 4.1 (b). We might encourage the cargo to choose stator A by relatively destabilizing the toehold of B in some manner, such as blocking it with another strand [58] or by having a different sequence. Leak currents will always exist, however, and even blunt-ended strand displacement (with no toehold) can occur [179].

The selection ratio of stator A to stator B, $\phi_{A/B}$, can be modelled by assuming that the toehold of the cargo can initially bind to stator i with a rate γ_i, whereupon the cargo successfully completes the step to stator i with probability f_i and detaches with probability $1 - f_i$. In this case,

$$\phi_{A/B} = \frac{\gamma_A f_A}{\gamma_B f_B}. \tag{4.4}$$

Assuming A and B are equidistant from the original stator, there is limited room to adjust γ_A/γ_B. The fact that $f_i \leq 1$ is important. It means that it will be essentially impossible to chose between two stators if both have sufficiently stable toeholds so that f_A and $f_B \sim 1$ (even if there is a large difference in absolute stability between the two). Furthermore, it means $\phi_{A/B}$ cannot be increased arbitrarily by increasing the stability of toehold A, limiting the maximum signal-to-noise ratio.

Increasing the failure rate of displacement by adjusting d (for both stators A and B equally) will tend to reduce f_i, and hence larger toehold stability will be required to achieve $f_i \sim 1$. The junction will then be able to distinguish between toeholds that are more stable than previously, and will have a higher maximum signal-to-noise ratio. We note that if the primary consequence of increasing d was to reduce the binding rate of the toehold to the next stator (rather than the success probability once bound), it would be impossible to improve the efficiency of the junction in this way.

Another possibility to change f_A and f_B is by the choice of the sequence, either in the toehold region or in the part of the stator strand that displaces the first stator to which the cargo was originally attached. By choosing the base composition of the

toehold, i.e. by making the sequence AT-rich or GC-rich, one can increase or decrease the possibility of unbinding. Another possibility is introducing mismatches in the stator strands.

4.4 Summary

In this chapter, we used the oxDNA model with the average-base parametrization to investigate an active DNA nanotechnology device: a unidirectional molecular motor [47, 53, 58]. In particular, we have studied the physics of the stepping process from one stator to the next, once the first has been cut by a nicking enzyme, with emphasis on the dependence of this process on the separation of stators.

We observed that the free-energy profiles of initial binding of the cargo's toehold to the second stator are fairly insensitive to stator separation. However, as displacement proceeds, there is a rise in the free energy when stators are separated by larger distances, associated with the need for ever shorter sections of DNA to extend across the gap between attachment points. Such a rise will tend to reduce the speed at which the cargo completes its step to the next stator once it is bound by its toehold. For a large enough rise, there will be a reduction in the probability of successful completion of a motor's step following initial attachment. These results suggest that the experimentally observed reduction in stepping rate when the distance between stators is doubled [53] is predominantly due to the increased difficulty of completing displacement, rather than a reduced probability for forming an initial contact between the next stator and the toehold of the cargo. We argue that such a reduction in successful stepping could be used to provide higher sensitivity at junctions where motors must chose between two adjacent stators.

We find that the rigidity with which stators are held in place can be important. At low stiffnesses, stators can move towards each other and relax some of the tension generated by displacement. At high stiffnesses, the stators' movement is limited and

the DNA must stretch across the full gap between attachment points instead, resulting in the aforementioned rise in free energy. These results suggest that the flexibility of the surface to which the stators are attached, and the nature of the attachment, could be significant in determining the properties of motor stepping. We estimate, however, that anchoring stators such as those studied here to a single duplex is enough to qualify as a stiff attachment.

Our results have been obtained with a coarse-grained model, and the free-energy landscapes and profiles we have measured do not completely determine kinetics, which also depends on non-equilibrium effects. Nonetheless, the increase in tension within the motor as displacement proceeds is based on very general mechanical and structural properties of DNA that are known to be well reproduced by the model. So it is reasonable to assume that large stator separation does indeed lead to a rise in free energy with displacement progress. Furthermore, such a rise makes later displacement intermediates harder to reach and will eventually reduce the probability of successful step completion once the second stator is bound to the cargo's toehold. For the system considered in this work, a separation of around 8 or 9 nm should be sufficient to demonstrate the effects of stator separation. The most important assumption that we have made is that the small fragment of the first stator and the nicking enzyme tend to dissociate before attachment to the next stator can occur. Even if this is not the case, however, their influence will be strongest during toehold attachment and displacement initiation, and should not prevent the rise in free energy associated with increased tension at the later stages of displacement.

Future work on this topic could include explicit simulations of the dynamics of motor stepping, although such simulations will probably be computationally demanding even for a model as simple as oxDNA. It would also be worthwhile to study the consequences of different sequences using the sequence-dependent model, and motifs such as internal mismatches, on the operation of the motor.

Chapter 5

Coarse-grained model of RNA

In this chapter, we propose a new off-lattice coarse-grained RNA model, oxRNA, that follows the top-down coarse-graining approach developed for the DNA model oxDNA which we introduced in Chapter 2. Given that oxDNA has been successfully used to model DNA nanotechnological systems, such as motors [197], tweezers [149], kissing hairpins [188], strand displacement [196] as well as for biophysical applications such as cruciforms [198], the pulling of double-stranded DNA [159], and the systems studied in Chapter 3, our goal is to derive a model of similar applicability for RNA. We aim to capture basic RNA structure, mechanics and thermodynamics with a model of similar simplicity to oxDNA. We replace each RNA nucleotide by a single rigid body with multiple interaction sites. The interactions between rigid bodies are parametrized to allow an A-helix to form from two single strands and to reproduce RNA thermodynamics as predicted by Turner's nearest-neighbor model for RNA. The resultant model goes beyond nearest-neighbor thermodynamics because it has the ability to capture topological, mechanical and spatial effects and allows for the study of kinetic properties of various processes within a molecular dynamics simulation framework. Like oxDNA, the oxRNA model uses only pairwise interactions to facilitate the use of cluster Monte Carlo algorithms for simulations. The simple representation, one rigid body per nucleotide, allows for efficient simulation of structures of sizes up to several hundred nucleotides on a single CPU as well as of rare events such as the dissociation

or the formation of a double helix.

In Section 5.1 we present our coarse-grained model and its parametrization. In Section 5.2 we test the thermodynamic, structural and mechanical properties of the model. The detailed description of the interaction potentials in our model is provided in Appendix A. We further illustrate the utility of the model through applications to pseudoknot thermodynamics, hairpin unzipping and kissing hairpins in Chapter 6.

5.1 The RNA model and its parametrization

5.1.1 RNA thermodynamics and the nearest-neighbor model

Similarly to oxDNA, oxRNA is parameterized to reproduce thermodynamics of duplex association as described by a nearest-neighbor model. The nearest-neighbor model for RNA, developed by Turner and collaborators, is similar to SantaLucia's model (SL model) for DNA thermodynamics, which we discussed in Section 2.3.1. The nearest-neighbor model for RNA thermodynamics (hereafter referred to as the NN-model) was parametrized in an extensive series of investigations [81, 79, 84, 80] to describe the thermodynamics of RNA duplex and hairpin formation. It is widely used in RNA secondary structure prediction [85, 90, 94, 89, 199]. The model treats RNA at the level of secondary structure, estimating enthalpic and entropic contributions to the stability from each pair of consecutive base pairs (bp) in a structure and including corrections for end effects and enclosed loops of unpaired bases. The parametrization used melting experiments of short duplexes and hairpins at 1 M $[Na^+]$. As in the case of the SL model, the results were fitted using a two-state assumption in which RNA either adopts the fully-formed structure or is completely disordered. The yield of the duplexes with a particular sequence is given by Eq. 2.2, with $\Delta G^{\ominus}$ determined by the NN-model for RNA in the same fashion as it is determined by the SL model for DNA.

Similarly, the yield of hairpins in the NN-model is

$$\frac{[C]}{[O]} = \exp(-\Delta G^{\ominus}(T)/RT), \tag{5.1}$$

where $[C]$ is the concentration of closed strands, and $[O]$ is the concentration of open strands.

The NN-model has been shown to reproduce the melting temperatures of RNA oligonucleotides with Watson-Crick base-pairing with 1.3 °C accuracy [79]. As was the case for the SL model for DNA, we will treat the NN-model as an accurate fit to the melting data for RNA and use its melting temperature predictions for fitting the oxRNA model.

5.1.2 The Representation

OxRNA uses a single rigid body with multiple interaction sites to represent a nucleotide. Each rigid body has a backbone, 3′-stacking, 5′-stacking, cross-stacking, and hydrogen-bonding interaction sites. The detailed description of the representation of a nucleotide is given in Appendix A.1 (Fig. A.1). In the pictures of oxRNA model we use a schematic ellipsoid to represent the stacking and hydrogen-bonding sites as this allows the orientation of the base to be clearly seen. The potentials between the nucleotides are effective interactions that are designed to capture the overall thermodynamic and structural consequence of the base-pairing and stacking interactions, rather than directly representing the microscopic contributions such as electrostatics, dispersion, exchange repulsion and hydrophobicity.

Following the top-down coarse-graining approach, we choose the functional forms of our coarse-grained interactions to reproduce directly experimentally measured properties of RNA. Any coarse-grained interaction is actually a free energy for the real system, rather than a potential energy, and therefore it is in principle state-point dependent. So it should come to no surprise that our potential contains an explicit dependence on the temperature, although for simplicity we try to limit this as much

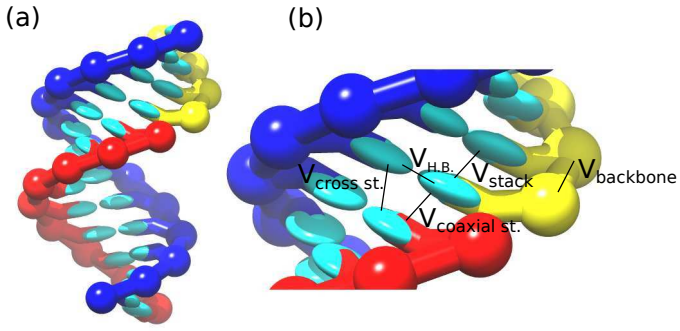

Figure 5.1: A schematic representation of (a) an A-RNA helix as represented by the model and of (b) the attractive interaction in oxRNA. The nucleotides can also interact with excluded-volume interactions.

as possible and only introduce it in one of the interaction terms (V_{stack}, as we will discuss later). Our coarse-graining aims to retain the relevant geometric degrees of freedom in order to still correctly capture the relative entropies of different states, despite not having temperature dependence in most of the interaction potentials [146].

The potential energy of the oxRNA model is

$$\begin{aligned} V_{\text{oxRNA}} &= \sum_{\langle ij \rangle} \left(V_{\text{backbone}} + V_{\text{stack}} + V'_{\text{exc}} \right) + \\ &+ \sum_{i,j \notin \langle ij \rangle} \left(V_{\text{H.B.}} + V_{\text{cross st.}} + V_{\text{exc}} + V_{\text{coaxial st.}} \right), \end{aligned} \tag{5.2}$$

where the first sum is taken over all the nucleotides that are neighbors along an RNA strand and the second sum is taken over all the non-nearest-neighbor pairs of nucleotides. All potentials are two-body potentials. There is a maximum distance beyond which all potentials are zero (with the exception of V_{backbone} which diverges to infinity as the distance between adjacent backbone sites approaches its maximum value). The interactions are schematically shown in Fig. 5.1. We discuss briefly the potentials here while the detailed description is given in Appendix A.2.

The backbone interaction, V_{backbone}, is an isotropic FENE (finitely-extensible non-linear elastic) potential and depends only on the distance between the backbone sites of the two adjacent nucleotides. This potential is used to mimic the covalent bonds in the RNA backbone that constrain this intramolecular distance. The nucleotides also

have repulsive excluded-volume interactions $V_{\rm exc}$ and $V_{\rm exc}^{'}$ that depend on the distance between the interaction sites, namely the backbone-backbone, stacking-stacking and stacking-backbone distances. The excluded-volume interactions ensure that strands cannot overlap, or pass through each other in a dynamical simulation.

The duplex is stabilized by hydrogen bonding ($V_{\rm H.B.}$), stacking ($V_{\rm stack}$) and cross-stacking ($V_{\rm cross\ st.}$) interactions. These potentials are highly anisotropic and depend on the distance between the relevant interaction sites as well as the mutual orientations of the nucleotides. The anisotropic potentials are of the form

$$V_{\rm H.B.} = \alpha_{ij} f_{\rm H.B.}\left(\mathbf{r}_{ij}, \mathbf{\Omega}_i, \mathbf{\Omega}_j\right) \tag{5.3}$$

$$V_{\rm stack} = \eta_{ij}(1+\kappa\, k_{\rm B}T) f_{\rm stack}\left(\mathbf{r}_{ij}, \mathbf{\Omega}_i, \mathbf{\Omega}_j\right) \tag{5.4}$$

$$V_{\rm cross\ st.} = \gamma f_{\rm cross\ st.}\left(\mathbf{r}_{ij}, \mathbf{\Omega}_i, \mathbf{\Omega}_j\right) \tag{5.5}$$

$$V_{\rm coaxial\ st.} = \mu f_{\rm coaxial\ st.}\left(\mathbf{r}_{ij}, \mathbf{\Omega}_i, \mathbf{\Omega}_j\right) \tag{5.6}$$

where the functions f are products of multiple terms, one of which depends on the distance between the relevant interaction sites and the remaining are angular modulation functions that are equal to one if the relevant angles between the nucleotides correspond to the minimum potential energy configuration, and smoothly go to zero as they depart from these values. The set of angles is different for each potential and includes angles between intersite vectors and orientations $\mathbf{\Omega}_i$ and $\mathbf{\Omega}_j$ of the nucleotides. The constant prefactors $\alpha_{ij}, \eta_{ij}, \gamma$, and μ set the strength of the interactions, with α_{ij}, η_{ij} being dependent on the nucleotides involved.

The hydrogen-bonding term $V_{\rm H.B.}$ is designed to capture the duplex stabilizing interactions between Watson-Crick and wobble base pairs. The potential reaches its minimum when two complementary nucleotides (AU, GC or GU) are coplanar, directly opposite and antiparallel with respect to each other and at the right distance.

The stacking interaction $V_{\rm stack}$ mimics the favorable interaction between adjacent bases, which results from a combination of hydrophobic, electrostatic and dispersion

effects. It acts only between nearest-neighbor nucleotides and its strength depends on both the distance between the respective 3′ and 5′ stacking sites of the nucleotides as well as their mutual orientations. It also depends on the vector between the backbone interaction sites in a way that ensures the inclination of the bases in the duplex structure matches that for A-RNA. We note that the nucleotides can also interact via the stacking interaction when they are in the single-stranded state. To ensure the right-handedness of the RNA helix in the duplex state, the stacking interaction has an additional modulation term that is equal to one if the nucleotides adopt a right-handed conformation and goes smoothly to zero in the left-handed conformation.

Similarly to oxDNA, the interaction strength of the stacking potential has a temperature-dependent contribution (the term $\kappa\, k_{\mathrm{B}}T$ in Eq. 5.4). This term was introduced in oxDNA [140] in order to correctly reproduce the thermodynamics of the stacking transition. We also found that retaining this temperature dependence enables oxRNA to reproduce more accurately the widths of the melting transitions, which are discussed in more detail in Section 5.1.4.

The cross-stacking potential, $V_{\text{cross st.}}$, is designed to capture the interactions between diagonally opposite bases in a duplex and has its minimum when the distance and mutual orientation between nucleotides correspond to the arrangement of a nucleotide and a 3′ neighbor of the directly opposite nucleotide in a A-helix. This interaction has been parametrized to capture the stabilization of an RNA duplex by a 3′ overhang [81]. OxRNA does not include any interaction with the 5′ neighbor of the directly opposite nucleotide, as 5′ overhangs are significantly less stabilizing than 3′ overhangs [81].

Finally, the coaxial stacking potential $V_{\text{coaxial st.}}$ represents the stacking interaction between nucleotides that are not nearest-neighbors on the same strand.

We note that, although oxRNA does not include an explicit term for electrostatic interactions between phosphates, these interactions are effectively incorporated into

the backbone repulsion. We chose to parametrize our model to the experimental data at 1 M [Na^+], where the electrostatic interaction is highly screened, making our approach reasonable. Furthermore, we are only able to capture those tertiary structure motifs that involve Watson-Crick and wobble base pairing or stacking, such as kissing hairpins or coaxial stacking of helices. In particular, oxRNA does not include Hoogsteen or sugar-edge hydrogen-bonded base pairs, or ribose zippers (interactions involving the 2′-OH group on the ribose sugar). In principle these interactions could be included, but for this version of the model we chose not to as there are no systematic thermodynamic data to which we could parametrize the relevant interaction strengths.

While the strengths of the hydrogen-bonding and stacking interactions depend on the identities of the interacting nucleotides, as in oxDNA, all nucleotides in oxRNA have the same size and shape. Therefore we do not expect oxRNA to capture detailed sequence-dependent structure of the A-helix.

The positions of the minima in the potential functions have been selected so that the model reproduces the structure of the A-RNA double helix, which RNA duplexes have been shown to adopt [9, 15] and which we describe in more detail in Section 5.2.1. The widths of the potential functions and the strengths of the interaction potentials were parametrized to reproduce RNA thermodynamics as described in Section 5.1.4.

5.1.3 Simulation methods

Algorithms

For the majority of our simulations, unless noted otherwise, we use the Virtual Move Monte Carlo algorithm (VMMC), often combined with umbrella sampling method, as described in Appendix B.2. We also implemented the oxRNA forcefield in a molecular dynamics (MD) code with an Andersen-like [200] thermostat, which is also briefly outlined in Appendix B.3.

The methods for calculating melting temperature are the same as used for the oxDNA model. In the course of parametrization of the interactions of the oxRNA model, we also employ the simulated annealing fitting algorithm that uses histogram reweighting method to calculate melting temperatures, which was introduced in Section 2.3.2.

5.1.4 Parametrization of the model

The anisotropic potentials V_{stack}, $V_{\text{H.B.}}$ and $V_{\text{cross st.}}$ have interaction strengths of the form $\eta_{ij}(1 + \kappa\, k_{\text{B}}T)$, α_{ij} and γ respectively, where the stacking interaction strength depends also on the simulation temperature T and i and j correspond to the types of interacting nucleotides (A, C, G, U). The magnitude of the temperature dependence of the stacking (κ) and the cross stacking interaction strength (γ) are set to be the same for all nucleotide types.

In the first step of the fitting procedure, we parametrize the model to reproduce the melting temperatures of the averaged NN-model, for which we define the enthalpy and entropy contribution per base-pair step by averaging contributions of all possible Watson-Crick base-pair steps in the NN-model. In calculating average melting temperatures of different motifs, such as hairpins or terminal mismatches and internal mismatches, the additional entropy and enthalpy contributions for a particular motif in the NN-model were again averaged over all possible combinations of bases. In the averaged NN-model, the melting temperature is hence independent of the particular sequence, but depends only on the lengths of the sequence and the particular secondary structure motif.

The fitting of the interaction strength parameters was done by a simulated annealing algorithm, as in the case of oxDNA, which was outlined in Section 2.3.2. The function to minimize (defined in Eq. 2.6 for the oxDNA parametrization) is the sum of absolute differences between the melting temperatures of a set of systems as calculated by oxRNA and as predicted by the NN-model.

First, the oxRNA model was parametrized to reproduce averaged NN-model melting temperatures of structures with only Watson-Crick base pairs. The interaction strength η_{ij} was hence set to η_{avg} for all base pair types i and j and α_{ij} was set to α_{avg} for Watson-Crick complementary nucleotides and 0 otherwise. The initial values for α_{avg}, η_{avg} and γ were first chosen by hand and then refined based on results of VMMC simulations in order to reproduce melting temperatures as predicted by the averaged NN-model of short duplexes of lengths 5, 6, 7, 8, 10 and 12 bp and of duplexes of lengths 5, 6, 8 bp with one overhanging nucleotide at either both $3'$ ends or both $5'$ ends. We set κ to be equal to 1.9756 (in the inverse of the energy unit used by the simulation code, as defined in Appendix A), the same value as was used by the oxDNA model. We found that leaving κ as a free fitting parameter did not lead to a significantly better fit to the considered sequences and motifs.

We note that for some applications, where one is more interested in the qualitative or generic nature of the studied system or one wants to average over all possible sequences, it might be more useful to study the system with a sequence independent model. We refer to such a model as the "average-base" model meaning that η_{ij} are set to η_{avg} for all types of bases and α_{ij} are set to α_{avg} for all Watson-Crick complementary base pairs (GC and AU) and 0 otherwise. If one is interested in sequence-specific effects, then a sequence-dependent parametrization is necessary, as in the case of the oxDNA model.

We used the final values of η_{avg}, α_{avg} as the initial values for fitting the sequence-dependent values η_{ij}, α_{ij}, with i and j being Watson-Crick or wobble base pairs (AU, GC, GU). The parameters were fitted to an ensemble that contained oligomers of the above mentioned sizes and hairpins with stem lengths 6, 8 and 10 and loop lengths 5, 6, 7, 8, 10 and 15. One hundred sequences with only Watson-Crick base pairs were randomly generated for each size, along with 533 further random duplexes of lengths 5 to 12 bp containing GU wobble base pairs. We excluded sequences with

neighboring wobble base pairs in the fitting process as these can lead to duplexes with particularly low melting temperatures (some of the base-pair steps containing wobble base pairs are actually destabilizing at room temperature [81]). We found that our model was unable to accurately fit melting temperatures of duplexes that contain neighboring GU/UG or UG/GU wobble base pairs, probably due to the fact that we do not account for the structural changes that these induce in the duplex.

We note that if one included only Watson-Crick base pairs in the sequence-dependent fitting (as was the case for the sequence-dependent parametrization of the oxDNA model), it would not be possible to distinguish between certain stacking interaction types. For instance, the contribution of AA and UU base stacking interactions always appear together in the AA/UU base pair step free-energy contribution in the NN-model. However, including wobble base pairs in the fitting ensemble provides additional information, for example the UU stacking contribution also appears in the AG/UU base-pair step. We therefore do not need to restrict the strength of stacking interaction to be the same for certain types of nucleotides, as was the case for the oxDNA model.

Finally, we parametrized the coaxial stacking interaction potential, $V_{\text{coaxial st.}}$, which captures the stacking interaction between two bases that are not neighbors along the same strand. Experiments have measured this interaction by a comparison of the melting of a 4-base strand with its complement, or with a hairpin with a 4-base overhang with the complementary sequence adjacent to the hairpin stem. They found for both DNA and RNA that the melting temperature increases for the 4 bp long strand attached to the overhang on the hairpin stem, which was attributed to the extra stabilizing interactions with the adjacent stem [82, 83, 74, 201]. The coaxial stacking free energy has been incorporated in the NN-model by assuming that the free-energy stabilizations in these experiments are similar in strength to the actual base-pair steps with the same sequence. The NN-model hence uses the same free

energy contribution for a base pair coaxially stacked on a subsequent base pair (as illustrated in Fig. 5.1) as it uses for a base pair step in an uninterrupted duplex. In order to parametrize these interactions for oxRNA, we performed melting simulations of a 5-base strand, which was able to associate with a complementary 5′ overhang on a longer duplex (which itself was stable). We fitted the interaction strength μ of the coaxial stacking interaction in our model so that it would match the prediction of the melting temperature by the averaged NN-model. We note that in our model, the contributing factors to stabilization are both the coaxial stacking interaction and the cross-stacking interaction between the 5-base strand and the hairpin.

5.2 Properties of the model

We describe the structural properties of the model and report the thermodynamics of duplexes, hairpins and other secondary structure motifs as represented by oxRNA. We further study some of its mechanical properties, namely the persistence length of a duplex, the force-extension curve for duplex stretching, and the overstretching transition.

5.2.1 Structure of the model

As mentioned in Section 5.1.4, the coarse-grained interactions were selected so that the model reproduces the A-form helix that RNA duplexes have been shown to adopt at physiological conditions [9, 11].

The A-RNA structure is significantly different from B-DNA, the prevalent duplex structure found in DNA molecules. These differences are mainly caused by the sugars in A-RNA adopting a more twisted conformation (C3′ *endo* pucker) as a result of the presence of an extra OH group on the sugar. The A-RNA duplex has a reported helical twist ranging [11] from 32.7° to 33.5° per base pair, corresponding to a pitch of 10.7 to 11 base pairs. The rise per base pair reported by X-ray measurements [11]

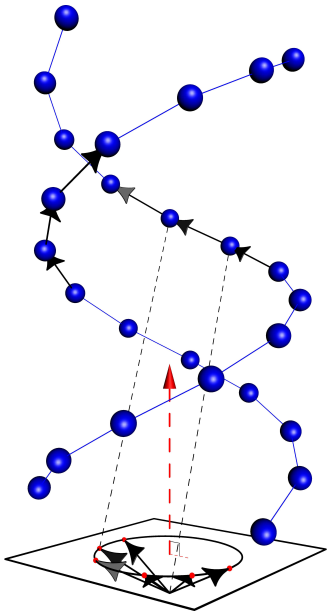

Figure 5.2: Fitting a helical axis to the duplex. The blue spheres show the positions of the backbone sites. The black arrows represent vectors pointing from a nucleotide backbone site to its neighbor's backbone site. When the vectors are placed onto the same origin, their endpoints would lie in a plane for the case of a perfect A-helical structure. A plane can hence be fitted through the endpoints of these vectors. A vector perpendicular to this plane is used as the helical axis (red dashed arrow).

is about 0.28 nm. The bases are displaced from the helical axis, i.e. the helical axis does not pass through the base pair mid-points as it is approximately the case for the B-DNA helix. Finally, the bases are not perpendicular to the helical axis, but are inclined at an angle of about 15.5°. Although the width of A-RNA is reported to be about 2.1 nm from X-ray crystal structures [12], Reference [202] uses an effective hydrodynamic diameter of 2.8 nm for the structure. The A-RNA helix has a narrow major groove (0.47 nm) and a wide minor groove (1.08 nm).

To characterize the structure of the oxRNA duplex, we simulate a 13-base-pair duplex at 25 °C using Monte Carlo simulation. We generated 30 000 decorrelated configurations that were analyzed in the following manner.

The helical axis was fitted for each saved configuration. The fitting was done in the following way, schematically illustrated in Fig. 5.2. For each base in the first strand, we took the vector pointing from its backbone site to the backbone site of

its 3′-neighbor and for each base in the second strand, we considered the vectors pointing to the 5′-neighbor's backbone site. For a perfect A-helical structure, the endpoints of all the vectors would all lie in the same plane if the origins of the vectors were all placed at the same point. The structure of the duplex is subject to thermal fluctuations and hence the plane has to be fitted through the endpoints of the vectors. The first and last two base pairs were not included in order to exclude end effects. The vector perpendicular to the plane was then taken to be the helical axis. The rise per base pair was measured as the distance between the projections of the midpoints of base pairs onto the helical axis. The length scale in the oxRNA model is defined so that the rise per base pair is 0.28 nm. The twist per base pair was measured as the angle between the projections of the vectors connecting bases in the base pairs onto the plane perpendicular to the helical axis. The mean turn per base pair in the model is 33.0°, corresponding to a pitch of 10.9 base pairs. The inclination, measured as the mean angle between the vector pointing from the center of mass of a nucleotide to its base and the plane perpendicular to the helical axis, is 16.1°.

The width of the helix is measured as twice the distance of the backbone from the axis, and includes the excluded volume interaction radius of the backbone site. The helix width in oxRNA is 2.5 nm. The major and minor grooves in oxRNA are 0.48 nm and 1.07 nm, respectively, where we measured the groove distances in a manner analogous to a method employed by the Curves+ software [203] for analyzing atomistic structures of DNA and RNA. For a selected nucleotide, we measured distances between its backbone site and points on a curve that was linearly interpolated through the backbone sites of the nucleotides on the opposite strand. The distances measured along the curve have two minima, one for each groove. The excluded volume interaction radius for each backbone site was subtracted from these measured distances.

Motif	$T_m - T_m(\mathrm{NN^{avg}})$	$T_m[^\circ\mathrm{C}]$
5-mer	0.6	26.4
6-mer	0.1	42.5
7-mer	-0.1	53.6
8-mer	-0.7	61.2
10-mer	-0.5	72.5
12-mer	-0.9	79.3
6-mer ($3'$ overhangs)	-1.2	49.8
6-mer ($5'$ overhangs)	-2.8	43.1
8-mer ($3'$ overhangs)	-0.7	65.6
8-mer ($5'$ overhangs)	-3.0	61.9
8-mer (terminal mismatch)	-2.0	56.0

Table 5.1: The melting temperatures of a series of duplexes for the average-base parametrization for oxRNA (T_m) compared to the averaged NN-model ($T_m(\mathrm{NN^{avg}})$). The melting temperatures were calculated from VMMC simulations and are for a strand concentration of 3.5×10^{-4} M. For structures with overhangs, two single-base overhangs were present either at the $3'$ or $5'$ ends. The 8-mer with a terminal mismatch had a non-complementary base pair at one of the ends of the duplex.

5.2.2 Thermodynamics of the model

In this section, we examine the thermodynamics of duplexes, hairpins, bulges, and internal and terminal mismatches as represented by oxRNA. We compare the melting temperatures as predicted by oxRNA with the melting temperatures calculated from the NN-model (denoted as $T_m(\mathrm{NN})$) for different sequences and different secondary structure motifs. To calculate the melting temperatures, we used the reweighting method which was introduced for parametrization of oxDNA in Section 2.3.2, with the states that were generated from VMMC simulations of melting for the average-base parametrization of oxRNA.

Duplex and hairpin melting

A comparison of the melting temperatures of the average-base parametrization of oxRNA with the thermodynamics of the averaged NN-model for structures involving only Watson-Crick base pairs is shown in Table 5.1. For this averaged model, the differences are roughly on the order of the accuracy of the NN-model itself.

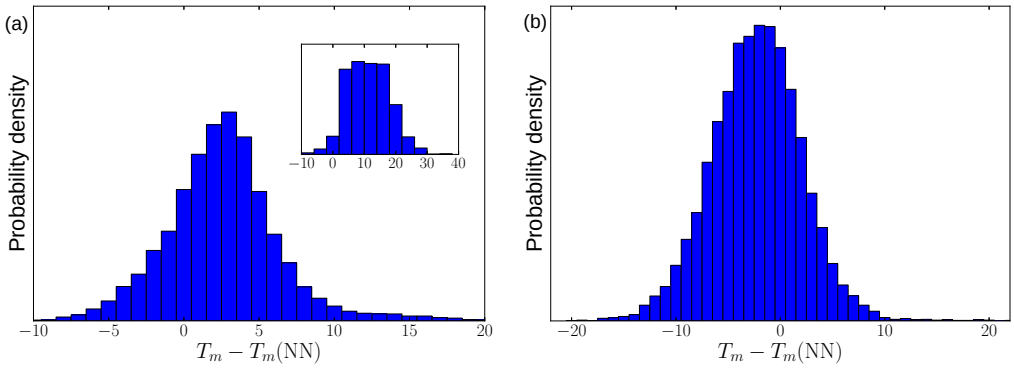

Figure 5.3: (a) The histogram of differences between melting temperatures as predicted by the oxRNA model (T_m) and by the NN-model (T_m(NN)) for a set of 20 255 randomly generated RNA duplexes of lengths 6, 7, 8, 10 and 12 with Watson-Crick and wobble base pairing. The main plot shows a histogram of values of T_m - T_m(NN) for duplexes that do not include GU/UG or UG/GU base pair steps. The inset shows a histogram of values of T_m - T_m(NN) for 1439 randomly generated sequences that contained at least one GU/UG or UG/GU base pair steps. (b) The histogram of differences between melting temperatures as predicted by the oxRNA model and by the NN-model for a set of 12 000 randomly generated hairpins with stems of lengths 6, 8, and 10 and loops with lengths of 5, 6, 7, 8, 10 and 15, where the stems only contain Watson-Crick base pairs

To test the sequence-dependent parametrization of the hydrogen-bonding and stacking strength of the interactions, we calculated the melting temperatures for randomly generated ensembles of RNA duplexes, different from the ensemble used for parametrization. A histogram of the differences in the melting temperature predicted by the sequence-dependent version of oxRNA (T_m) and those calculated from the nearest neighbor model (T_m(NN)) is shown in Fig. 5.3(a). The main histogram is for duplexes with both Watson-Crick and wobble base pairing, but not containing GU/UG or UG/GU base pair steps. For convenience, the generated ensembles of sequences also do not include any self-complementary sequences because the calculation of their melting temperatures requires a different finite size correction [167, 166]. The average difference in melting temperatures is 2.0 °C, with an average absolute deviation of 3.3 °C. The histogram in the inset of Fig. 5.3(a) is for sequences containing at least one GU/UG or UG/GU base pair step. The average difference in melting

temperatures for the ensemble is 9.3 °C and the absolute average deviation is 9.6 °C. We note that in the NN-model, the free-energy contribution of these base pair steps is positive at 37 °C, meaning that they actually destabilize the duplex. However, in the oxRNA model, the cross-stacking, stacking and hydrogen-bonding interactions are always stabilizing interactions and the interaction strength of two hydrogen-bonded nucleotides does not depend on the identity of their respective neighbors on the strand. Our coarse-grained model hence cannot capture the free-energy contributions of GU/UG and UG/GU base pair steps. One could imagine adding multi-body interactions, but for the sake of computational efficiency and maintaining the consistency of our coarse-graining methodology, we do not do so in this study. Another option might be to introduce a structural perturbation of the helix caused by the GU base pairs.

The histogram in Fig. 5.3(b) shows the difference between melting temperatures calculated by oxRNA and those predicted by the NN-model for an ensemble of randomly generated hairpins. The average melting temperature difference was −2.8 °C with the average absolute deviation being 4.1 °C.

The transition widths for duplex and hairpin formation were calculated for the averaged model as the difference between the temperatures at which the yield is 0.8 and 0.2, respectively. This quantity is important because the widths of the transition determine the change of the duplex melting temperature with concentration. It can be shown [150] that the derivative of the melting temperature as a function of strand concentration is proportional to the width of the transition. The melting simulations of oxRNA with the average-base parametrization were compared with the width predicted by the averaged NN-model. For the duplexes of lengths 6, 8, 10, and 12 the width was on average underestimated by 0.8 °C, but was overestimated for a 5-mer by 0.5 °C. The width of the transition for the averaged NN-model decreases from 20.5 °C for a 5-mer to 9.2 °C for a 12-mer. For a set of hairpins with stems of length

Motif	$\langle \Delta T_m \rangle$	$\langle \lvert \Delta T_m \rvert \rangle$	$\Delta T_m^{\text{duplex}}(\text{NN})$	$\Delta T_m^{\text{duplex}}$
Bulge (1 b)	-3.8	4.3	-13.0	-19.5
Bulge (2 b)	-2.1	4.3	-17.5	-22.4
Bulge (3 b)	-4.4	6.0	-19.6	-27.0
Bulge (6 b)	-0.9	5.2	-25.4	-29.1
Internal mis. (1 b)	10.0	10.0	-18.0	-10.9
Internal mis. (2 b)	8.8	9.4	-27.2	-21.2
Internal mis. (3 b)	16.3	16.4	-45.0	-31.5

Table 5.2: Melting temperatures for bulge and internal mismatch motifs in a 10-mer. $\langle \Delta T_m \rangle = \langle T_m - T_m(\text{NN}) \rangle$ is the average difference between the melting temperature of the oxRNA model (T_m) and the melting temperature as predicted by the NN-model ($T_m(\text{NN})$). $\langle \lvert \Delta T_m \rvert \rangle = \langle \lvert T_m - T_m(\text{NN}) \rvert \rangle$ is the average absolute difference in melting temperatures. $\Delta T_m^{\text{duplex}}(\text{NN})$ and $\Delta T_m^{\text{duplex}}$ are the average differences in melting temperature between the sequences with a secondary structure motif and a duplex with the same sequence but with no bulge or internal mismatch as predicted by the NN-model and oxRNA respectively. Each of the motifs considered is destabilizing, resulting in a decrease of the melting temperature. The averages were taken over an ensemble of randomly generated sequences (1000 for each motif) that had 10 complementary Watson-Crick base pairs for the bulges, and 9, 8, and 7 complementary base pairs for internal mismatches of size 1, 2 and 3 bases, respectively. The bulges that we consider were of the size 1, 2, 3 and 6 bases. All the melting temperature calculations were calculated for an equal strand concentration of $3.5 \times 10^{-4}\,\text{M}$.

6, 8, and 10 bp and with loops of lengths 5, 6, 7, 8, 10 and 15, the width of the melting transition was on average underestimated by 1.5 °C. The width of the hairpin transition decreases from about 12 °C for stems of length 6 to approximately 8 °C for stems of length 10 in the averaged NN-model. However, the trend of increasing width with decreasing size is always captured by the oxRNA model.

Finally, we note that we could have parametrized the sequence-dependent model only to duplex melting temperatures, as for oxDNA, which would then have led to a larger underestimate of hairpin melting temperatures, presumably because our representation of the strand is too simple to exactly capture the entropy and enthalpy of the loop formation. We hence chose to parametrize to the ensemble of duplexes and hairpins because hairpins are a prominent secondary structure motif in RNA. Comparisons with the parametrization of oxDNA will be discussed in Sec. 5.3.

Thermodynamics of secondary structure motifs

Given that our aim is to design a model that goes beyond describing hybridization of simple duplexes, it is important to assess how well it is able to reproduce the thermodynamic variation induced by common secondary structure motifs such as bulges or mismatches. To this end, we have studied the melting temperature changes induced by internal mismatches, terminal mismatches and bulges of different lengths.

To assess the effects of bulges, we consider a systems of two strands, one with 10 bases and the other with 10 complementary bases and extra bases that create a bulge motif. We considered bulges of lengths 1, 2, 3 and 6, positioned in the center. For each sequence considered we calculated the melting temperatures by reweighting a set of 6000 states that were sampled from a melting simulation using the average-base parametrization. For each bulge length, we considered 1000 randomly generated sequences with Watson-Crick base pairing in the complementary part.

We further evaluated the melting temperatures for randomly generated 10-mers which contained 1, 2 or 3 consecutive mismatches (and therefore had 9, 8 or 7 complementary Watson-Crick base pairs). The average difference and absolute average deviation for the considered bulges and mismatches are shown in Table 5.2. The melting temperatures of duplexes with bulges are underestimated by a few degrees. However, the model presently significantly overestimates the stability of internal mismatches by the order of 10 °C or more. Even though the mismatching base pairs do not gain stabilization from hydrogen-bonding interactions (which are zero for bases that are not complementary), they still have favorable cross-stacking and stacking interactions, which have their minimum energy configuration in an A-helical configuration, which the oxRNA model can still form with the mismatches presents. We further note that our model represents each nucleotide by the same rigid body structure, so the mismatching base pairs do not cause any distortion to the duplex structure in our model. Improving the model to more accurately represent the sec-

ondary structures with mismatching nucleotides could be the subject of future work on the improvement of the oxRNA model.

5.2.3 Mechanical properties of the model

Persistence length

The persistence length L_p of dsRNA molecule measured in experiments is reported to be between 58 nm to 80 nm [204, 205, 14], corresponding to 206–286 bp (assuming 0.28 nm rise per base pair). The first studies of the persistence length of dsRNA used electron micrographic, gel-based and hydrodynamic measurements (reviewed in Ref. [205]) and reported the persistence length to be between 70 to 100 nm, in salt conditions ranging from 6 mM $[Mg^{2+}]$ and 0.01 M to 0.5 M $[Na^{+}]$. A more recent single-molecule experimental study [204] in 0.01 M $[Na^{+}]$ and 0.01 M $[K^{+}]$ buffer measured the persistence length by analyzing force-extension curves in magnetic tweezers experiments as well as by analyzing atomic force microscopy images of the RNA duplexes. The two methods yielded consistent values with the measured persistence length estimated as 63.8 and 62.2 nm respectively, corresponding to 227 and 222 bp. Finally, a recent single molecule force-extension study [14] of dsDNA and dsRNA at salt concentrations ranging from 0.15 M to 0.5 M $[Na^{+}]$ found the dsRNA persistence length to decrease from 67.7 to 57.7 nm with increasing salt concentration, and the extrapolation of measured persistence lengths to higher salt concentrations approaches asymptotically 48 nm (171 bp).

To measure the persistence length in our model, we simulated an 142-bp long double-stranded RNA with the average-base oxRNA model, and measured the correlations in the orientation of a local helical axis along the strand. The local axis vector $\hat{\mathbf{l}}_i$ is fitted through the i-th base pair and its nearest neighbors, using the approach described in Section 5.2.1, but considering only the nearest neighbors. We found the results to be robust even when 2 or 3 next nearest neighbors were included in the construction of the local axis. To account for edge effects, a section of ten base

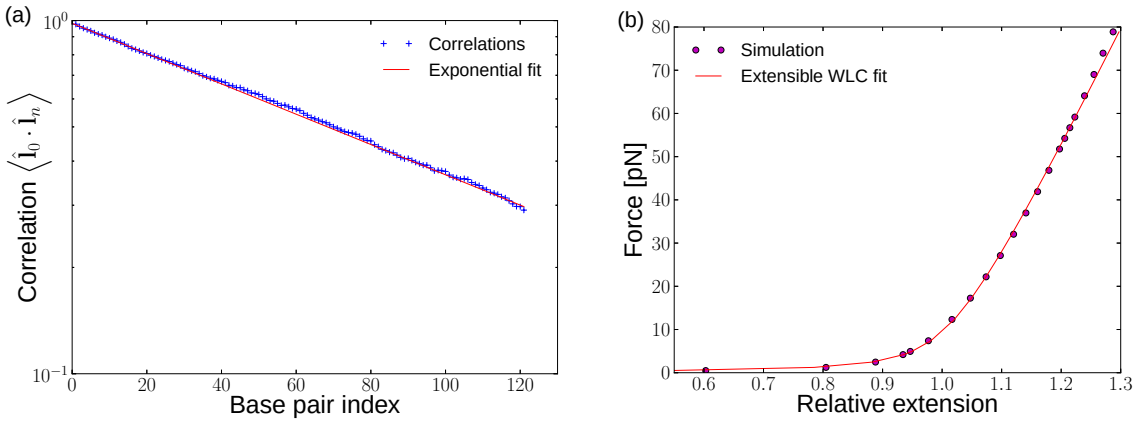

Figure 5.4: (a) Semi-logarithmic plot of the correlation function for the direction of the local helix axis along the duplex as defined in Eq. 5.7 and an exponential fit to the data. (b) The force-extension curve of a 81-bp section of a 99-bp duplex. The extension is normalized with respect to the contour length of the 81-bp duplex (with a rise of 2.8 nm per bp) and a fit to the data using the extensible wormlike chain model defined in Eq. 5.8 is also plotted.

pairs at each end of the duplex was ignored when accumulating averages. For an infinitely long, semi-flexible polymer in which the correlations decay exponentially with separation along the strand, the persistence length L_p can be obtained from [206]

$$\left\langle \hat{\mathbf{l}}_0 \cdot \hat{\mathbf{l}}_n \right\rangle = \exp\left(-\frac{n\,\langle r\rangle}{L_p}\right) \tag{5.7}$$

where $\langle r\rangle$ is the rise per base pair. This correlation function is shown in Fig. 5.4(a) along with the exponential fit from which we estimated the persistence length of our model to be about 101 base pairs, somewhat lower than the 171 bp persistence length expected at this salt concentration. Our model's persistence length is hence smaller than the experimentally measured values, but still within a factor of 2. OxRNA hence captures the correct order of magnitude for the persistence length and as long as one studies structures whose duplex segments are smaller than the persistence length of the model, it should provide a physically reasonable description.

We note that the persistence length is quite hard to correctly reproduce. We expect this issue to hold for other coarse-grained RNA models as well. To our knowledge, the persistence length has not been measured yet in these models.

Force-extension properties

As a further test of the mechanical properties of the model, we measured the extension of a 99-bp RNA duplex as a function of applied force for the average-base model. We used a constant force acting on the center of mass of the first and last nucleotides in one of the two strands in the duplex. We focus on the average extension between the 11^{th} and 91^{st} nucleotide on this strand in order to avoid end effects, such as from fraying of base pairs, in our measurements.

We compare our force-extension data with an extensible worm-like chain model [156], which provides the following expression for the projection of the end-to-end distance $\mathbf{R}$ along the direction of the force $\hat{\mathbf{z}}$:

$$\langle \mathbf{R} \cdot \hat{\mathbf{z}} \rangle = L_0 \left(1 + \frac{F}{K} - \frac{k_{\mathrm{B}}T}{2FL_0} \left(1 + A \coth A \right) \right) \tag{5.8}$$

where

$$A = \sqrt{\frac{FL_0^2}{L_p k_{\mathrm{B}} T}},$$

K is the extension modulus and L_0 is the contour length. This expression comes from an expansion around the fully extended state, and thus it is expected to be valid at forces high enough for the polymer to be almost fully extended.

It was shown experimentally [14] that this extensible worm-like chain model describes the behavior of dsRNA prior to the overstretching transition. In particular, at 0.5 M [Na^+], the extensible worm-like chain model fit to the experimentally measured force-extension curve yielded $L_p = 57.7\,\mathrm{nm}$ and $K = 615\,\mathrm{pN}$.

The force-extension curve for oxRNA is shown in Fig. 5.4(b). We used data from forces between 2.4 pN and 69 pN for our fitting and allowed L_0, K and L_p to be free parameters. From the fit we obtained $L_0 = 23.4\,\mathrm{nm}$ (84 bp), $K = 296\,\mathrm{pN}$ and $L_p = 26.0\,\mathrm{nm}$ (93 bp). We note, however, that Eq. 5.8 is not a robust fit for our model: changing the fitting interval and thus including or excluding points at either high or low forces significantly changes the resulting values of the fitting parameters,

even though the residual error in the fit remains small. The persistence length, for instance, can change by more than a factor of two. In the interval we selected for the fit, the L_p obtained approximately corresponds to that obtained from the correlation function in Fig. 5.4(a), but given the sensitivity of the fit (which was not observed for the oxDNA force-extension curve [150]), the errors on these extracted parameter values should be taken to be quite large. Another issue to keep in mind is that the inclination angle is also affected by the force. At 10 pN, this is roughly a 1 to 2 degree change, but close to the point where the chain starts to significantly overstretch (as discussed in the next section), the inclination has disappeared, and the bases are almost perpendicular to the axis. It is likely that this deformation is not entirely physical. However, the physical structure of stretched RNA is not experimentally known. In DNA, the structure of the extended state is a very active topic of research.

Overstretching

Both DNA and RNA duplexes are known to undergo an overstretching transition at high enough force. Recent experiments [14] for different salt concentrations find 63.6 (2.0) pN for RNA overstretching at 0.15 M [Na^+] up to 65.9 (3.3) pN at 0.5 M [Na^+]. Following the approach taken in the study of DNA overstretching with the oxDNA model [159], we used the average-base oxRNA model to run VMMC simulations of a 99-bp RNA duplex with equal and opposite forces applied to the first and last nucleotide of one strand. To aid equilibration, only native base pairs were allowed to form, i.e. no misbonds in the duplexes or intrastrand base pairs are present in the unpeeled strand.

The simulations were started from a partially unpeeled state to sample states which have between 65 and 71 bp. The obtained free-energy profiles as a function of the number of base pairs are shown in Fig. 5.5(a). As the force increases, the slope of the free energy profiles changes from negative (states with more base pairs are favored) to positive (it is favorable for duplexes to unpeel). By estimating the force at which

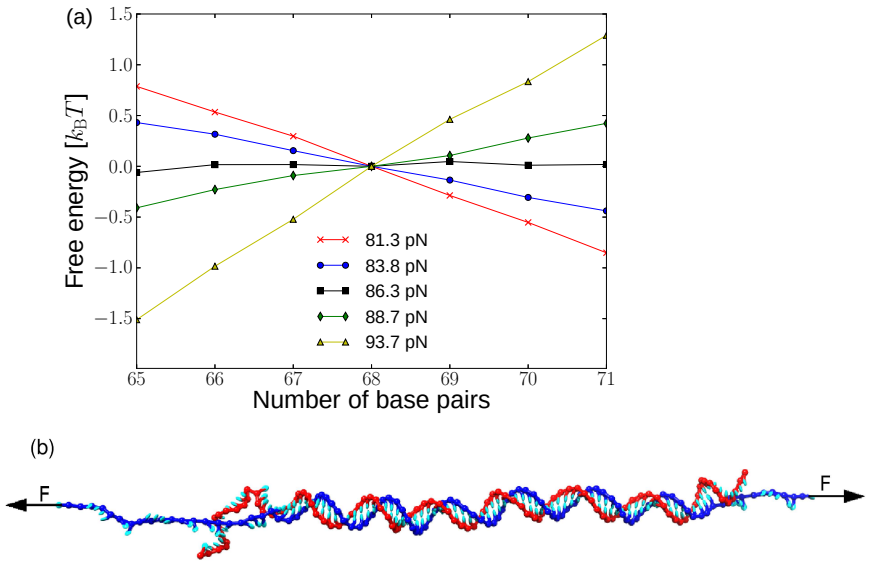

Figure 5.5: (a) Free energy as a function of the number of base pairs in the duplex for different forces, where we set the free energy to be 0 for 68 bp for each force considered. At the overstretching force, the slope of the free-energy profile is 0. (b) Snapshot from a VMMC simulation at $F = 86.3\,\mathrm{pN}$, showing unpeeling from the ends.

the slope becomes zero, we obtained 86.2 pN as the overstretching force. We note that our model was parametrized for 1 M [Na^+], whereas the overstretching experiment was done at 0.5 M. Furthermore, by not allowing any formation of secondary structure in the unpeeled strands, we overestimate the overstretching force in the model, because these intramolecular base pairs stabilize the unpeeled state. For the oxDNA model, it was shown that allowing secondary structure decreases the overstretching force by about 3 pN [159]. We would expect the stabilization to be slightly higher for RNA, as RNA base pairs are more stable. Our model hence overestimates the value of the overstretching force by about 16-20 pN. The overestimation of the overstretching force is partly due to the higher extensibility of the duplex in oxRNA, which is aided by the loss of inclination in the duplex when higher forces are applied, as we already discussed in the previous section.

5.3 Overview of the oxRNA model and comparison with the coarse-graining of oxDNA

We have developed a new off-lattice coarse-grained model for RNA, oxRNA, which aims to capture basic thermodynamic, structural and mechanical properties of RNA structures with a minimal representation and pair-wise interaction potentials. OxRNA is specifically developed to allow for efficient simulations of large structures, and for reactions involving multiple strands of RNA, which are important for applications in RNA nanotechnology. Our coarse-graining strategy is closely linked to the previous successful coarse-graining of DNA with oxDNA. Rather than focusing mainly on reproducing structure, as many other previous models have done, here we tried to systematically compare to a whole suite of different properties.

We employed a "top-down" coarse-graining approach, where we aim to reproduce free-energy differences between different states (such as opened and closed state of a hairpin) as measured in experiment. OxRNA represents each nucleotide (i.e., sugar, phosphate and base) as a single rigid body with multiple interaction sites. The model is capable of representing RNA structures such as hairpins and duplexes and is designed to reproduce the A-helical form of duplex RNA. We used a histogram reweighting method, developed in Chapter 2, to parametrize the model to reproduce the thermodynamics of short duplexes and hairpins. Currently, the oxRNA model captures Watson-Crick and wobble base-pairing interactions as well as various types of stacking interaction. However it was not designed to capture non-canonical interactions such as Hoogsteen or sugar-edge hydrogen-bonded base pairs, or ribose zippers. Nevertheless, it can reproduce some important tertiary interaction motifs, in particular coaxial stacking of helices, kissing loop interactions, and pseudoknots, which we will study in Chapter 6.

The model is currently parametrized for a salt concentration of 1 M, as this corresponds to the conditions for the melting experiments used for the nearest-neighbor

model to which our model was parametrized. Explicit electrostatic interactions are not included, because they are very short-ranged at high salt and thus can be incorporated into the short-ranged excluded volume terms in the potential. This excluded volume also prevents a strand from crossing itself or other strands, forbidding topologically impossible trajectories in kinetic simulations. It is possible to use the same coarse-graining techniques to parametrize the model at lower salt concentrations. However, as the screening lengths become longer, different longer-ranged forms for the interactions may need to be used to capture the correct physics. We note, however, that nanotechnology experiments *in vitro* are usually carried out in high salt conditions. But of course experiments *in vivo* will need to be described by a model parametrized to similar solution conditions, which might be a non-trivial challenge to overcome.

To test our model, we investigated the thermodynamics of short duplexes and hairpins with different sequence content, as well as various other secondary structures such as bulges, internal and terminal mismatches. We found that in comparison with oxDNA, parametrizing RNA thermodynamics is a more challenging task. We found, for example, that with the sequence-dependent oxDNA parametrization we agree with the SL model predictions for duplexes to within a standard deviation of 0.85 °C, whereas for oxRNA we found for the duplexes with Watson-Crick and wobble base pairs (shown in the main plot of Fig. 5.3(a)) a standard deviation of 4.07 °C when compared to the NN-model. The larger deviation for the duplexes in the RNA model might be partially caused by the fact that besides duplexes, we also included hairpins in the fitting ensemble. Furthermore, experimental melting temperatures of a duplex of a given length can differ by as much as 70 °C between weak and strong sequences with Watson-Crick base pairs, as opposed to 50 °C for DNA. So sequence effects are stronger in RNA. Including wobble base pairs presents further challenges, as some base pair steps that include two or more neighboring wobble base pairs have a positive

contribution to the free energy of duplex. Although it is not possible to capture this effect with the current representation of our model, adding the structural effects of wobble base pairs on the double helix may provide means to improve this aspect of the model.

Finally, we found that oxRNA overestimates the stability of duplexes with mismatches in internal loops considerably more than oxDNA does. This could lead to an overestimation of the stability of misfolded structures and complicate the study of the folding of RNA strands that have multiple metastable states with mismatches. Nevertheless, even though oxRNA does not reproduce the exact melting temperatures for structures with internal mismatches and bulges, we do observe, as expected, a decrease in melting temperatures of a duplex with internal mismatches or bulges with respect to the fully complementary strands. The observed changes in melting temperatures are within the same order of magnitude as predicted by the NN-model for the same motifs and capture correctly the direction of the change.

We have tested the mechanical properties of the RNA duplex as represented by the model and found its persistence length to be about half of the reported persistence length of RNA molecules at high salt conditions. The model hence has the correct order of magnitude for the duplex persistence length and provides sufficiently accurate mechanical behavior for most applications that are suited to oxRNA, where individual double helical sections are likely to be much shorter than the persistence length.

At this point it is interesting to reflect on some similarities and differences between the coarse-graining of oxDNA and oxRNA. Although oxRNA can clearly reproduce a good number of properties of RNA, quantitative differences with experiment for the melting temperatures of certain motifs are larger than they are for oxDNA. Moreover, it was more difficult to simultaneously reproduce the thermodynamics and the correct persistence length or the force-extension curves. One reason for these differences may be that RNA itself exhibits more complex behavior than DNA, and

so is harder to coarse-grain. It is intuitively obvious that the compromises made to increase tractability mean that not all properties of the underlying system can be simultaneously captured by a more simplified description, a "representability problems" phenomenon [147, 148] that we mentioned in Chapter 1. One consequence of representability problems is that in general, fitting too strongly to one set of input data (say structure, as is often done for other RNA models) will often lead to larger errors in other quantities, say thermodynamics. We tried to compromise between different requirements for oxRNA. However, in order to make further progress, more detailed fitting may not be enough. Instead a more complex and most likely less tractable representation of the full interactions may need to be chosen. For example, for RNA it remains to be seen whether our single nucleotide-level model can be easily extended to generate a better representation of both structure and thermodynamics, or whether, say, a more complex model is needed to achieve the next level of accuracy. Clearly, including electrostatic effects for lower salt-concentrations, or implementing tertiary structure contacts, for example non-canonical base pairing interaction (such as Hoogsteen base pairs) and hydrogen bonding between a sugar group and a base will also need an extension of the current representation.

Given the challenges and complexity of RNA modelling, it is unsurprising that oxRNA performs less well than oxDNA for the whole ensemble of motifs. However, we believe that it is a non-trivial achievement to create a model that can semi-quantitatively reproduce such a wide range of the thermodynamic data. The properties of our model have been comprehensively tested on a variety of systems and we note that it is also currently the only RNA model with reported mechanical properties, which were tested by measuring its persistence length, force-extension curve and duplex overstretching.

In Chapter 6, we further examine the versatility of oxRNA by studying a range of different applications.

Chapter 6

Examples of systems studied with a coarse-grained RNA model

In this chapter, we use the oxRNA model, developed in Chapter 5, to study the folding thermodynamics of a pseudoknot, the formation of a kissing loop complex, the structure of a hexagonal RNA nanoring, and the unzipping of a hairpin motif.

6.1 The thermodynamics of a pseudoknot

Pseudoknots are a common structural motif in RNA. If a strand is represented as a circle and base pair contacts are represented as chords, then its structure is pseudoknotted if the chords cross. Most secondary structure prediction tools cannot treat pseudoknotted structures and thus cannot be used to study systems where they are relevant, although some progress has been made in developing efficient algorithms for this task [208, 209].

Since oxRNA provides an explicitly three-dimensional representation of the RNA strands, it can be used to simulate the folding and thermodynamics of RNA structures that involve pseudoknots. In this section, we use our model to study the well known MMTV pseudoknot [210]. The sequence and the three-dimensional representation of the MMTV pseudoknot by oxRNA are shown in Fig. 6.1. The MMTV pseudoknot's thermodynamic properties were previously studied in experiment [210] as well as with another coarse-grained RNA model [136]. Moreover, the MMTV pseudoknot's

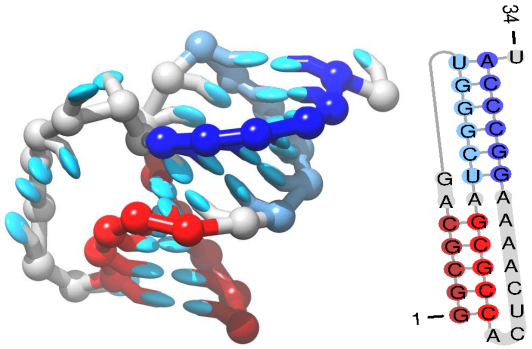

Figure 6.1: A snapshot of the MMTV pseudoknot as represented by oxRNA. Stem 1 (shown in blue) has 6 base pairs whereas stem 2 (shown in red) has 5 base pairs. A schematic representation (created with the Pseudoviewer software [207]) of the secondary structure with the sequence is shown on the right.

structure has also been investigated by NMR experiments [211] and two stems were identified in the folded structure: stem 1 with 6 base pairs and stem 2 with 5 base pairs, as schematically shown in Fig. 6.1.

To study the thermodynamics of the system, we ran VMMC simulations of oxRNA for 3.4×10^{11} steps at 75 °C. Umbrella sampling, using the number of base pairs in each of the pseudoknot stems as order parameters, was employed to enhance thermodynamic sampling. We also used histogram reweighting to extrapolate our results to other temperatures. The occupation probabilities of the unfolded state, a single hairpin with stem 1 or stem 2 (denoted as hairpin 1 and hairpin 2), and the pseudoknot are shown in Fig. 6.2(a). Our simulations also allow us to extract the heat capacity C_V from the expression

$$C_\mathrm{V} = \frac{\partial \langle U \rangle}{\partial T} \tag{6.1}$$

where we use a cubic interpolation to our simulation data for $\langle U \rangle$ in order to compute the derivative with respect to T. The results are shown in Fig. 6.2(b).

The experimentally measured C_V at 1 M [Na^+] has two peaks, one at 73.5 °C and the other at 95.0 °C [210]. It was hypothesized that the two peaks correspond to the transition from an unstructured strand to a hairpin structure and a second transition from a hairpin structure to the full pseudoknot. Analysis of our yields supports this

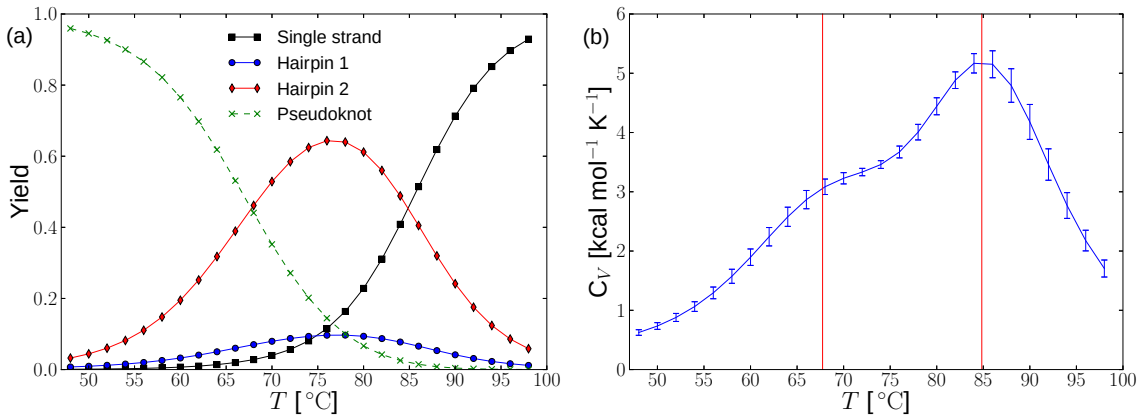

Figure 6.2: (a) Equilibrium yields and (b) C_{V} as a function of temperature for the MMTV pseudoknot. In (a) the pseudoknot and hairpins are defined as having at least 1 native base pair in the relevant stems, whereas the unstructured single-stranded state has no native base pairs. In (b) the error bars are the standard deviations derived from 5 independent simulations. The red vertical lines indicate the temperatures at which we observe equal yields of pseudoknot and hairpin 2 (67.7 °C) and hairpin 2 and the unstructured single strand (84.8 °C).

claim, showing a pseudoknot to hairpin 2 transition at 67.7 °C and transition from hairpin 2 to a single strand with no bonds in stem 1 or stem 2 at 84.8 °C. The higher temperature transition coincides with a peak in the heat capacity, whilst the lower temperature transition gives rise to a shoulder. While our simulations reproduce qualitatively the behavior of the experimental system, the position of the transitions is not exactly the same as the ones measured experimentally. This is not surprising, as we have shown in Section 5.2.2 that the model generally underestimates the melting temperatures of hairpins.

It is of further interest to analyze the free-energy landscape of the system (Fig. 6.3). Perhaps unsurprisingly, our results suggest that the minimum free-energy pathway for folding this pseudoknot from a single strand involves first forming one of the stems (forming stem 2 first is more likely as it is more stable and has a higher yield at the considered temperature) and then closing the second stem, instead of simultaneously forming both of them.

One subtlety concerns the formation of the GU base pair between the seventh and

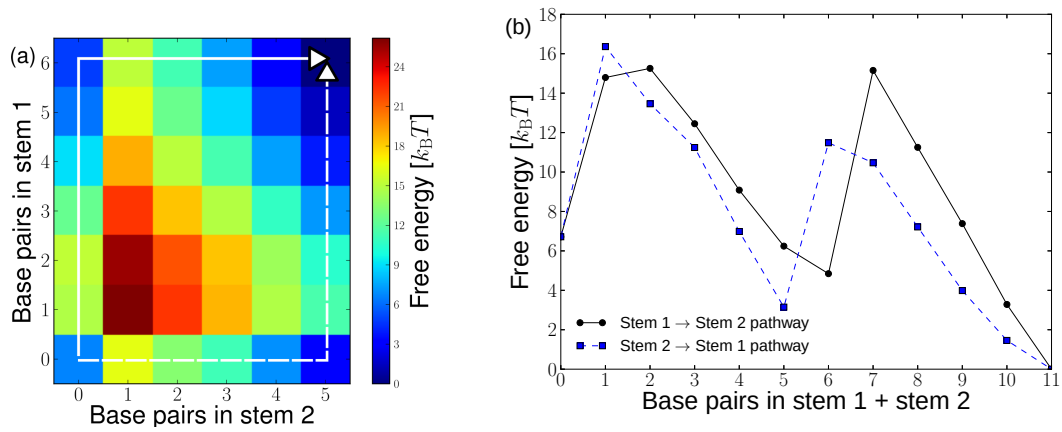

Figure 6.3: (a) A free-energy landscape for pseudoknot formation at 48 °C. White lines denote minimum free-energy pathways. (b) Free-energy profiles along the paths indicated in the free-energy landscape of (a). Dashed and solid lines correspond in both pictures. Only native base pairs contribute to the order parameters.

thirty-fourth nucleotide. The NMR study [211] did not observe the presence of this GU base pair in the pseudoknot structure. However, in our simulations, we find some structures where this base pair forms (thus extending the size of stem 1 from six to seven base pairs), although it has a $5\,k_BT$ free energy penalty at 48 °C with respect to a pseudoknot state which had only six bases in stem 1. Including this additional base pair within the definition of stem 1 had only a small effect on the calculated yields (the positions of the equal yields points changed by less than 0.3 °C) and we saw at most $0.5\,k_BT$ free-energy change for the folding pathways. We thus did not include this extra base pair in the definition of stem 1.

The experimental NMR study [211] of the structure of the MMTV pseudoknot found that the two stems of the pseudoknot are bent with respect to each other at about 112°, and the AA mismatch between the sixth and the fourteenth nucleotides to be not stacked. As can be seen in Fig. 6.1, in oxRNA, this mismatch is typically stacked leading to an angle closer to 160°. We think that this stacking of the stems reflects the overestimation of the stability of mismatches in simpler motifs (see Table 5.2).

In summary, our model is able to describe the thermodynamics of the pseudoknot

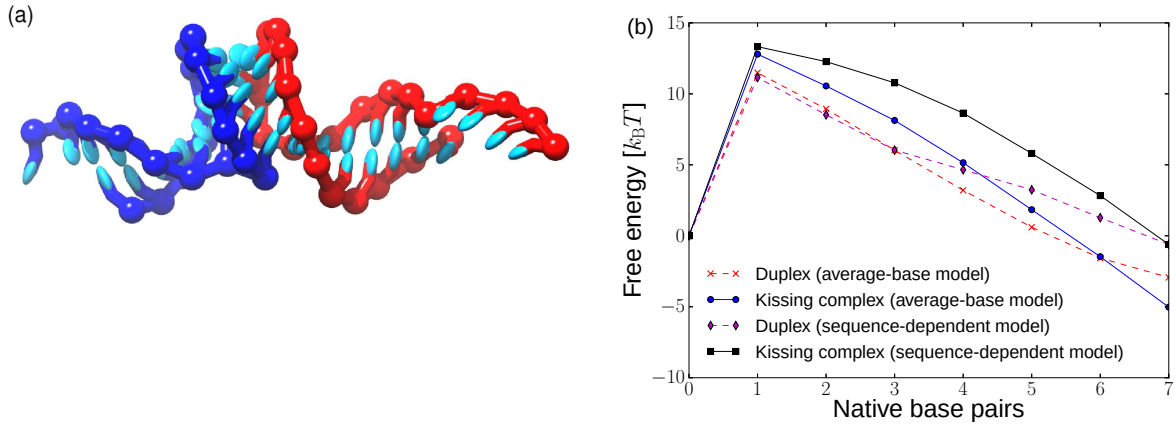

Figure 6.4: (a) A typical configuration of a kissing complex between two hairpins that have a complementary 7-base loops. (b) Free-energy profiles at 45 °C for forming the kissing complex and a 7-bp duplex with the same sequence as the hairpin loops. Results are shown for the average-base and the sequence-dependent parametrization of oxRNA.

folding and predict the stabilities of the two stems, supporting the hypothesis that the peak in heat capacity at higher temperature corresponds to the pseudoknot to hairpin 2 transition. The overall secondary structure of the pseudoknot is correct in our model, which also helps to understand the tertiary structure even though we found the angle between the two stems to be larger than the one reported from experiment.

6.2 Kissing hairpin complex

A kissing complex is a naturally occurring motif in RNA structures [15] and consists of two hairpins that have complementary loops and can thus bind to each other. An example of such a complex as represented by oxRNA is shown in Fig. 6.4(a). The kinetics and thermodynamics of forming an RNA kissing complex with 7 bases in the loops was experimentally studied in Ref. [212] at varying salt concentrations, including 1M [Na^+], the concentration at which our model was parametrized.

To examine the capability of oxRNA to describe kissing complexes, we studied the melting of this kissing complex using both the average-base and the sequence-dependent parametrization of oxRNA and found the transition at a point which is

approximately consistent with the observed experimental behavior. The 7-base loops in the two hairpins have fully complementary sequences (5′-GGAAAUG-3′ and its Watson-Crick complementary sequence). All melting simulations were run in a volume corresponding to an equal strand concentration of 3.5×10^{-4} M.

For the average-base representation we found the melting temperature of the kissing hairpins to be 62.7 °C which compared to 53.6 °C for a 7-bp duplex with the same sequences as the loops. For the sequence-dependent model, we found the melting temperature of the kissing complex to be 44.8 °C, similar to 45.2 °C for this 7-bp duplex. The free-energy profiles for both average-base and sequence-dependent models at 45 °C are shown in Fig. 6.4(b).

For most sequences, we find that the kissing complex is more stable than the separate duplexes with respect to the unbound state. We find that with increasing temperature, the kissing complex loses less stability with respect to the unbound state than a duplex at the same temperature and strand concentration. This trend can be rationalized in terms of the fact that a constrained loop loses less configurational entropy upon binding than an unconstrained single strand does. Furthermore, the kissing complex also gets an extra enthalpic stabilization from a coaxial stacking interaction between the loop and the stem nucleotides. These two effects help explain why, on average, the kissing hairpins are more stable, especially at higher temperatures. However, the kissing hairpins do not satisfy the enthalpic contributions as well as a perfectly formed duplex does. Thus, at low temperature, the duplex can be more stable. Which of these effects dominates depends on the sequence, and if the melting happens before the switch of which motif is more stable, then the duplex will have a higher melting temperature, which we find for a minority of sequences at our strand concentration. For the sequence above, the melting temperatures are very close. Note that the hairpin loops are sufficiently short that kissing complex formation is not inhibited by the topological requirement that the linking number between

the loops must remain zero. This contrasts with previous simulations of DNA kissing complexes between hairpins that have 20-base complementary loops [188].

The thermodynamics of this kissing complexes was studied in Ref. [212] using isothermal titration calorimetry (ITC) at 1 M [Na^+]. Up to 35 °C the authors found evidence of a transition to a kissing complex after the injection of the reactants, but did not observe a transition at 45 °C. The authors claim to measure only a 0.6 kcal/mol change in the ΔG for forming the kissing complex between 10 °C and 35 °C while observing a significant increase (19.5 kcal/mol) in ΔH along with a compensating increase in ΔS. Such behavior is not observed in our simulations, where we see a classic quasi-two-state transition in which ΔH and ΔS change slowly with temperature, similarly to a duplex association. If the yields of the kissing complexes in our simulations were extrapolated to the concentrations used in the experiment, we would predict a yield of 35% at 35 °C and 5% at 45 °C. We note that the thermodynamic parameters in [212] were derived with the assumption that the transition was fully saturated after the injection of the reactants in the ITC experiment, which is incompatible with the experimentally inferred value of ΔG. If the transitions were in fact not fully saturated, then it is possible that the experiments are consistent with a more typical quasi-two-state transition as observed in our model with a melting temperature approximately consistent with that found by us.

It is also interesting to use oxRNA to probe the structure of this kissing complex because it is a motif that has been used in RNA nanotechnology. Molecular dynamics simulations of the kissing complex using an all-atom representation (Amber) measured the angle between the helical stems at 300 K and 0.5 M monovalent salt to be approximately 120° [213]. Based on this finding, a hexagonal ring nanostructure that can be assembled from six RNA building blocks was proposed. Each of the proposed building blocks is an RNA strand that in the folded state has a stem and two hairpin loops. The sequences in the loops are designed to bind to the com-

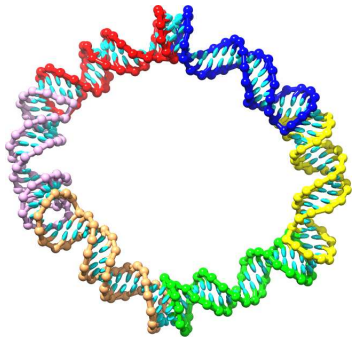

Figure 6.5: The hexagonal RNA nanoring of Ref. [213], as represented by oxRNA. The structure is composed of six strands, with a total of 264 nucleotides, connected by kissing loops.

plementary block to allow the assembly of a hexagonal "nanoring" via the kissing complex interaction. This computationally proposed RNA nanoring design was later experimentally realized by self-assembly [214]. The nanoring can be functionalized by including siRNA sequences either in the hairpin stems or by appending siRNA sequences onto the stems. Experiments in blood serum have shown that the nanoring protects the loop regions of the assembly blocks from single-strand RNA endonucleolytic cleavage, making the nanoring a promising tool for *in vivo* nanotechnology applications [214].

Simulations of oxRNA at 25 °C allowed us to measure the angle between the helical axes of the hairpin stems. We found this angle to fluctuate around an average value of 133.9° with a standard deviation of 14.8°. Hence, an octagon would probably be the most relaxed nanoring for oxRNA, and therefore favored by enthalpy. Smaller rings would be favored by translational entropy.

To illustrate the capabilities of our oxRNA model, we constructed the hexagonal RNA nanoring of Ref. [213] (Fig. 6.5) by starting a simulation with six folded hairpins blocks and introducing a harmonic potential between the complementary loop regions, which helped the kissing interactions to form. Once the nanoring was completed, the harmonic traps were removed and the assembled structure was relaxed in a molecular dynamics simulation. We found the angle between the stems of the kissing hairpins in

the nanoring to fluctuate around a mean of 124.9°. The thermal fluctuations around the mean value had a standard deviation of 14.4°, which is similar to that of the single kissing complex.

A typical relaxation time (i.e. the number of simulation steps necessary for the self-correlation function to go to zero) for the angle between adjacent kissing complexes or the energy of the assembled nanoring structure corresponds to about a minute of CPU time on a standard 2.2 GHz processor. This example shows that structural investigations of systems of the order of hundreds of bases are well within the capabilities of the oxRNA model using a single CPU, and if multiple CPUs are used, or a GPU chip is used [215] then structural properties and fluctuations around equilibrium can be studied for systems on the order of thousands of base pairs, as can be done for oxDNA [146].

6.3 Hairpin unzipping

RNA hairpin unzipping has been used in experiment to study the thermodynamics of base pairing and the mechanical properties of RNA, with the kinetics of the process also being of interest [216, 217, 218]. Unzipping the same sequence under different salt and temperature conditions can provide systematic data on the physical properties of RNA that, for example, could be used to improve the parametrization of thermodynamic models of RNA.

Here we consider the unzipping of the CD4 hairpin (shown schematically in Fig. 6.6(a)), which has a 20-bp stem and 4 bases in the loop. It was studied experimentally by pulling at different rates and measuring the unzipping force [216, 217, 218] for each trajectory. While it is possible to simulate pulling the hairpin ends at a given rate in the oxRNA model, direct comparisons with experimentally observed unzipping forces are difficult because for a coarse-grained model there is not a straightforward way to map the simulation time to the experimental one. Furthermore, to obtain

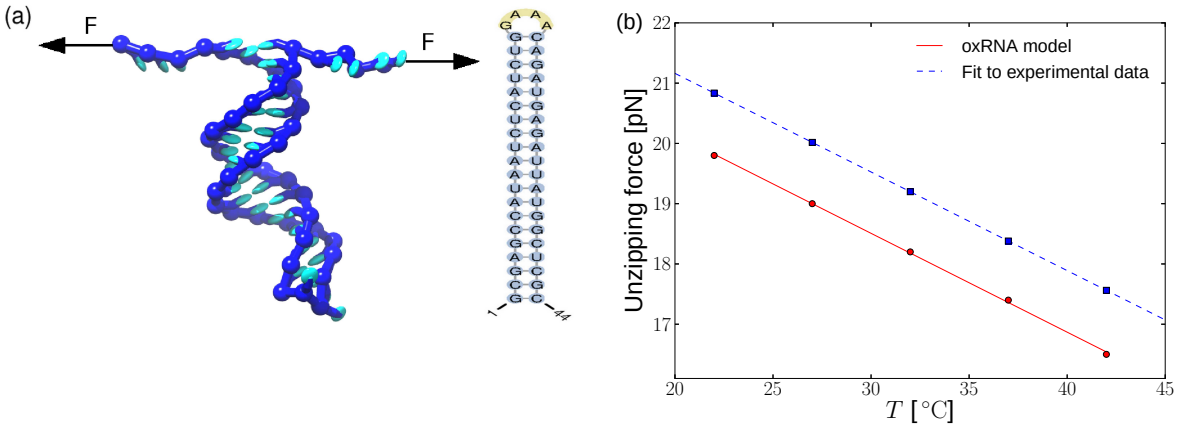

Figure 6.6: (a) A snapshot from an oxRNA simulation of unzipping of the CD4 hairpin and a schematic representation of the CD4 hairpin sequence created with the PseudoViewer software [207]. (b) The unzipping forces for the CD4 hairpin as a function of temperature for the oxRNA model (full circles) along with a linear fit to the data. For comparison, the fit to the experimentally measured unzipping force at 1M $[Na^+]$ at temperatures 22, 27, 32, 37 and 42 °C is also shown.

an estimate of the average unzipping force for a given pulling rate, one needs to average over multiple trajectories, and generating such trajectories can be quite time consuming, especially for very slow pulling rates.

A more suitable experimental setup for comparison with a coarse-grained simulation comes from Ref. [218], where the authors performed force-clamp experiments at different temperatures and salt concentrations. In the experimental setup, they kept the forces applied to the first and last base of the hairpin constant and measured the folding and unfolding rate of the hairpin, from which they inferred the free-energy difference between the open and closed state. By interpolating results for a range of applied forces, they were able to obtain the unzipping force, which was defined as the force for which the free-energy difference was zero.

To compare to the experimental results, we ran VMMC simulations of the CD4 hairpin with a constant force of 19.7 pN applied to the first and last nucleotide at 23 °C. We then extrapolated the free energies of the open and closed state by the histogram reweighting method to forces ranging from 16.2 to 20.7 pN and to the tem-

peratures at which the experiments were carried out (22, 27, 32, 37 and 42 °C). We only allowed bonds between the native base pairs to avoid sampling of metastable secondary structures that would slow down our simulations. We considered the hairpin to be closed if at least one of the bonds in the stem was present. For each temperature considered, we performed a linear interpolation of the free-energy difference between closed and open state as a function of force to obtain the unzipping force for which the difference is 0. The unzipping forces we obtain are shown in Fig. 6.6(b), along with a fit. We also show for comparison the fit to the experimentally observed unzipping forces [218] at 1 M [Na^+], expressed in the form

$$F_{\text{unzip}}(T) = a - cT, \tag{6.2}$$

where F_{unzip} is the unzipping force at temperature T. The values obtained in the experiment [218] were $a = 69.1\,\text{pN}$ and $c = -0.164\,\text{pN/K}$. Fitting Eq. 6.2 to our simulation data, we obtained the same value for c and 68.2 pN for a. The values of the fitting parameters varied by at most 6% between the fits to unzipping forces obtained from three independent sets of generated states. Thus, oxRNA is able to reproduce the unzipping force with 5% (1 pN) accuracy and fully captures the trend with temperature.

6.4 Summary and possible further applications

In this chapter, we provided some applications of the oxRNA model, introduced in Chapter 5, to illustrate its strength and potential utility. In particular, we investigated the thermodynamics of pseudoknot folding and the thermodynamics and structure of a kissing hairpin complex. We also showed that oxRNA can be used to model large nanostructures like an RNA nanoring composed of 264 nucleotides on a single CPU. The computational cost of oxRNA is very similar to oxDNA where simulations of a DNA origami motif with 12 391 nucleotides (shown in Fig. 1.3(b)) have been achieved

[146]. Finally, our model is able to reproduce experimental results for the mechanical unzipping of a hairpin quite closely, to within an accuracy of 1 pN, illustrating oxRNA's potential to study mechanical properties of RNA constructs.

Although we did not describe applications with detailed dynamics (the simulations are typically quite involved and so beyond the scope of this work), we want to emphasize that oxRNA is particularly well-suited for such tasks. For example, oxDNA has been used to study the detailed dynamics of hybridization [219], toehold-mediated strand displacement [196] and hairpin formation [146]. Studying similar processes would be very feasible for oxRNA. For example, it should be possible to use the model to obtain estimates of the rates of a strand displacement reactions as a function of length of the toehold as well as temperature. OxRNA can further be used to study the self-assembly and mechanical properties of nanostructures such as the RNA nanoring and to investigate both the thermodynamics and the kinetics of such systems.

Further possible interesting applications of oxRNA could include the study of the force-extension properties of a hairpin which contains various secondary structure motifs such as bulges and internal loops or which has regions with variable sequence strengths (such as AU-rich and GC-rich regions). Studying force-extension response of large folded RNA structures, such as ribosomes, would also be possible.

Although the model is currently only parametrized at high salt concentration, oxRNA can be also used to qualitatively study processes of biological relevance, for instance, the folding pathways of RNA strands or *in vivo* applications of nanotechnology.

Chapter 7

Conclusions and outlook

In this thesis we introduced sequence-dependent interactions into the coarse-grained DNA model of Ouldridge, Doye and Louis [149, 150, 140], called oxDNA, and tested its performance by studying multiple DNA systems with sequence-dependent phenomena: the heterogeneous stacking transition in single-stranded DNA, the fraying of a duplex, the decrease in the melting temperature of a hairpin with a long loop as the stacking strength between the bases in the loop is increased, the force-extension curves of homogeneous DNA single strands with strong and weak stacking between bases, and the structure of a kissing loop complex.

We then used the coarse-grained model of DNA to study an active DNA nanodevice, the burnt-bridges motor. We explored the free energy profiles of the DNA motor strand stepping from one stator to the next one as a function of the distance between the stators, the strengths of the attachment of the stators to a surface, and the length of the toehold region of the motor. Our results provide insight into the function of this nanodevice and have implications, for example, for the design of junctions on a motor track.

Finally, we introduced a new nucleotide-level coarse-grained model of RNA, called oxRNA, which aims to reproduce structural, mechanical, and thermodynamic properties of RNA. We used the parametrization methods developed for introducing sequence-dependence into oxDNA to parametrize the thermodynamics of the model.

We found that the model is able to reproduce semi-quantitatively a range of properties of RNA. We studied the melting temperatures of duplexes, hairpins and structures with bulges and internal and terminal mismatches. We also explored the force-extension properties and overstretching of RNA duplexes. We further demonstrated the use of the RNA model for the thermodynamics of folding of a pseudoknot, the association of a kissing complex, sampling of configurations of an RNA nanoring and unzipping of a hairpin at different temperatures.

While the agreement with experimental data for the oxRNA model is satisfactory for a range of applications, we note that we found the parametrization of the model to be more challenging that the parametrization of DNA, where closer agreement with the secondary structure thermodynamics as well as experimentally reported mechanical properties was found. It is possible that in order to improve the accuracy of the oxRNA model, it will be necessary to use a more detailed representation, such as two or three particles per nucleotide. It is not clear if pairwise interactions would be sufficient for such a model, or three-body potentials, for example functions of dihedral angles, would need to be introduced as well. That would likely lead to a decrease in the computational efficiency.

Although we used mainly the Virtual Move Monte Carlo algorithm for most simulations in this thesis, both oxDNA and oxRNA models are also designed for molecular dynamics simulations, which allow for efficient studies of large systems as well as for extracting relative rates for different processes. Further interesting projects using the oxRNA and oxDNA models hence include the molecular dynamics study of a toehold-mediated strand displacement with RNA strands [70] and the study of kinetics and thermodynamics of self-assembly of systems such as DNA Lego [27], DNA and RNA cages [22, 29, 69], and tiles [34, 67]. Folding pathways and force-extension properties of large DNA and RNA single-stranded structures, including those with biological relevance such as ribosomes, can also be of interest.

Apart from applying the models to study a variety of interesting systems, there are several possible ways of extending the oxDNA and oxRNA models further, which could include:

- Introducing salt dependence into the oxDNA and oxRNA models. One possibility would be to reparametrize the strengths of the base pairing and stacking interactions to reflect the destabilization of duplexes with the increasing salt concentration. However, such an approach will not be able to capture some other effects of the salt concentration on the system behavior, such as the increase of the persistence length of a duplex. A more plausible first line of approach is to use Debye-Hückel theory, with the charges placed on the backbone sites of the nucleotides in the oxDNA and oxRNA models. To reflect the fact that ions condensate around the charges on the backbone of the DNA/RNA strands, one would need to introduce an effective reduced charge into the Debye-Hückel potential, as was done for example for the RNA model of Denesyuk *et al.* [136]. We note that recently a Debye-Hückel potential has been introduced into the oxDNA model by Wang and Pettitt [220], who showed for one particular DNA sequence that the decrease in the melting temperature of the duplex reproduces the predictions of SantaLucia's model at salt concentrations ranging from 0.5 M down to 0.1 M. It would be further necessary to check melting temperatures for a range of duplex lengths and study the change in mechanical properties of duplexes and single strands to assess the accuracy of the introduced potential. It is possible that such an approach would be also successful for the oxRNA model.

- Extension of the oxDNA model to capture a sequence-dependent structure and flexibility. The flexibility and structure of a double strand as a function of DNA sequence content was studied experimentally [161], in fully-atomistic simulations [173, 174, 221, 163], and by analyzing structure of DNA-protein crystals

[160]. It would be of interest to see if it is possible to capture these effects with our coarse-grained model, for instance by introducing sequence-dependent equilibrium positions and widths of stacking and hydrogen bonding interactions. Refs. [173, 221] have used simulations with the Amber force-field of a large set of different sequences to parametrize an elastic model for DNA. As a first step towards sequence-dependent elasticity, we could use a similar dataset to parametrize the positions and widths or our oxDNA interactions. Alternatively, data from DNA crystals could also be used. But however it is achieved, this project will likely need a substantial amount of work.

- Development of a hybrid RNA/DNA model. As outlined in Chapter 1, DNA/RNA hybrids are increasingly popular constructs in nanotechnology and hence a hybrid coarse-grained model would present a computationally efficient tool to study such systems. Similarly to the nearest-neighbor models for DNA and RNA, there is a set of parameters available for the nearest-neighbor model for RNA/DNA duplexes [72], which could be used for parametrization using the techniques developed in Chapter 2. The potentials would need to capture the behavior of RNA and DNA systems and at the same time also correctly reproduce the thermodynamics and structure of a hybrid duplex. At present, it is not clear whether introducing hybrid RNA/DNA base pairing and cross-stacking potentials into oxDNA and oxRNA would be sufficient to achieve such a model.

- Further improvements of oxRNA could include the introduction of non-canonical interactions that stabilize the tertiary structure, such as Hoogsteen or sugar-edge interactions and ribose zippers. While the thermodynamics of Watson-Crick and wobble base pairing was carefully studied for the parametrization of the nearest-neighbor model of RNA, there is not such a set of data available for tertiary structure interactions. Knowledge-based parametrization might hence

be necessary, using the database of known RNA structures and fine-tuning the parameters of the interactions in order to obtain the known folded structure of a particular sequence as the free-energy minimum for our model as well.

In summary, we have seen that our top-down modelling approach can be successfully applied to derive coarse-grained models of both DNA and RNA that can describe a wide range of phenomena for both natural and artificial systems. OxDNA and oxRNA currently present computationally efficient models for simulations of systems composed of up to thousands of nucleotides, while still accurately representing physical properties of double- and single-stranded states. For models at this level of complexity, it is our contention that they are the most versatile and carefully tested systems on the market. In addition, our automated parametrization method can also be in principle applied to other coarse-grained models. Given the success of top-down approach for nucleic acids, one might be also interested in applying such a method to study some protein systems, or DNA/RNA-protein complexes.

Nucleic acid nanotechnology is moving rapidly forward, with structures of increasing size and complexity being developed. These anticipated advances will present more data to compare to for our models. At the same time, the reliable design of such systems also creates the need for better understanding of the biophysical properties of DNA and RNA. Our coarse-grained models can hence also be used to guide or inspire such experimental efforts.

References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 2002.

[2] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids. *Nature*, 171:737–738, 1953.

[3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[4] J. S. Mattick. A new paradigm for developmental biology. *J. Exp. Biol.*, 210:1526–1547, 2007.

[5] S. Washietl, J. S. Pedersen, J. O. Korbel, C. Stocsits, A. R. Gruber, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Reiche, A. Tanzer, et al. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.*, 17:852–864, 2007.

[6] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489:101–108, 2012.

[7] N. C. Seeman. Nucleic acid junctions and lattices. *J. Theor. Biol.*, 99:237–247, 1982.

[8] P. Guo. The emerging field of RNA nanotechnology. *Nature Nanotech.*, 5:833–842, 2010.

[9] W. Saenger. *Principles of Nucleic Acid Structure*. Springer-Verlag, 1984.

[10] S. Pitchiaya and Y. Krishnan. First blueprint, now bricks: DNA as construction material on the nanoscale. *Chem. Soc. Rev.*, 35:1111–1121, 2006.

[11] S. Neidle. *Principles of Nucleic Acid Structure*. Elsevier, 2010.

[12] S. Neidle. *Oxford Handbook of Nucleic Acid Structure*. Oxford University Press, 1999.

[13] L. Jaeger and A. Chworos. The architectonics of programmable RNA and DNA nanostructures. *Curr. Opin. Struc. Biol.*, 16:531–543, 2006.

[14] E. Herrero-Galán, M. E. Fuentes-Perez, C. Carrasco, J. M. Valpuesta, J. L. Carrascosa, F. Moreno-Herrero, and J. R. Arias-Gonzalez. Mechanical Identities of RNA and DNA Double Helices Unveiled at the Single-Molecule Level. *J. Am. Chem. Soc.*, 135:122–131, 2013.

[15] D. Elliott and M. Ladomery. *Molecular Biology of RNA*. Oxford University Press, 2011.

[16] T. R. Cech and B. Bass. Biological catalysis by RNA. *Annu. Rev. Biochem.*, 55:599–629, 1986.

[17] M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, 271:1247–1254, 1996.

[18] P. Wang, S. Hyeon K., C. Tian, C. Hao, and C. Mao. RNA-DNA hybrid origami: folding of a long RNA single strand into complex nanostructures using short DNA helper strands. *Chem. Commun.*, 49:5462–5464, 2013.

[19] S. H. Ko, M. Su, C. Zhang, A. E. Ribbe, W. Jiang, and C. Mao. Synergistic self-assembly of RNA and DNA molecules. *Nat. Chem.*, 2:1050–1055, 2010.

[20] K. A. Afonin, M. Viard, A. N. Martins, S. J. Lockett, A. E. Maciag, E. O. Freed, E. Heldman, L. Jaeger, R. Blumenthal, and B. A. Shapiro. Activation of different split functionalities on re-association of RNA-DNA hybrids. *Nature Nanotech.*, 8:296–304, 2013.

[21] P. W. K. Rothemund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440:297–302, 2006.

[22] R. P. Goodman, I. A. T. Sharp, C. F. Tardin, C. M. Erben, R. M. Berry, C. F. Schmidt, and A. J. Turberfield. Rapid Chiral Assembly of Rigid DNA Building Blocks for Molecular Nanofabrication. *Science*, 310:1661–1665, 2005.

[23] S. M. Douglas, H. Dietz, T. Liedl, B. Högberg, F. Graf, and W. M. Shih. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature*, 459:414–418, 2009.

[24] S. M. Douglas, I. Bachelet, and G. M. Church. A Logic-Gated Nanorobot for Targeted Transport of Molecular Payloads. *Science*, 335:831–834, 2012.

[25] C. E. Castro, F. Kilchherr, D.-N. Kim, E. L. Shiao, T. Wauer, P. Wortmann, M. Bathe, and H. Dietz. A primer to scaffolded DNA origami. *Nat. Methods*, 8:221–229, 2011.

[26] V. Linko and H. Dietz. The enabled state of DNA nanotechnology. *Curr. Opin. Biotech.*, 24:555–561, 2013.

[27] B. Wei, M. Dai, and P. Yin. Complex shapes self-assembled from single-stranded DNA tiles. *Nature*, 485:623–626, 2012.

[28] Y. Ke, L. L. Ong, W. M. Shih, and P. Yin. Three-Dimensional Structures Self-Assembled from DNA Bricks. *Science*, 338:1177–1183, 2012.

[29] F. A. Aldaye and H. F. Sleiman. Modular Access to Structurally Switchable 3D Discrete DNA Assemblies. *J. Am. Chem. Soc.*, 129:13376–13377, 2007.

[30] H. Yan, S. H. Park, G. Finkelstein, J. H. Reif, and T. H. LaBean. DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires. *Science*, 301:1882–1884, 2003.

[31] E. Winfree, F. R. Liu, L. A. Wenzler, and N. C. Seeman. Design and self-assembly of two-dimensional DNA crystals. *Nature*, 394:539, 1998.

[32] J. Malo, J. C. Mitchell, C. Venien-Bryan, J. R. Harris, H. Wille, D. J. Sherrat, and A. J. Turberfield. Engineering a 2D Protein-DNA Crystal. *Angew. Chem. Int. Ed.*, 44:3057–3061, 2005.

[33] J. Zheng, J. J. Birktoft, Y. Chen, T. Wang, R. Sha, P. E. Constantinou, S. L. Ginell, C. Mao, and N. C. Seeman. From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal. *Nature*, 461:74, 2009.

[34] P. W. K. Rothemund, N. Papadakis, and E. Winfree. Algorithmic Self-Assembly of DNA Sierpinski Triangles. *PLoS Biol*, 2:e424, 2004.

[35] D. Y. Zhang, A. Turberfield, B. Yurke, and E. Winfree. Engineering Entropy-Driven Reactions and Networks Catlayzed by DNA. *Science*, 318:1121–1125, 2007.

[36] B. Yurke, A. J. Turberfield, A. P. Mills, F. C. Simmel, and J. Neumann. A DNA-fueled molecular machine made of DNA. *Nature*, 406:605–608, 2000.

[37] R. P. Goodman, M. Heilemann, S. Doose, C. M. Erben, A. N. Kapanidis, and A. J. Turberfield. Reconfigurable, braced, three-dimensional DNA nanostructures. *Nature Nanotech.*, 3:93–96, 2008.

[38] E. S. Andersen, M. Dong, M. M. Nielsen, K. Jahn, R. Subramani, W. Mamdouh, M. M. Golas, B. Sander, H. Stark, C. L. P. Oliveira, J. S. Pedersen, V. Birkedal, F. Besenbacher, K. V. Gothelf, and J. Kjems. Self-assembly of a nanoscale DNA box with a controllable lid. *Nature*, 459:73–76, 2009.

[39] P. K. Lo, P. Karam, F. Aldaye, C. K. McLaughlin, G. Hamblin, G. Cosa, and H. F. Sleiman. Loading and Selective Release of Cargo in DNA Nanotubes with Longitudinal Variation. *Nature Chem.*, 2:319–328, 2010.

[40] D. Han, S. Pal, Y. Liu, and H. Yan. Folding and cutting DNA into reconfigurable topological nanostructures. *Nature Nanotech.*, 5:712–717, 2010.

[41] W. B. Sherman and N. C. Seeman. A precisely controlled DNA biped walking device. *Nano Lett.*, 4:1203–1207, 2004.

[42] J.-S. Shin and N. A. Pierce. A synthetic DNA walker for molecular transport. *J. Am. Chem. Soc.*, 126:10834–10835, 2004.

[43] T. Liedl and F. C. Simmel. switching the conformation of a DNA molecule with a chemical oscillator. *Nano Lett.*, 5:1894–1898, 2005.

[44] J. Cheng, S. Sreelatha, R. Hou, A. Efremov, R. Liu, J. R. C. van der Maarel, and Z. Wang. Bipedal Nanowalker by Pure Physical Mechanisms. *Phys. Rev. Lett.*, 109:238104, 2012.

[45] P. Yin, H. Yan, X. G. Daniell, A. J. Turberfield, and J. H. Reif. A Unidirectional DNA Walker That Moves Autonomously along a Track. *Angew. Chem. Int. Ed.*, 43:4906–4911, 2004.

[46] Y. Chen, M. Wang, and C. Mao. An Autonomous DNA Nanomotor Powered by a DNA Enzyme. *Angew. Chem. Int. Ed.*, 43:3554–3557, 2004.

[47] J. Bath, S. J. Green, and A. J. Turberfield. A Free-Running DNA Motor Powered by a Nicking Enzyme. *Angew. Chem.*, 117:4432–4435, 2005.

[48] Y. Tian, Y. He, Y. Chen, P. Yin, and C. Mao. A DNAzyme That Walks Processively and Autonomously along a One-Dimensional Track. *Angew. Chem. Int. Ed.*, 44:4355–4358, 2005.

[49] S. Venkataraman, R. M. Dirks, P. W. K. Rothemund, E. Winfree, and N. A. Pierce. An autonomous polymerization motor powered by DNA hybridization. *Nature Nanotech.*, 2:490–494, 2007.

[50] S. J. Green, J. Bath, and A. J. Turberfield. Coordinated chemomechanical cycles: a mechanism for autonomous molecular motion. *Phys. Rev. Lett.*, 101:238101, 2008.

[51] J. Bath, S. J. Green, K. E. Allan, and A. J. Turberfield. Mechanism for a directional, processive and reversible DNA motor. *Small*, 5:1513–1516, 2009.

[52] T. Omabegho, R. Sha, and N. C. Seeman. A bipedal DNA brownian motor with coordinated legs. *Science*, 324:67–71, 2009.

[53] S. F. J. Wickham, M. Endo, Y. Katsuda, K. Hidaka, J. Bath, H. Sugiyama, and A. J. Turberfield. Direct observation of stepwise movement of a synthetic molecular transporter. *Nature Nanotech.*, 6:166–169, 2011.

[54] R. A. Muscat, J. Bath, and A. J. Turberfield. A Programmable Molecular Robot. *Nano Lett.*, 11:982–987, 2011.

[55] L. M. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266:1021–1024, 1994.

[56] G. Seelig, D. Soloveichik, D. Y. Zhang, and E. Winfree. Enzyme-Free Nucleic Acid Logic Circuits. *Science*, 314:1585–1588, 2006.

[57] L. Qian, E. Winfree, and J. Bruck. Neural network computation with DNA strand displacement cascades. *Nature*, 475:368–372, 2011.

[58] S. F. J. Wickham, J. Bath, Y. Katsuda, M. Endo, K. Hidaka, H. Sugiyama, and A. J. Turberfield. A DNA-based molecular motor that can navigate a network of tracks. *Nature Nanotech.*, 7:169–173, 2012.

[59] L. Di Michele and E. Eiser. Developments in understanding and controlling self assembly of DNA-functionalized colloids. *Phys. Chem. Chem. Phys.*, 15:3115–3129, 2013.

[60] T. Schmatko, B. Bozorgui, N. Geerts, D. Frenkel, E. Eiser, and W. C. K. Poon. A finite-cluster phase in λ-DNA-coated colloids. *Soft Matter*, 3:703–706, 2007.

[61] C. Knorowski, S. Burleigh, and A. Travesset. Dynamics and Statics of DNA-Programmable Nanoparticle Self-Assembly and Crystallization. *Phys. Rev. Lett.*, 106:215501, 2011.

[62] S. Biffi, R. Cerbino, F. Bomboi, E. M. Paraboschi, R. Asselta, F. Sciortino, and T. Bellini. Phase behavior and critical activated dynamics of limited-valence DNA nanostars. *Proc. Natl. Acad. Sci. USA*, 110:15633–15637, 2013.

[63] L. Rovigatti, F. Smallenburg, F. Romano, and F. Sciortino. Gels of DNA Nanostars Never Crystallize. *ACS Nano*, in press, 2014.

[64] C. Laing and T. Schlick. Computational approaches to 3D modeling of RNA. *J. Phys.: Condens. Mat.*, 22:283101, 2010.

[65] C. Laing and T. Schlick. Computational approaches to RNA structure prediction, analysis, and design. *Curr. Opin. Struc. Biol.*, 21:306–318, 2011.

[66] W. Grabow and L. Jaeger. RNA modularity for synthetic biology. *F1000prime reports*, 5, 2013.

[67] A. Chworos, I. Severcan, A. Y. Koyfman, P. Weinkam, E. Oroudjev, H. G. Hansma, and L. Jaeger. Building programmable jigsaw puzzles with RNA. *Science*, 306:2068–2072, 2004.

[68] B. Cayrol, C. Nogues, A. Dawid, I. Sagi, P. Silberzan, and H. Isambert. A Nanostructure Made of a Bacterial Noncoding RNA. *J. Am. Chem. Soc.*, 131:17270–17276, 2009.

[69] K. A. Afonin, E. Bindewald, A. J. Yaghoubian, N. Voss, E. Jacovetty, B. A. Shapiro, and L. Jaeger. In vitro assembly of cubic RNA-based scaffolds designed in silico. *Nature Nanotech.*, 5:676–682, 2010.

[70] L. M. Hochrein, M. Schwarzkopf, M. Shahgholi, P. Yin, and N. A. Pierce. Conditional Dicer Substrate Formation via Shape and Sequence Transduction with Small Conditional RNAs. *J. Am. Chem. Soc.*, 135:17322–17330, 2013.

[71] S. Guo, N. Tschammer, S. Mohammed, and P. Guo. Specific delivery of therapeutic RNAs to cancer cells via the dimerization mechanism of phi29 motor pRNA. *Hum. Gene Ther.*, 16:1097–1110, 2005.

[72] N. Sugimoto, S.-i. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, T. Ohmichi, M. Yoneyama, and M. Sasaki. Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. *Biochemistry*, 34:11211–11216, 1995.

[73] N. C. Horton and B. C. Finzel. The Structure of an RNA/DNA Hybrid: A Substrate of the Ribonuclease Activity of HIV-1 Reverse Transcriptase. *J. Mol. Biol.*, 264:521–533, 1996.

[74] J. SantaLucia, Jr. and D. Hicks. The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.*, 33:415–40, 2004.

[75] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, 29:65–88, 2007.

[76] T. Liedl, B. Högberg, J. Tytell, D. E. Ingbe, and W. M. Shih. Self-assembly of three-dimensional prestressed tensegrity structures from DNA. *Nature Nanotech.*, 5:520–524, 2010.

[77] J. SantaLucia, Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, 17:1460–5, 1998.

[78] M. J. Serra and D. H. Turner. Predicting thermodynamic properties of RNA. *Method. Enzymol.*, 259:242–261, 1995.

[79] T. Xia, J. SantaLucia, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–14735, 1998.

[80] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, 101:7287–7292, 2004.

[81] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[82] A. E. Walter and D. H. Turner. Sequence dependence of stability for coaxial stacking of RNA helixes with Watson-Crick base paired interfaces. *Biochemistry*, 33:12715–12719, 1994.

[83] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.

[84] Z. J. Lu, D. H. Turner, and D. H. Mathews. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res.*, 34:4912–4924, 2006.

[85] J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce. NUPACK: Analysis and design of nucleic acid systems. *J. Comput. Chem.*, 32:170–173, 2011.

[86] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[87] J. S. Reuter and D. H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11:129, 2010.

[88] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31:3406–3415, 2003.

[89] I. L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31:3429–3431, 2003.

[90] N. R. Markham and M. Zuker. UNAFold. *Methods. Mol. Bio.*, 453:3–31, 2008.

[91] N. R. Markham and M. Zuker. DINAMelt web server for nucleic acid meting prediction. *Nucleic Acids Res.*, 33:W577–W581, 2005.

[92] N. L. Novere. MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics*, 17:1226–1227, 2001.

[93] C. Flamm, W. Fontana, I. L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

[94] A. Xayaphoummine, T. Bucher, and H. Isambert. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.*, 33:W605–W610, 2005.

[95] J. Šponer, K. E. Riley, and P. Hobza. Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys. Chem. Chem. Phys.*, 10:2595–2610, 2008.

[96] A. Pérez, A. Noy, F. Lankaš, F. J. Luque, and M. Orozco. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res.*, 32:6144–6151, 2004.

[97] P. Hobza and J. Šponer. Structure, Energetics, and Dynamics of the Nucleic Acid Base Pairs: Nonempirical Ab Initio Calculations. *Chem. Rev.*, 99:3247–3276, 1999.

[98] J. Šponer, P. Jurečka, I. Marchan, F. J. Luque, M. Orozco, and P. Hobza. Nature of Base Stacking: Reference Quantum-Chemical Stacking Energies in Ten Unique B-DNA Base-Pair Steps. *Chem. Eur. J.*, 12:2854–2865, 2006.

[99] D. Svozil, P. Hobza, and J. Šponer. Comparison of Intrinsic Stacking Energies of Ten Unique Dinucleotide Steps in A-RNA and B-DNA Duplexes. Can We Determine Correct Order of Stability by Quantum-Chemical Calculations? *J. Phys. Chem. B*, 114:1191–1203, 2010.

[100] J. Šponer, P. Jurečka, and P. Hobza. Accurate Interaction Energies of Hydrogen-Bonded Nucleic Acid Base Pairs. *J. Am. Chem. Soc.*, 126:10142–10151, 2004.

[101] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[102] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[103] A. Pérez, F. J. Luque, and M. Orozco. Frontiers in Molecular Dynamics Simulations of DNA. *Acc. Chem. Res.*, 45:196–205, 2012.

[104] M. Krepl, M. Zgarbov, P. Stadlbauer, M. Otyepka, P. Ban, J. Koa, T. E. Cheatham, P. Jureka, and J. poner. Reference Simulations of Noncanonical Nucleic Acids with Different Variants of the AMBER Force Field: Quadruplex DNA, Quadruplex RNA, and Z-DNA. *J. Chem. Theory Comput.*, 8:2506–2520, 2012.

[105] J. Yoo and A. Aksimentiev. In situ structure and dynamics of DNA origami determined through molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*, 110:20099–20104, 2013.

[106] A. T. Guy, T. J. Piggot, and S. Khalid. Single-stranded DNA within nanopores: conformational dynamics and implications for sequencing; a molecular dynamics simulation study. *Biophys. J.*, 103:1028–1036, 2012.

[107] P. Minary, M. E. Tuckerman, and G. J. Martyna. Dynamical spatial warping: A novel method for the conformational sampling of biophysical structure. *SIAM J. Sci. Comput.*, 30:2055–2083, 2008.

[108] A. Y. Sim, M. Levitt, and P. Minary. Modeling and design by hierarchical natural moves. *Proc. Natl. Acad. Sci. USA*, 109:2890–2895, 2012.

[109] K. Drukker, G. Wu, and G. C. Schatz. Model Simulations of DNA Denaturation Dynamics. *J. Chem. Phys.*, 114:579–590, 2001.

[110] M. Sales-Pardo, R. Guimera, A. A. Moreira, J. Widom, and L. Amaral. Mesoscopic Modelling fo Nucleic Acid Chain Dynamics. *Phys. Rev. E*, 71:051902, 2005.

[111] M. Kenward and K. D. Dorfman. Brownian Dynamics simulations of single-stranded DNA hairpins. *J. Chem. Phys.*, 130:095101, 2009.

[112] T. E. Ouldridge, I. G. Johnston, A. A. Louis, and J. P. K. Doye. The self-assembly of DNA Holliday junctions studied with a minimal model. *J. Chem. Phys.*, 130:065101, 2009.

[113] T. A. Knotts, IV, N. Rathore, D. Schwartz, and J. J. de Pablo. A Coarse Grain Model for DNA. *J. Chem. Phys.*, 126:084901, 2007.

[114] E. J. Sambriski, D. C. Schwartz, and J. J. de Pablo. A mesoscale model of DNA and its renaturation. *Biophys. J.*, 96:1675–1690, 2009.

[115] M. C. Linak, R. Tourdot, and K. D. Dorfman. Moving beyond Watson–Crick models of coarse grained DNA dynamics. *J. Chem. Phys.*, 135:205102, 2011.

[116] J. C. Araque, A. Z. Panagiotopoulos, and M. A. Robert. Lattice model of oligonucleotide hybridization in solution. I. Model and thermodynamics. *J. Chem. Phys.*, 134:165103, 2011.

[117] A.-M. Florescu and M. Joyeux. Thermal and mechanical denaturation properties of a DNA model with three sites per nucleotide. *J. Chem. Phys.*, 135:085105, 2011.

[118] A. Morriss-Andrews, J. Rottler, and S. S. Plotkin. A systematically coarse-grained model for DNA and its predictions for persistence length, stacking, twist and chirality. *J. Chem. Phys.*, 132:035105, 2010.

[119] A. V. Savin, M. A. Mazo, I. P. Kikot, L. I. Manevitch, and A. V. Onufriev. Heat conductivity of the DNA double helix. *Phys. Rev. B*, 83:245406, 2011.

[120] P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano. A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *J. Chem. Theory Comput.*, 6:1711–1725, 2010.

[121] A. Savelyev and G. A. Papoian. Molecular Renormalization Group Coarse-Graining of Polymer Chains: Application to Double-Stranded DNA. *Biophys. J.*, 96:4044–4052, 2009.

[122] N. B. Becker and R. Everaers. From rigid base pairs to semiflexible polymers: Coarse-graining DNA. *Phys. Rev. E*, 76:021923, 2007.

[123] F. Lankaš. *Innovations in Biomolecular Modeling and Simulations*, volume 2 of *RSC Biomolecular Sciences*. The Royal Society of Chemistry, 2012.

[124] D. M. Hinckley, G. S. Freeman, J. K. Whitmer, and J. J. de Pablo. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *J. Chem. Phys.*, 139:144903, 2013.

[125] M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. B. Altman. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, 15:189–199, 2009.

[126] Z. Xia, D. R. Bell, Y. Shi, and P. Ren. RNA 3D Structure Prediction by Using a Coarse-Grained Model and Experimental Data. *J. Phys. Chem. B*, 117:3135–3144, 2013.

[127] O. Taxilaga-Zetina, P. Pliego-Pastrana, and M. D. Carbajal-Tinoco. RNA pseudo-knots simulated with a one-bead coarse-grained model. *J. Chem. Phys.*, 140:115106, 2014.

[128] A. M. Mustoe, H. M. Al-Hashimi, and C. L. Brooks III. Coarse Grained Models Reveal Essential Contributions of Topological Constraints to the Conformational Free Energy of RNA Bulges. *J. Phys. Chem. B*, 118:2615–2627, 2014.

[129] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452:51–55, 2008.

[130] R. Das, J. Karanicolas, and D. Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, 7:291–294, 2010.

[131] M. Paliy, R. Melnik, and B. A. Shapiro. Coarse-graining RNA nanostructures for molecular dynamics simulations. *Physical Biology*, 7:036001, 2010.

[132] S. Pasquali and P. Derreumaux. HiRE-RNA: A High Resolution Coarse-Grained Energy Model for RNA. *J. Phys. Chem. B*, 114:11957–11966, 2010.

[133] T. Cragnolini, P. Derreumaux, and S. Pasquali. Coarse-Grained Simulations of RNA and DNA Duplexes. *J. Phys. Chem. B*, 117:8047–8060, 2013.

[134] F. Ding, S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude, and N. V. Dokholyan. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, 14:1164–1173, 2008.

[135] C. Hyeon and D. Thirumalai. Mechanical Unfolding of RNA hairpins. *Proc. Natl. Acad. Sci. USA*, 102:6789–6794, 2005.

[136] N. A. Denesyuk and D. Thirumalai. Coarse-Grained Model for Predicting RNA Folding Thermodynamics. *J. Phys. Chem. B*, 117:4901–4911, 2013.

[137] C. Hyeon, R. I. Dima, and D. Thirumalai. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure*, 14:1633–1645, 2006.

[138] S. Cao and S.-J. Chen. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*, 11:1884–1897, 2005.

[139] D. Jost and R. Everaers. Prediction of RNA multiloop and pseudoknot conformations from a lattice-based, coarse-grain tertiary structure model. *J. Chem. Phys.*, 132:095101–095101, 2010.

[140] T. E. Ouldridge. *Coarse-grained modelling of DNA and DNA nanotechnology.* PhD thesis, Oxford University, 2011.

[141] T. Dauxois, M. Peyrard, and A. R. Bishop. Dynamics and thermodynamics of a nonlinear model for DNA denaturation. *Phys. Rev. E*, 47:684–695, 1993.

[142] C. Nisoli and A. R. Bishop. Thermomechanics of DNA: Theory of Thermal Stability under Load. *Phys. Rev. Lett.*, 107:068102, 2011.

[143] S. Cocco and R. Monasson. Statistical Mechanics of Torque Induced Denaturation of DNA. *Phys. Rev. Lett.*, 83:5178–5181, 1999.

[144] G. Weber. Mesoscopic model parametrization of hydrogen bonds and stacking interactions of RNA from melting temperatures. *Nucleic Acids Res.*, 41:e30–e30, 2013.

[145] S. Khalid, P. J. Bond, J. Holyoake, R. W. Hawtin, and M. S. Sansom. DNA and lipid bilayers: self-assembly and insertion. *J.Roy. Soc. Interface*, 5:241–250, 2008.

[146] J. P. K. Doye, T. E. Ouldridge, A. A. Louis, F. Romano, P. Šulc, C. Matek, B. E. K. Snodin, L. Rovigatti, J. S. Schreck, R. M. Harrison, and W. P. J. Smith. Coarse-graining DNA for simulations of DNA nanotechnology. *Phys. Chem. Chem. Phys.*, 15:20395–20414, 2013.

[147] A. A. Louis. Beware of density dependent pair potentials. *J. Phys.: Condens. Matter*, 14:9187, 2002.

[148] A. A. Louis. No free lunch for effective potentials: general comment for Faraday FD144. *arXiv preprint arXiv:1001.1097*, 2010.

[149] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye. DNA nanotweezers studied with a coarse-grained model of DNA. *Phys. Rev. Lett.*, 104:178101, 2010.

[150] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *J. Chem. Phys.*, 134:085101, 2011.

[151] P. Šulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis. Sequence-dependent thermodynamics of a coarse-grained DNA model. *J. Chem. Phys.*, 137:135101, 2012.

[152] P. Šulc, T. E. Ouldridge, F. Romano, J. P. K. Doye, and A. A. Louis. Simulating a burnt-bridges DNA motor with a coarse-grained DNA model. *Natural Computing*, 2013.

[153] P. Šulc, F. Romano, T. E. Ouldridge, J. P. K. Doye, and A. A. Louis. A nucleotide-level coarse-grained model of RNA. *arXiv preprint arXiv:1403.4180*, 2014.

[154] C. Calladine and H. Drew. *Understanding DNA: the molecule & how it works*. Academic Press, 1997.

[155] P. J. Hagerman. Flexibility of DNA. *Annu. Rev. Biophys. Biophys. Chem.*, 17:265–286, 1988.

[156] T. Odijk. Stiff Chains and Filaments under Tension. *Macromolecules*, 28:7016–7018, 1995.

[157] M. D. Wang, H. Yin, R. Landick, J. Gelles, and S. M. Block. Stretching DNA with optical tweezers. *Biophys. J.*, 72:1335–1346, 1997.

[158] J. R. Wenner, M. C. Williams, I. Rouzina, and V. A. Bloomfield. Salt Dependence of the Elasticity and Overstretching Transition of single DNA Molecules. *Biophys. J.*, 82:3160–3169, 2002.

[159] F. Romano, D. Chakraborty, J. P. K. Doye, T. E. Ouldridge, and A. A. Louis. Coarse-grained simulations of DNA overstretching. *J. Chem. Phys.*, 138:085101, 2013.

[160] W. K. Olson, A. A. Gorin, X.-J. Lu, L. M. Hock, and V. B. Zhurkin. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, 95:11163–11168, 1998.

[161] S. Geggier and A. Vologodskii. Sequence dependence of DNA bending rigidity. *Proc. Natl. Acad. Sci. USA*, 107:15421–15426, 2010.

[162] B. Basham, G. P. Schroth, and P. S. Ho. An A-DNA triplet code: thermodynamic rules for predicting A- and B-DNA. *Proc. Natl. Acad. Sci. USA*, 92:6464–6468, 1995.

[163] A. Noy and R. Golestanian. The chirality of DNA: elasticity cross-terms at base-pair level including A-tracts and the influence of ionic strength. *J. Phys. Chem. B*, 114:8022–8031, 2010.

[164] V. Ortiz and J. J. de Pablo. Molecular Origins of DNA Flexibility: Sequence Effects on Conformational and Mechanical Properties. *Phys. Rev. Lett.*, 106:238107, 2011.

[165] J. M. Huguet, C. V. Bizarro, N. Forns, S. B. Smith, C. Bustamante, and F. Ritort. Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proc. Natl. Acad. Sci. USA*, 107:15431–15436, 2010.

[166] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye. Extracting bulk properties of self-assembling systems from small simulations. *J. Phys.: Condens. Matter*, 22:104102, 2010.

[167] T. E. Ouldridge. Inferring bulk self-assembly properties from simulations of small systems with multiple constituent species and small systems in the grand canonical ensemble. *J. Chem. Phys.*, 137:144105, 2012.

[168] D. H. De Jong, L. V. Schäfer, A. H. De Vries, S. J. Marrink, H. J. Berendsen, and H. Grubmüller. Determining equilibrium constants for dimerization reactions from molecular dynamics simulations. *J. Comput. Chem.*, 32:1919–1928, 2011.

[169] A. M. Ferrenberg and R. H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.*, 61:2635–2638, 1988.

[170] D. Landau and K. Binder. *A Guide to Monte Carlo Simulations in Statistical Physics*. Cambridge University Press, 2005.

[171] G. Torrie and J. P. Valleau. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comp. Phys.*, 23:187–199, 1977.

[172] W.-S. Chen, W.-H. Chen, Z. Chen, A. A. Gooding, K.-J. Lin, and C.-H. Kiang. Direct Observation of Multiple Pathways of Single-Stranded DNA Stretching. *Phys. Rev. Lett.*, 105:218104, 2010.

[173] F. Lankaš, O. Gonzalez, L. M. Heffler, G. Stoll, M. Moakher, and J. H. Maddocks. On the parameterization of rigid basepair models of DNA from molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, 11:10565–10588, 2009.

[174] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, I. Cheatham, Thomas, S. Dixit, B. Jayaram, F. Lankaš, C. Laughton, J. H. Maddocks, A. Michon, R. Osman, M. Orozco, A. Perez, T. Singh, N. Špacková, and J. Šponer. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.*, 38:299–313, 2010.

[175] D. Norouzi, F. Mohammad-Rafiee, and R. Golestanian. Effect of Bending Anisotropy on the 3D Conformation of Short DNA Loops. *Phys. Rev. Lett.*, 101:168103, 2008.

[176] A. Noy and R. Golestanian. Length Scale Dependence of DNA Mechanical Properties. *Phys. Rev. Lett.*, 109:228101, 2012.

[177] J. Holbrook, M. Capp, R. Saecker, and M. Record. Enthalpy and Heat Capacity Changes for Formation of an Oligomeric DNA Duplex: Interpretation in Terms of Coupled Processes of Formation and Association of Single-Stranded Helices. *Biochemistry*, 38:8409–8422, 1999.

[178] S. Nonin, J.-L. Leroy, and M. Gueron. Terminal Base Pairs of Oligodeoxynucleotides: Imino Proton Exchange and Fraying. *Biochemistry*, 34:10652–10659, 1995.

[179] D. Y. Zhang and E. Winfree. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *J. Am. Chem. Soc.*, 131:17303–17314, 2009.

[180] N. L. Goddard, G. Bonnet, O. Krichevsky, and A. Libchaber. Sequence Dependent Rigidity of Single Stranded DNA. *Phys. Rev. Lett.*, 85:2400–2403, 2000.

[181] S. B. Smith, Y. Cui, and C. Bustamante. Overstretching B-DNA: the elastic response of individual double-stranded and sinle-stranded DNA molecules. *Science*, 271:795–799, 1996.

[182] G. Mishra, D. Giri, and S. Kumar. Stretching of a single-stranded DNA: Evidence for structural transition. *Phys. Rev. E*, 79:031930, 2009.

[183] Y. Seol, G. M. Skinner, K. Visscher, A. Buhot, and A. Halperin. Stretching of Homopolymeric RNA Reveals Single-Stranded Helices and Base-Stacking. *Phys. Rev. Lett.*, 98:158103, 2007.

[184] Y. Seol, G. M. Skinner, and K. Visscher. Elastic Properties of a Single-Stranded Charged Homopolymeric Ribonucleotide. *Phys. Rev. Lett.*, 93:118102, 2004.

[185] M.-N. Dessinges, B. Maier, Y. Zhang, M. Peliti, D. Bensimon, and V. Croquette. Stretching Single Stranded DNA, a Model Polyelectrolyte. *Phys. Rev. Lett.*, 89:248102, 2002.

[186] Y. Zhang, H. Zhou, and Z.-C. Ou-Yang. Stretching Single-Stranded DNA: Interplay of Electrostatic, Base-Pairing, and Base-Pair Stacking Interactions. *Biophys. J.*, 81:1133–1143, 2001.

[187] A. Montanari and M. Mézard. Hairpin Formation and Elongation of Biomolecules. *Phys. Rev. Lett.*, 86:2178–2181, 2001.

[188] F. Romano, A. Hudson, J. P. K. Doye, T. E. Ouldridge, and A. A. Louis. The effect of topology on the structure and free energy landscape of DNA kissing complexes. *J. Chem. Phys.*, 136:215102, 2012.

[189] J. Bois, S. Venkataraman, H. M. T. Choi, A. J. Spakowitz, Z. Wang, and N. A. Pierce. Topological constraints in nucleic acid hybridization kinetics. *Nucleic Acids Res.*, 33:4090–4095, 2005.

[190] R. M. Dirks and N. A. Pierce. Triggered amplification by hybridization chain reaction. *Proc. Natl. Acad. Sci. USA*, 101:15275–15278, 2004.

[191] P. Yin, H. M. Choi, C. R. Calvert, and N. A. Pierce. Programming biomolecular self-assembly pathways. *Nature*, 451:318–323, 2008.

[192] S. J. Green, D. Lubrich, and A. J. Turberfield. DNA Hairpins: Fuel for Autonomous DNA Devices. *Biophys. J.*, 91:2966–2975, 2006.

[193] D. F. Heiter, K. D. Lunnen, and G. G. Wilson. Site-specific DNA-nicking Mutants of the Heterodimeric Restriction Endonuclease R.BbvCI. *J. Mol. Biol.*, 348:631–640, 2005.

[194] J. Bath and A. J. Turberfield. DNA nanomachines. *Nature Nanotech.*, 2:275–284, 2007.

[195] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13:1011–1021, 1992.

[196] N. Srinivas, T. E. Ouldridge, P. Šulc, J. M. Schaeffer, B. Yurke, A. A. Louis, J. P. K. Doye, and E. Winfree. On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Res.*, 41:10641–10658, 2013.

[197] T. E. Ouldridge, R. L. Hoare, A. A. Louis, J. P. K. Doye, J. Bath, and A. J. Turberfield. Optimizing DNA Nanotechnology through Coarse-Grained Modeling: A Two-Footed DNA Walker. *ACS Nano*, 7:2479–2490, 2013.

[198] C. Matek, T. E. Ouldridge, A. Levy, J. P. K. Doye, and A. A. Louis. DNA Cruciform Arms Nucleate through a Correlated but Asynchronous Cooperative Mechanism. *J. Phys. Chem. B*, 116:11616–11625, 2012.

[199] S. Bellaousov, J. S. Reuter, M. G. Seetin, and D. H. Mathews. RNAstructure: web servers for RNA secondary structure prediction and analysis. *Nucleic Acids Res.*, 41:W471–W474, 2013.

[200] J. Russo, P. Tartaglia, and F. Sciortino. Reversible gels of patchy particles: Role of the valence. *J. Chem. Phys.*, 131:014504, 2009.

[201] D. V. Pyshnyi and E. M. Ivanova. The influence of nearest neighbours on the efficiency of coaxial stacking at contiguous stacking hybridization of oligodeoxyribonucleotides. *Nucleosides Nucleotides Nucleic Acids*, 23:1057–1064, 2004.

[202] P. Kebbekus, D. E. Draper, and P. Hagerman. Persistence length of RNA. *Biochemistry*, 34:4354–4357, 1995.

[203] R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, and K. Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, 37:5917–5929, 2009.

[204] J. Abels, F. Moreno-Herrero, T. Van der Heijden, C. Dekker, and N. Dekker. Single-molecule measurements of the persistence length of double-stranded RNA. *Biophys. J.*, 88:2737–2744, 2005.

[205] P. J. Hagerman. Flexibility of RNA. *Annu. Rev. Biophys. Biomol. Struct.*, 26:139–156, 1997.

[206] S. F. Edwards and M. Doi. *The theory of polymer dynamics*. Oxford University Press, 1986.

[207] Y. Byun and K. Han. PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res.*, 34:W416–W422, 2006.

[208] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[209] M. Bon and H. Orland. TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res.*, 39:e93–e93, 2011.

[210] C. A. Theimer and D. P. Giedroc. Contribution of the intercalated adenosine at the helical junction to the stability of the gag-pro frameshifting pseudoknot from mouse mammary tumor virus. *RNA*, 6:409–421, 2000.

[211] L. X. Shen and I. Tinoco Jr. The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. *J. Mol. Biol.*, 247:963–978, 1995.

[212] N. Salim, R. Lamichhane, R. Zhao, T. Banerjee, J. Philip, D. Rueda, and A. L. Feig. Thermodynamic and kinetic analysis of an RNA kissing interaction and its resolution into an extended duplex. *Biophys. J.*, 102:1097–1107, 2012.

[213] Y. G. Yingling and B. A. Shapiro. Computational Design of an RNA Hexagonal Nanoring and an RNA Nanotube. *Nano Lett.*, 7:2328–2334, 2007.

[214] W. W. Grabow, P. Zakrevsky, K. A. Afonin, A. Chworos, B. A. Shapiro, and L. Jaeger. Self-assembling RNA nanorings based on RNAI/II inverse kissing complexes. *Nano Lett.*, 11:878–887, 2011.

[215] L. Rovigatti, P. Šulc, I. Z. Reguly, and F. Romano. A comparison between parallelization approaches in molecular dynamics simulations on GPUs. *arXiv preprint arXiv:1401.4350*, 2014.

[216] M. Manosas, D. Collin, and F. Ritort. Force-Dependent Fragility in RNA Hairpins. *Phys. Rev. Lett.*, 96:218301, 2006.

[217] C. Bizarro, A. Alemany, and F. Ritort. Non-specific binding of Na+ and Mg2+ to RNA determined by force spectroscopy methods. *Nucleic Acids Res.*, 40:6922–6935, 2012.

[218] W. Stephenson, S. Keller, R. Santiago, J. E. Albrecht, P. N. Asare-Okai, S. A. Tenenbaum, M. Zuker, and P. T. Li. Combining temperature and force to study folding of an RNA hairpin. *Phys. Chem. Chem. Phys.*, 16:906–917, 2014.

[219] T. E. Ouldridge, P. Šulc, F. Romano, J. P. K. Doye, and A. A. Louis. DNA hybridization kinetics: zippering, internal displacement and sequence dependence. *Nucleic Acids Res.*, 41:8886–8895, 2013.

[220] Q. Wang and B. M. Pettitt. Modeling DNA Thermodynamics under Torsional Stress. *Biophys. J.*, 106:1182–1193, 2014.

[221] O. Gonzalez, D. Petkevičiūtė, and J. Maddocks. A sequence-dependent rigid-base model of DNA. *J. Chem. Phys.*, 138:055102, 2013.

[222] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[223] S. Whitelam, E. H. Feng, M. F. Hagan, and P. L. Geissler. The role of collective motion in examples of coarsening and self-assembly. *Soft Matter*, 5:1251–1262, 2009.

[224] D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 1996.

[225] T. Schlick. *Molecular modeling and simulation: An interdisciplinary guide*. Springer, 2010.

[226] H. C. Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, 72:2384–2393, 1980.

[227] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76:637–649, 1982.

[228] J. Lapham, J. P. Rife, P. B. Moore, and D. M. Crothers. Measurement of diffusion constants for nucleic acids by NMR. *J. Biomol. NMR*, 10:252–262, 1997.

Appendix A

Potentials and nucleotide representation in the oxRNA model

The oxRNA model and its potentials were introduced in Section 5.1 and here we provide a detailed description of the interaction potentials and the nucleotides. We first describe the representation of each nucleotide as a rigid body in Section A.1 and then give the explicit expression for each of the potential terms in Section A.2. All the values of the potential parameters are in an internal unit system of the downloadable simulation code where 1 distance unit = 8.4 Å and 1 energy unit = 41.4 pN nm = $10\,k_{\mathrm{B}}T$ for $T = 300\,\mathrm{K}$. In molecular dynamics simulations, we set 1 mass unit to correspond to the average weight of a nucleotide, 321.4 AMU, which gives us the simulation time unit corresponding to 3.06×10^{-12} s in SI units.

A.1 Representation

Each nucleotide in the oxRNA model is represented as a single rigid body with multiple interaction sites. Each nucleotide has backbone, hydrogen-bonding, cross-stacking and 3′- and 5′-stacking interaction sites. The position and orientation of each nucleotide is uniquely specified by its center of mass position and the perpendicular unit vectors $\mathbf{a}_3$ and $\mathbf{a}_1$, where $\mathbf{a}_1$ is a unit vector pointing from the center of mass to the hydrogen-bonding site and $\mathbf{a}_3$ is defined in Fig. A.1. In a duplex configuration, the

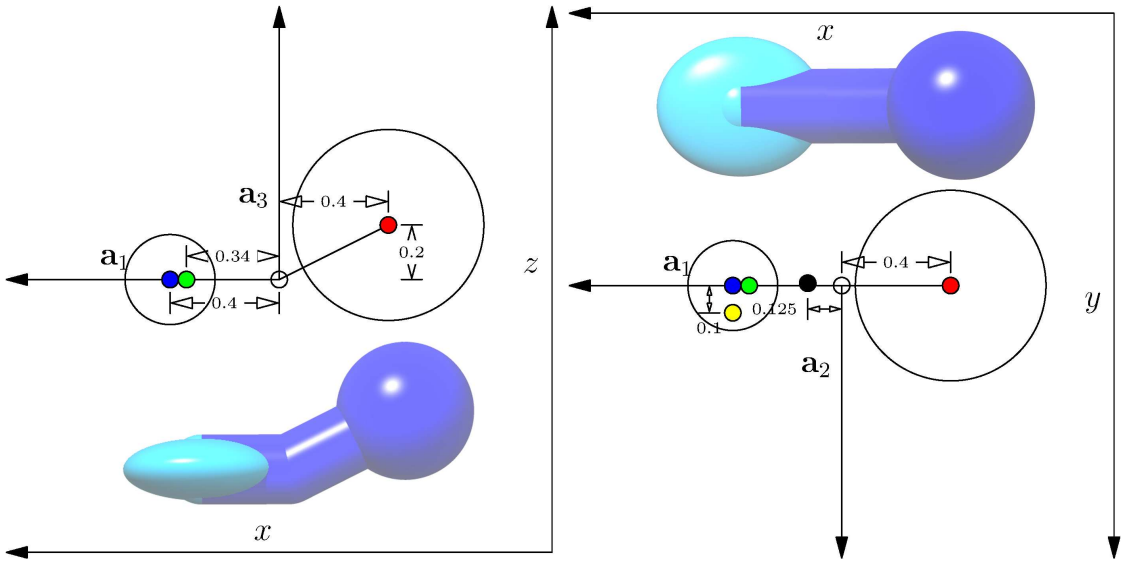

Figure A.1: A schematic representation of the nucleotides as represented by the oxRNA model. The red circle indicates the backbone site, the blue circle is the hydrogen-bonding site and the green circle is the coaxial stacking site. The yellow circle is the 3′-stacking site, and the black circle is the 5′ stacking site. The unfilled circle from which the $\mathbf{a}_3$, $\mathbf{a}_2$ and $\mathbf{a}_1$ vectors originate is the center of mass. All distances are given in a unit system where 1 distance unit = 8.4 Å. The left image shows the projection of a single nucleotide where the $\mathbf{a}_2$ vector is pointing towards the reader, and the image on the right shows a projection where the $\mathbf{a}_3$ vector is pointing towards the reader. For comparison, we also show the schematic representation of the nucleotide that is used in producing pictures of oxRNA configurations. The backbone site is represented by a sphere because of the isotropic nature of the its interactions, whereas the base is represented by an ellipsoid whose principal axes are parallel to $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ respectively.

$\mathbf{a}_3$ vector would be pointing towards the 5′-neighboring nucleotide. We further define $\mathbf{a}_2 = \mathbf{a}_3 \times \mathbf{a}_1$. The nucleotide as represented by oxRNA is schematically shown in Fig. A.1. The small colored circles indicate the position of the interaction sites, while the large circles around hydrogen bonding and backbone sites indicate the interaction radius of the excluded-volume interactions.

The interaction potentials are functions of the distances between the relevant interaction sites as well as the angles between intersite vectors and the respective

orientation vectors $\mathbf{a}_3$, $\mathbf{a}_1$, $\mathbf{p}_{3'}$ and $\mathbf{p}_{5'}$, where we define

$$\mathbf{p}_{3'} = -0.46\mathbf{a}_1 - 0.53\mathbf{a}_2 + 0.71\mathbf{a}_3 \quad \text{(A.1)}$$

$$\mathbf{p}_{5'} = -0.1\mathbf{a}_1 - 0.84\mathbf{a}_2 + 0.53\mathbf{a}_3. \quad \text{(A.2)}$$

We also define the following vectors which are then used in the definitions of the potentials in the oxRNA model (Eq. 5.2):

- $\delta\mathbf{r}_{\text{backbone}}$: the vector between backbone sites of the nucleotides. If the nucleotides are nearest neighbors, it is pointing towards the nucleotide's $3'$-neighbor's backbone site

- $\delta\mathbf{r}_{\text{HB}}$: the vector between the hydrogen-bonding sites of the interacting nucleotides, pointing from the first nucleotide towards the second one

- $\delta\mathbf{r}_{\text{coaxial st.}}$: the vector between the coaxial stacking sites of the interacting nucleotides, pointing from the first nucleotide towards the second one

- $\delta\mathbf{r}_{\text{stack}}$: the vector pointing from the $3'$-stacking site of a nucleotide to the $5'$-stacking site of its $3'$-neighbor

- $\delta\mathbf{r}_{\text{back}-\text{base}}/\delta\mathbf{r}_{\text{base}-\text{back}}$: the vector pointing from the backbone/hydrogen-bonding site of the first nucleotide to the hydrogen-bonding/backbone site of the second nucleotide

We further define the following angles that are used in the potential functions:

$$\theta_1 = \arccos\left(-\mathbf{a}_1 \cdot \mathbf{b}_1\right) \tag{A.3}$$

$$\theta_2 = \arccos\left(-\mathbf{b}_1 \cdot \widehat{\delta\mathbf{r}}_{\mathrm{HB}}\right) \tag{A.4}$$

$$\theta_3 = \arccos\left(\mathbf{a}_1 \cdot \widehat{\delta\mathbf{r}}_{\mathrm{HB}}\right) \tag{A.5}$$

$$\theta_4 = \arccos\left(\mathbf{a}_3 \cdot \mathbf{b}_3\right) \tag{A.6}$$

$$\theta_5 = \arccos\left(\mathbf{a}_3 \cdot \widehat{\delta\mathbf{r}}_{\text{coaxial st.}}\right) \tag{A.7}$$

$$\theta_6 = \arccos\left(-\mathbf{b}_3 \cdot \widehat{\delta\mathbf{r}}_{\text{coaxial st.}}\right) \tag{A.8}$$

$$\theta_{5'} = \arccos\left(\mathbf{a}_3 \cdot \widehat{\delta\mathbf{r}}_{\text{stack}}\right) \tag{A.9}$$

$$\theta_{6'} = \arccos\left(-\mathbf{b}_3 \cdot \widehat{\delta\mathbf{r}}_{\text{stack}}\right) \tag{A.10}$$

$$\theta_7 = \arccos\left(-\mathbf{b}_3 \cdot \widehat{\delta\mathbf{r}}_{\mathrm{HB}}\right) \tag{A.11}$$

$$\theta_8 = \arccos\left(\mathbf{a}_3 \cdot \widehat{\delta\mathbf{r}}_{\mathrm{HB}}\right) \tag{A.12}$$

$$\theta_9 = \arccos\left(-\mathbf{p}_{3'} \cdot \widehat{\delta\mathbf{r}}_{\text{backbone}}\right) \tag{A.13}$$

$$\theta_{10} = \arccos\left(-\mathbf{q}_{5'} \cdot \widehat{\delta\mathbf{r}}_{\text{backbone}}\right) \tag{A.14}$$

$$\cos\left(\phi_1\right) = \widehat{\delta\mathbf{r}}_{\text{backbone}} \cdot \mathbf{a}_2 \tag{A.15}$$

$$\cos\left(\phi_2\right) = \widehat{\delta\mathbf{r}}_{\text{backbone}} \cdot \mathbf{b}_2 \tag{A.16}$$

$$\cos\left(\phi_3\right) = \widehat{\delta\mathbf{r}}_{\text{coaxial st.}} \cdot \left(\widehat{\delta\mathbf{r}}_{\text{backbone}} \times \mathbf{a}_1\right) \tag{A.17}$$

$$\cos\left(\phi_4\right) = \widehat{\delta\mathbf{r}}_{\text{coaxial st.}} \cdot \left(\widehat{\delta\mathbf{r}}_{\text{backbone}} \times \mathbf{b}_1\right), \tag{A.18}$$

where we use the notation $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$ to define the orientation vectors of the second nucleotide participating in the interaction (the orientation vectors of the first nucleotide are denoted as $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$). The vector $\mathbf{q}_{5'}$ corresponds to the $\mathbf{p}_{5'}$ vector of the 3′-neighbor of the interacting nucleotide, i.e. using the same definition as in Eq. A.2, but substituting $\mathbf{b}$ for $\mathbf{a}$.

A.2 Potentials

The oxRNA potential consists of a sum of potential functions designed to represent different physical interactions, with some of the potentials being products of multiple potential functions. The functions that are used in the potentials are:

- FENE spring (used in V_{backbone}):

$$V_{\text{FENE}}(r, \epsilon, r^0, \Delta) = -\frac{\epsilon}{2} \ln \left(1 - \frac{(r - r^0)^2}{\Delta^2} \right). \tag{A.19}$$

- Morse potential (used in V_{stack} and $V_{\text{H.B.}}$):

$$V_{\text{Morse}}(r, \epsilon, r^0, d) = \epsilon \big(1 - \exp\left(-d(r - r^0)\right) \big)^2. \tag{A.20}$$

- Harmonic potential (used in $V_{\text{cross st.}}$ and $V_{\text{coaxial st.}}$):

$$V_{\text{harm}}(r, k, r^0) = \frac{k}{2} \left(r - r^0 \right)^2. \tag{A.21}$$

- Lennard-Jones potential (used in excluded volume potentials V'_{exc} and V_{exc}):

$$V_{\text{LJ}}(r, \epsilon, \sigma) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right]. \tag{A.22}$$

- Quadratic modulation terms (used for angular modulation of the anisotropic potentials $V_{\text{H.B.}}, V_{\text{cross st.}}, V_{\text{stack}}$ and $V_{\text{coaxial st.}}$):

$$V_{\text{mod}}(\theta, a, \theta^0) = 1 - a(\theta - \theta^0)^2. \tag{A.23}$$

- Quadratic smoothing terms for truncation (used in all potentials with the exception of V_{backbone}) in order to make the potentials differentiable functions that are equal to 0 beyond some specific cutoff distance:

$$V_{\text{smooth}}(x, b, x^c) = b(x^c - x)^2. \tag{A.24}$$

The smoothed functions used in the potentials have the following form:

- The radial part of the stacking and hydrogen-bonding potentials:

$$f_1(r,\epsilon,d,r^0,r^c,r^{low},r^{high}) = \begin{cases} V_{\text{Morse}}(r,\epsilon,r^0,d) - V_{\text{Morse}}(r^c,\epsilon,r^0,d)) & \text{if } r^{low} < r < r^{high}, \\ \epsilon V_{\text{smooth}}(r,b^{low},r^{c,low}) & \text{if } r^{c,low} < r < r^{low}, \\ \epsilon V_{\text{smooth}}(r,b^{high},r^{c,high}) & \text{if } r^{high} < r < r^{c,high}, \\ 0 & \text{otherwise.} \end{cases} \tag{A.25}$$

- The radial part of the cross-stacking and coaxial stacking potentials:

$$f_2(r,k,r^0,r^c,r^{low},r^{high}) = \begin{cases} V_{\text{harm}}(r,k,r^0) - V_{\text{harm}}(r^c,k,r^0) & \text{if } r^{low} < r < r^{high}, \\ kV_{\text{smooth}}(r,b^{low},r^{c,low}) & \text{if } r^{c,low} < r < r^{low}, \\ kV_{\text{smooth}}(r,b^{high},r^{c,high}) & \text{if } r^{high} < r < r^{c,high}, \\ 0 & \text{otherwise.} \end{cases} \tag{A.26}$$

- The radial part of the excluded volume potential:

$$f_3(r,\epsilon,\sigma,r^\star) = \begin{cases} V_{\text{LJ}}(r,\epsilon,\sigma) & \text{if } r < r^\star, \\ \epsilon V_{\text{smooth}}(r,b,r^c) & \text{if } r^\star < r < r^c, \\ 0 & \text{otherwise.} \end{cases} \tag{A.27}$$

- The angular modulation factor used in stacking, hydrogen-bonding, cross-stacking and coaxial stacking:

$$f_4(\theta,a,\theta^0,\Delta\theta^\star) = \begin{cases} V_{\text{mod}}(\theta,a,\theta^0) & \text{if } \theta^0 - \Delta\theta^\star < \theta < \theta^0 + \Delta\theta^\star, \\ V_{\text{smooth}}(\theta,b,\theta^0 - \Delta\theta^c) & \text{if } \theta^0 - \Delta\theta^c < \theta < \theta^0 - \Delta\theta^\star, \\ V_{\text{smooth}}(\theta,b,\theta^0 + \Delta\theta^c) & \text{if } \theta^0 + \Delta\theta^\star < \theta < \theta^0 + \Delta\theta^c, \\ 0 & \text{otherwise.} \end{cases} \tag{A.28}$$

- Another modulating term which is used to impose right-handedness:

$$f_5(x,a,x^\star) = \begin{cases} 1 & \text{if } x > 0, \\ V_{\text{mod}}(x,a,0) & \text{if } x^\star < x < 0, \\ V_{\text{smooth}}(x,b,x^c) & \text{if } x^c < x < x^\star, \\ 0 & \text{otherwise.} \end{cases} \tag{A.29}$$

We note that for given parameters of the main part of the smoothed functions (for example, ϵ, r_0, d and r_c for the V_{Morse} part of f_1), the parameters of the smoothed

cutoff regions (b^{low}, b^{high}, $r^{c,low}$, $r^{c,high}$ for f_1) are uniquely determined by ensuring differentiability of the function at the boundaries (r^{low} and r^{high} for f_1) and thus they are not explicitly listed in the function arguments and are not provided in the tables of values of model parameters.

The potentials are then

$$V_{\rm backbone} = V_{\rm FENE}(\delta r_{\rm backbone}, \epsilon_{\rm backbone}, \delta r^0_{\rm backbone}, \Delta_{\rm backbone}). \tag{A.30}$$

$$\begin{aligned} V_{\rm stack}\,(i,j) &= \eta(i,j)(1+\kappa\, k_{\rm B}T) f_1(\delta r_{\rm stack}, \epsilon_{\rm stack}, d_{\rm stack}, \delta r^0_{\rm stack}, \delta r^c_{\rm stack}, \delta r^{low}_{\rm stack}, \delta r^{high}_{\rm stack}) \\ &\times \quad f_4(\theta_{5'}, a_{\rm stack,5}, \theta^0_{\rm stack,5}, \Delta\theta^\star_{\rm stack,5})\ f_4(\theta_{6'}, a_{\rm stack,6}, \theta^0_{\rm stack,6}, \Delta\theta^\star_{\rm stack,6}) \\ &\times \quad f_4(\theta_9, a_{\rm stack,9}, \theta^0_{\rm stack,9}, \Delta\theta^\star_{\rm stack,9})\ f_4(\theta_{10}, a_{\rm stack,10}, \theta^0_{\rm stack,10}, \Delta\theta^\star_{\rm stack,10}) \\ &\times \quad f_5(\cos(\phi_1), a_{\rm stack,1}, \cos(\phi_1)^\star_{\rm stack})\ f_5(\cos(\phi_2), a_{\rm stack,2}, \cos(\phi_2)^\star_{\rm stack}). \end{aligned} \tag{A.31}$$

$$\begin{aligned} V_{\rm H.B.}(i,j) &= \alpha(i,j) f_1(\delta r_{\rm HB}, \epsilon_{\rm HB}, d_{\rm HB}, \delta r^0_{\rm HB}, \delta r^c_{\rm HB}, \delta r^{low}_{\rm HB}, \delta r^{high}_{\rm HB}) \\ &\times \quad f_4(\theta_1, a_{\rm HB,1}, \theta^0_{\rm HB,1}, \Delta\theta^\star_{\rm HB,1})\ f_4(\theta_2, a_{\rm HB,2}, \theta^0_{\rm HB,2}, \Delta\theta^\star_{\rm HB,2}) \\ &\times \quad f_4(\theta_3, a_{\rm HB,3}, \theta^0_{\rm HB,3}, \Delta\theta^\star_{\rm HB,3})\ f_4(\theta_4, a_{\rm HB,4}, \theta^0_{\rm HB,4}, \Delta\theta^\star_{\rm HB,4}) \\ &\times \quad f_4(\theta_7, a_{\rm HB,7}, \theta^0_{\rm HB,7}, \Delta\theta^\star_{\rm HB,7})\ f_4(\theta_8, a_{\rm HB,8}, \theta^0_{\rm HB,8}, \Delta\theta^\star_{\rm HB,8}). \end{aligned} \tag{A.32}$$

$$\begin{aligned} V_{\rm cross\ st.} &= \gamma f_2(\delta r_{\rm HB}, k_{\rm cross}, \delta r^0_{\rm cross}, \delta r^c_{\rm cross}, \delta r^{low}_{\rm cross}, \delta r^{high}_{\rm cross}) f_4(\theta_1, a_{\rm cross,1}, \theta^0_{\rm cross,1}, \Delta\theta^\star_{\rm cross,1}) \\ &\times \quad f_4(\theta_2, a_{\rm cross,2}, \theta^0_{\rm cross,2}, \Delta\theta^\star_{\rm cross,2})\ f_4(\theta_3, a_{\rm cross,3}, \theta^0_{\rm cross,3}, \Delta\theta^\star_{\rm cross,3}) \\ &\times \quad \left(f_4(\theta_7, a_{\rm cross,7}, \theta^0_{\rm cross,7}, \Delta\theta^\star_{\rm cross,7}) + f_4(\pi-\theta_7, a_{\rm cross,7}, \theta^0_{\rm cross,7}, \Delta\theta^\star_{\rm cross,7})\right) \\ &\times \quad \left(f_4(\theta_8, a_{\rm cross,8}, \theta^0_{\rm cross,8}, \Delta\theta^\star_{\rm cross,8}) + f_4(\pi-\theta_8, a_{\rm cross,8}, \theta^0_{\rm cross,8}, \Delta\theta^\star_{\rm cross,8})\right). \end{aligned} \tag{A.33}$$

$$\begin{aligned} V_{\rm coaxial\,st.} &= \mu f_2(\delta r_{\rm coaxial\,st.}, k_{\rm coax}, \delta r^0_{\rm coax}, \delta r^c_{\rm coax}, \delta r^{low}_{\rm coax}, \delta r^{high}_{\rm coax})\ f_4(\theta_4, a_{\rm coax,4}, \theta^0_{\rm coax,4}, \Delta\theta^\star_{\rm coax,4}) \\ &\times \quad \left(f_4(\theta_1, a_{\rm coax,1}, \theta^0_{\rm coax,1}, \Delta\theta^\star_{\rm coax,1}) + f_4(2\pi-\theta_1, a_{\rm coax,1}, \theta^0_{\rm coax,1}, \Delta\theta^\star_{\rm coax,1})\right) \\ &\times \quad \left(f_4(\theta_5, a_{\rm coax,5}, \theta^0_{\rm coax,5}, \Delta\theta^\star_{\rm coax,5}) + f_4(\pi-\theta_5, a_{\rm coax,5}, \theta^0_{\rm coax,5}, \Delta\theta^\star_{\rm coax,5})\right) \\ &\times \quad \left(f_4(\theta_6, a_{\rm coax,6}, \theta^0_{\rm coax,6}, \Delta\theta^\star_{\rm coax,6}) + f_4(\pi-\theta_6, a_{\rm coax,6}, \theta^0_{\rm coax,6}, \Delta\theta^\star_{\rm coax,6})\right) \\ &\times \quad f_5(\cos(\phi_3), a_{\rm coax,3'}, \cos(\phi_3)^\star_{\rm coax})\ f_5(\cos(\phi_4), a_{\rm coax,4'}, \cos(\phi_4)^\star_{\rm coax}). \end{aligned} \tag{A.34}$$

$$\begin{aligned} V_{\rm exc} &= f_3(\delta r_{\rm backbone}, \epsilon_{\rm exc}, \sigma_{\rm backbone}, \delta r^\star_{\rm backbone}) + f_3(\delta r_{\rm HB}, \epsilon_{\rm exc}, \sigma_{\rm base}, \delta r^\star_{\rm base}) \\ &+ \quad f_3(\delta r_{\rm back-base}, \epsilon_{\rm back-base}, \sigma_{\rm back-base}, \delta r^\star_{\rm back-base}) \\ &+ \quad f_3(\delta r_{\rm base-back}, \epsilon_{\rm back-base}, \sigma_{\rm back-base}, \delta r^\star_{\rm back-base}) \end{aligned} \tag{A.35}$$

Interaction	Parameters		
	cross-stacking: $V_{\text{cross. st.}}$		
$f_2(\delta r_{\text{HB}})$	$k_{\text{cross}} = 1.0$	$r^0_{\text{cross}} = 0.5$	$\delta r^c_{\text{cross}} = 0.6$
$\gamma = 59.96$	$\delta r^{low}_{\text{cross}} = 0.42$	$\delta r^{high}_{\text{cross}} = 0.58$	
$f_4(\theta_1)$	$a_{\text{cross},1} = 2.25$	$\theta^0_{\text{cross},1} = 0.505$	$\Delta\theta^\star_{\text{cross},1} = 0.58$
$f_4(\theta_2)$	$a_{\text{cross},2} = 1.70$	$\theta^0_{\text{cross},2} = 1.266$	$\Delta\theta^\star_{\text{cross},2} = 0.68$
$f_4(\theta_3)$	$a_{\text{cross},3} = 1.70$	$\theta^0_{\text{cross},3} = 1.266$	$\Delta\theta^\star_{\text{cross},3} = 0.68$
$f_4(\theta_7) + f_4(\pi - \theta_7)$	$a_{\text{cross},7} = 1.70$	$\theta^0_{\text{cross},7} = 0.309$	$\Delta\theta^\star_{\text{cross},7} = 0.68$
$f_4(\theta_8) + f_4(\pi - \theta_8)$	$a_{\text{cross},8} = 1.70$	$\theta^0_{\text{cross},8} = 0.309$	$\Delta\theta^\star_{\text{cross},8} = 0.68$
	coaxial stacking: $V_{\text{coaxial st.}}$		
$f_2(\delta r_{\text{coax}})$	$k_{\text{coax}} = 1.0$	$\delta r^0_{\text{coax}} = 0.5$	$\delta r^c_{\text{coax}} = 0.6$
$\mu = 80.0$	$\delta r^{low}_{\text{coax}} = 0.42$	$\delta r^{high}_{\text{coax}} = 0.58$	
$f_4(\theta_1) + f_4(2\pi - \theta_1)$	$a_{\text{coax},1} = 2.00$	$\theta^0_{\text{coax},1} = 2.592$	$\Delta\theta^\star_{\text{coax},1} = 0.65$
$f_4(\theta_4)$	$a_{\text{coax},4} = 1.30$	$\theta^0_{\text{coax},4} = 0.151$	$\Delta\theta^\star_{\text{coax},4} = 0.8$
$f_4(\theta_5) + f_4(\pi - \theta_5)$	$a_{\text{coax},5} = 0.90$	$\theta^0_{\text{coax},5} = 0.685$	$\Delta\theta^\star_{\text{coax},5} = 0.95$
$f_4(\theta_6) + f_4(\pi - \theta_6)$	$a_{\text{coax},6} = 0.90$	$\theta^0_{\text{coax},6} = 0.685$	$\Delta\theta^\star_{\text{coax},6} = 0.95$
$f_5(\cos(\phi_3))$	$a_{\text{coax},3'} = 2.00$	$\cos(\phi_3)^\star_{\text{coax}} = -0.65$	
$f_5(\cos(\phi_4))$	$a_{\text{coax},4'} = 2.00$	$\cos(\phi_4)^\star_{\text{coax}} = -0.65$	

Table A.1: Values of parameters in the model. In this table, all energies and lengths are in terms of the simulation units of energy and distance. When more than one function is listed for an interaction, the total interaction is a product of all the terms.

The excluded volume interaction between bonded neighbors, V'_{exc}, is the same as V_{exc} with the exception that it does not include the first term which depends on $\delta r_{\text{backbone}}$, because the neighbors already interact with the FENE potential through the V_{backbone} interaction that ensures that the backbone sites do not come too close.

The parameters of the interaction potentials are specified in tables A.1 and A.2.

Interaction	**Parameters**		
	backbone spring: V_{backbone}		
$V_{\text{FENE}}(\delta r_{\text{backbone}})$	$\epsilon_{\text{backbone}} = 2$	$\Delta_{\text{backbone}} = 0.25$	$\delta r^0_{\text{backbone}} = 0.76$
	hydrogen bonding: $V_{\text{H.B.}}$		
$f_1(\delta r_{\text{HB}})$	$\epsilon_{\text{HB}} = 1.0$	$d_{\text{HB}} = 8$	$\delta r^0_{\text{HB}} = 0.4$
$\alpha^{\text{avg}} = 0.87$	$\alpha(\text{A},\text{U}) = 0.82$	$\alpha(\text{G},\text{U}) = 0.51$	$\alpha(\text{G},\text{C}) = 1.06$
	$\delta r^c_{\text{HB}} = 0.75$	$\delta r^{low}_{\text{HB}} = 0.34$	$\delta r^{high}_{\text{HB}} = 0.70$
$f_4(\theta_1)$	$a_{\text{HB},1} = 1.50$	$\theta^0_{\text{HB},1} = 0$	$\Delta\theta^\star_{\text{HB},1} = 0.70$
$f_4(\theta_2)$	$a_{\text{HB},2} = 1.50$	$\theta^0_{\text{HB},2} = 0$	$\Delta\theta^\star_{\text{HB},2} = 0.70$
$f_4(\theta_3)$	$a_{\text{HB},3} = 1.50$	$\theta^0_{\text{HB},3} = 0$	$\Delta\theta^\star_{\text{HB},3} = 0.70$
$f_4(\theta_4)$	$a_{\text{HB},4} = 0.46$	$\theta^0_{\text{HB},4} = \pi$	$\Delta\theta^\star_{\text{HB},4} = 0.70$
$f_4(\theta_7)$	$a_{\text{HB},7} = 4.00$	$\theta^0_{\text{HB},7} = \pi/2$	$\Delta\theta^\star_{\text{HB},7} = 0.45$
$f_4(\theta_8)$	$a_{\text{HB},8} = 4.00$	$\theta^0_{\text{HB},8} = \pi/2$	$\Delta\theta^\star_{\text{HB},8} = 0.45$
	stacking: V_{stack}		
$f_1(\delta r_{\text{stack}})$	$\epsilon_{\text{stack}} = 1.0$	$d_{\text{stack}} = 6$	$\delta r^0_{\text{stack}} = 0.43$
	$\delta r^c_{\text{stack}} = 0.93$	$\delta r^{low}_{\text{stack}} = 0.35$	$\delta r^{high}_{\text{stack}} = 0.78$
	$\eta^{\text{avg}} = 1.402$	$\kappa = 1.9756$	
$\eta(\text{G},\text{C}) = 1.276$	$\eta(\text{C},\text{G}) = 1.603$	$\eta(\text{G},\text{G}) = 1.494$	$\eta(\text{C},\text{C}) = 1.473$
$\eta(\text{G},\text{A}) = 1.621$	$\eta(\text{U},\text{C}) = 1.167$	$\eta(\text{A},\text{G}) = 1.394$	$\eta(\text{C},\text{U}) = 1.471$
$\eta(\text{U},\text{G}) = 1.286$	$\eta(\text{C},\text{A}) = 1.583$	$\eta(\text{G},\text{U}) = 1.571$	$\eta(\text{A},\text{C}) = 1.210$
$\eta(\text{A},\text{U}) = 1.385$	$\eta(\text{U},\text{A}) = 1.246$	$\eta(\text{A},\text{A}) = 1.316$	$\eta(\text{U},\text{U}) = 1.175$
$f_4(\theta_{5'})$	$a_{\text{stack},5} = 0.90$	$\theta^0_{\text{stack},5} = 0$	$\Delta\theta^\star_{\text{stack},5} = 0.95$
$f_4(\theta_{6'})$	$a_{\text{stack},6} = 0.90$	$\theta^0_{\text{stack},6} = 0$	$\Delta\theta^\star_{\text{stack},6} = 0.95$
$f_4(\theta_9)$	$a_{\text{stack},9} = 1.3$	$\theta^0_{\text{stack},9} = 0$	$\Delta\theta^\star_{\text{stack},9} = 0.8$
$f_4(\theta_{10})$	$a_{\text{stack},10} = 1.3$	$\theta^0_{\text{stack},10} = 0$	$\Delta\theta^\star_{\text{stack},10} = 0.8$
$f_5(\cos(\phi_1))$	$a_{\text{stack},1} = 2.00$	$\cos(\phi_1)^\star_{\text{stack}} = 0.65$	
$f_5(\cos(\phi_2))$	$a_{\text{stack},2} = 2.00$	$\cos(\phi_2)^\star_{\text{stack}} = 0.65$	
	excluded volume: V_{exc}		
$f_3(\delta r_{\text{backbone}})$	$\epsilon_{\text{exc}} = 2.00$	$\sigma_{\text{backbone}} = 0.70$	$\delta r^\star_{\text{backbone}} = 0.675$
$+f_3(\delta r_{\text{HB}})$	$\epsilon_{\text{exc}} = 2.00$	$\sigma_{\text{base}} = 0.33$	$\delta r^\star_{\text{base}} = 0.32$
$+f_3(\delta r_{\text{back-base}})$	$\epsilon_{\text{exc}} = 2.00$	$\sigma_{\text{back-base}} = 0.515$	$\delta r^\star_{\text{back-base}} = 0.50$
$+f_3(\delta r_{\text{base-back}})$	$\epsilon_{\text{exc}} = 2.00$	$\sigma_{\text{back-base}} = 0.515$	$\delta r^\star_{\text{back-base}} = 0.50$

Table A.2: Further parameter values in the model. In this table, all energies and lengths are in terms of the simulation units of energy and distance. When more than one function is listed for an interaction, the total interaction is a product of all the terms, with the exception of V_{exc}, which is a sum of the respective terms.

Appendix B

Simulation methods

To study DNA and RNA systems with the oxDNA and oxRNA models, we run simulations that sample from the Boltzmann distribution of an NVT ensemble, i.e. we keep the number of particles, the simulation box volume and the temperature constant in our simulations. Our simulation code implements Monte Carlo methods (Metropolis Monte Carlo [222] and Virtual Move Monte Carlo [223]) as well as Molecular dynamics code with an Andersen-like thermostat [200]. We briefly introduce the algorithms in this appendix.

B.1 Metropolis Monte Carlo algorithm

The Metropolis Monte Carlo sampling algorithm is a very popular algorithm used to sample configurations from a canonical ensemble at temperature T, where the probability of observing the system in a configuration given by the position and orientation of the particles $\left(\mathbf{r}^N, \mathbf{\Omega}^N\right)$ is proportional to $\exp(-\beta V\left(\mathbf{r}^N, \mathbf{\Omega}^N\right))$ where $\beta = 1/k_{\mathrm{B}}T$ and V is the potential function of the simulated system of N particles. In each step of the Metropolis Monte Carlo algorithm, we randomly select a nucleotide and a proposed move. The proposed move can either be a rotation around a random axis which passes through the backbone site of the nucleotide (by an angle drawn from a normal distribution with zero mean and standard deviation 0.2 radians) or a translation through vector with a random direction with a length drawn from a normal

distribution with zero mean and standard deviation 0.1 simulation code length unit. The proposed move is accepted with a probability

$$P(i \to j) = \min\left(1, \exp\left(-\beta\left(E_j - E_i\right)\right)\right) \tag{B.1}$$

where E_i and E_j are the total energies of the studied system before and after the proposed move respectively. While it is fairly easy to implement the Metropolis Monte Carlo sampling, it was found during the development of the oxDNA model [140] that the sampling of the ensemble is not very efficient with this simple algorithm: collective motion is suppressed, as each nucleotide is moved individually and in order to move a physical cluster (for example a single strand) in one direction, all of the nucleotides have to be moved separately in the same direction. Long translations are unlikely to be accepted, as moving the nucleotide to a large distance from the others is likely to result in a large increase in the backbone potential V_{backbone}. In order to sample DNA and RNA systems, we therefore use a cluster-move Monte Carlo algorithm, which can move in one step a cluster with multiple particles. In particular, we use the Virtual Move Monte Carlo algorithm, which is briefly introduced in the following section.

B.2 VMMC algorithm and umbrella sampling

B.2.1 Virtual Move Monte Carlo algorithm

For the majority of the simulations described in this thesis, unless noted otherwise, we use the Virtual Move Monte Carlo algorithm (VMMC) (specifically the variant described in the Appendix of [223]) to study the thermodynamics of oxDNA and oxRNA. VMMC is a cluster-move Monte Carlo algorithm, in which new configurations are proposed by moving dynamically selected clusters of particles (nucleotides in our case). The proposed moves used in our VMMC simulations are either rotation around a randomly chosen nucleotide's backbone site or a linear translation, with the same parameters as described in Section B.1. After the selection of the initial nucleotide,

the VMMC algorithm forms a cluster which will recursively include, with a certain probability, the neighbors of the selected nucleotide. The nucleotides get recruited into the cluster depending on the change in their interaction energies before and after the proposed move. Hence, the algorithm takes into account not only the interaction strength before the move, but also to some extent the gradient of the interaction strength. Detailed balance needs to be ensured, which makes the computation of the acceptance probability a more complicated expression than for the Metropolis Monte Carlo algorithm (see [223] for a derivation).

Despite the higher computational cost of generating the cluster, VMMC accelerates the equilibration of dilute, strongly-interacting systems relative to Monte Carlo described in Section B.1. A whole strand can be included in the generated cluster and hence translated or rotated in a single move, thus speeding up the diffusion.

Currently, a comprehensive comparison of VMMC with other cluster-move Monte Carlo algorithms is not available, but we assume that the VMMC algorithm is efficient for DNA and RNA systems mainly because the change of interaction energy with the proposed move is taken into account during the cluster generation. For instance, in the VMMC cluster generation, two interacting nucleotides are more likely to be included in a single cluster if a move that would break the interaction is proposed. Therefore, if a large move is proposed, the whole strand is likely to be included. On the other hand, if the interaction between two nucleotides does not vary substantially before and after the proposed move, it is likely they will not be included in the same cluster, which can be helpful for example in duplex equilibration, where only a part of a strand can be slightly rotated in a single move while it still maintains hydrogen bonds with the complementary strands.

Since the canonical ensemble sampling with VMMC was found very efficient for many previous applications of oxDNA [146], we also choose to use it for oxRNA, given the similarity of the models.

In our VMMC simulations we usually compute the free energy of the system as a function of the number of base pairs between the interacting strands. If there are only two potentially interacting strands (or two regions of a single strand, as in the case of a hairpin stem), the order parameter would typically be the number of base pairs b between the strands. We define a base pair as being formed if the hydrogen-bonding energy ($V_{\mathrm{H.B.}}$ in Eqs. 2.1 and 5.2) between the two bases is more negative than $-1.0\,k_{\mathrm{B}}T$ for $T = 300\,\mathrm{K}$, which corresponds to about 15% of typical hydrogen-bonding energies in the oxDNA model and to about 18% of the typical hydrogen-bonding energy for a base pair in oxRNA. The free energy F of a state with b formed base pairs is related to the probability P that the system is found in such a state during simulations by

$$F\left(b\right)/k_{\mathrm{B}}T = -\log P\left(b\right) + \log P_0\,, \tag{B.2}$$

where P_0 is an arbitrary normalization. A relatively high free energy thus corresponds to a relatively unlikely state.

B.2.2 Umbrella sampling

DNA and RNA systems studied in this thesis often have few thermodynamically relevant macrostates (for example a hairpin and a singe-stranded state). Sampling all relevant states can be difficult even with an efficient algorithm such as VMMC, as these (meta)stable states are often separated by high free-energy barriers that are difficult to cross. Simulations can therefore get stuck for large portions of the simulation time in a local free-energy minimum. In order to overcome this problem we combine the VMMC algorithm with the umbrella sampling method [171, 224].

Instead of sampling from the Boltzmann distribution, we sample from configurations with a weight $w(b)\exp(-\beta V\left(\mathbf{r}^N, \mathbf{\Omega}^N\right))$, where $w(b)$ is an arbitrary biasing potential, defined as a function of a specific order parameter (reaction coordinate) b (typically the number of bonds between the interacting strands). $w(b)$ can be chosen

to raise the probability of visiting unlikely transition intermediates, thereby accelerating equilibration between local free-energy minima. To obtain an unbiased estimate of the free energy, one must correct for the applied bias as follows:

$$F\left(b\right)/k_{\mathrm{B}}T = -\log\left(\frac{P_{\mathrm{biased}}\left(b\right)}{w\left(b\right)}\right) + \log P_0, \tag{B.3}$$

where $P_{\mathrm{biased}}\left(b\right)$ is the probability with which states appear in the biased simulation. We note that in order to obtain the correct estimate of the free energy, the states sampled with the biasing potential $w(b)$ have to be representative of the states that would be sampled in a simulation without the biasing potential.

We further note that in more complicated systems, apart from considering order parameters based on the number of bonds between potentially interacting regions of strands, it is also possible to additionally introduce distance between strands (or their regions) as an order parameter. In such a case, we define a set of intervals for the possible values of the minimum distance between the selected complementary nucleotides in the strands, with a specific weight assigned to each particular interval. In order to facilitate the study of, for example, the association of two strands, higher weights are assigned to the states with a smaller distance between the complementary nucleotides.

In systems studied with the umbrella sampling method, we typically first run several simulations to determine $w(b)$. The initial weights $w(b)$ are chosen by experience and then iteratively adjusted by hand to enhance sampling of the relevant states. The weights are adapted so that the system spends approximately the equal amount of time in all relevant states respectively, which is achieved by setting the weights to be inversely proportional to the (unbiased) number of iterations spent in a given state in the initial simulations. The correct choice of weights then enables efficient sampling of relevant states and hence the free energy profile can be obtained faster than with an unbiased simulations. Once satisfactory weights $w(b)$ are obtained, long simulations are run to collect the final data.

B.2.3 Estimating melting temperatures from VMMC simulations

For many of the systems considered in this thesis, we use the VMMC algorithm with umbrella sampling to obtain the melting temperature for the oxDNA or oxRNA model. For the case of dimer association, such as the melting of a duplex, one needs to sample many times the transitions between single-stranded and duplex states. In our simulations, we define the duplex state as having at least one of the bonds between the complementary strands formed. Larger weights need to be assigned to the states with only few bonds between the strands, as these states are free-energetically unfavorable, but the system needs to efficiently sample them in order to be able to pass from the duplex state to the single-stranded state and *vice versa*.

The melting temperature is defined as the temperature at which the yield of duplexes in the bulk is 50%. In our melting temperature simulations, we only simulate a single duplex in the simulation box and hence finite-size effects need to be taken into account [167, 166, 168]. If the two strands are not self-complementary (i.e. are heterodimers), the ratio of the number of sampled duplex states to the number of sampled single-stranded states is equal to 2 (and to 1 for homodimers) at the melting temperature.

For each umbrella sampling VMMC simulation, we count the (unbiased) numbers of steps spent in the single-stranded and duplex states. We use histogram reweighting, as outlined in Section 2.3.2, to extrapolate these numbers to a range of temperatures in the neighborhood of the temperature at which the VMMC simulation was run. By interpolating the calculated ratios for each temperature, we obtain the melting temperature as the temperature for which the ratio is 2 (or 1 in the case of homodimers).

For hairpin melting simulations, finite-size effects are not present since it is a unimolecular transitions. The melting temperature of a hairpin is hence obtained when the ratio of the number of simulation steps spent in the single-stranded state

and the number of steps in the hairpin form is equal to 1. We define the strand to be in the hairpin state if at least one bond is present in the stem.

We note that, with the appropriate choice of weights, on the order of 10^9 VMMC steps are necessary to obtain the melting temperature of an 8-mer within $1 - 2\,^{\circ}\mathrm{C}$ precision. Such a simulation takes several days on a single 2.2 GHz CPU. Multiple simulations can however be run in parallel and their results combined.

B.3 Molecular dynamics

Broadly speaking, Molecular dynamics simulation codes integrate Newton's equations of motion to follow the time evolution of a studied system [225]. The oxDNA and oxRNA potentials are designed to be differentiable functions in order to allow for such a simulation.

Since the nucleotides in oxDNA and oxRNA are represented as rigid bodies, we solve the equations of motion for both linear and angular momenta. We set the mass m of the nucleotides to be equal to 1 in simulation units for oxDNA and oxRNA. We assume that the moment of inertia I for our nucleotides corresponds to that of a sphere and we also set it to be equal to 1 in simulation units. The average mass of a nucleotide corresponds to 315.75 AMU for DNA and 321.4 AMU for RNA. Since the simulation energy unit of the simulation code is the same for both models (as defined in [140] and Appendix A), the simulation code time unit is 3.03×10^{-12} s for oxDNA and 3.06×10^{-12} s for oxRNA.

In order to use Molecular dynamics codes to sample from the NVT ensemble, one needs to introduce a thermostat, i.e. a coupling to a heat bath at temperature T. Our simulation code implements a thermostat which is based on the Andersen thermostat [226]. The detailed description of the algorithm used is provided in the appendix of [200]. We refer to this thermostat as "Andersen-like". This thermostat evolves the system according to Newton's equations for a number of steps N_{Newt}, using the

Velocity Verlet algorithm [225, 227]. It then resets the velocity of each nucleotide with probability p_v and the angular velocity of each nucleotide with a probability p_ω. The newly assigned velocities and angular velocities are drawn from a Boltzmann distribution for a given temperature T. By default, we choose $N_{\mathrm{Newt}} = 103$ and $D = 2.5$ in simulation units (corresponding to $6.0 \times 10^{-7}\,\mathrm{m^2 s^{-1}}$ and $5.8 \times 10^{-7}\,\mathrm{m^2 s^{-1}}$ in SI units for oxDNA and oxRNA respectively) for the translational diffusion coefficient and $D_{\mathrm{rot}} = 3D$ for the rotational diffusion coefficient. The corresponding probabilities of reseting velocities can be obtained from [200]:

$$p_v = \frac{2k_{\mathrm{B}}TN_{\mathrm{Newt}}\delta t}{k_{\mathrm{B}}TN_{\mathrm{Newt}}\delta t + 2mD} \tag{B.4}$$

$$p_\omega = \frac{2k_{\mathrm{B}}TN_{\mathrm{Newt}}\delta t}{k_{\mathrm{B}}TN_{\mathrm{Newt}}\delta t + 2ID_{\mathrm{rot}}} \tag{B.5}$$

On time scales longer than $N_{\mathrm{Newt}}\delta t/p_v$, where δt is the integration time step, the dynamics is diffusive.

For an integration step δt corresponding to 15.2 fs (0.005 in simulation time units), we measured the diffusion coefficient for an MD simulation of the oxDNA model with the Andersen-like thermostat with the above parameters. For a 14-bp DNA duplex we found $D_{\mathrm{sim}} = 2.1 \times 10^{-8}\,\mathrm{m^2 s^{-1}}$, which is significantly higher than the experimental measurements of $D_{\mathrm{exp}} = 1.19 \times 10^{-10}\,\mathrm{m^2 s^{-1}}$ [228]. We intentionally set the diffusion coefficient higher for our coarse-grained models in order to accelerate the diffusion process and access longer time-scales when studying complex processes.