Modeling Complex Structures in Nucleic Acids

A dissertation
submitted to the faculty of the graduate school
of the University of Minnesota
by

Margaret C. Linak

In partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Advisor: Kevin D. Dorfman

April, 2012

Acknowledgements

Science is not done in a vacuum, it is influenced by and in turn influences the surrounding world. Similarly scientists and researchers owe a great debt not only to the generations of thinkers, inventors, and researchers that proceeded them but also to all those that currently surround them. I would like to thank the many people who have helped, guided, and influenced me.

I owe a great deal to colleagues, teachers, friends, and members of my family who through their own research, comments, and questions have encouraged, supported, and enlightened me. I would like to express my very great appreciation to my advisor, Dr. Kevin Dorfman; his patient guidance, enthusiastic encouragements, and useful critiques of my work have undoubtedly made me a better researcher. I would also like to offer sincere thanks to the other professors who have continued my education. The faculty of this department have provided me with a tremendous graduate training: they have taught me how to think in novel ways; have provided me with interesting scientific opportunities; and have immersed me in an academic environment that has forged me into the person I am today. Further, I am grateful to Dr. Martin Kenward who first introduced me to the world of simulations and to the other members of the Dorfman research group for their vast technical skills, patience, and interesting, albeit off topic, discussions.

During my graduate education I was also fortunate enough to be exposed to graduate students from many other disciplines and these interactions have unquestionably broadened my understanding not only of chemical engineering but also of the greater scientific and social context of my research. I would like to thank the members of the BioTechnology Institute at the University of Minnesota; not only did you first teach me the basics of biology but through your help I have truly come to understand the wider biological context surrounding my work. In addition, I was able to augment my

technical lessons with instruction and research in public policy. I am indebted to the students and professors that helped me bridge the, once thought great, divide between the realms of science and policy. Finally, I grateful to the members of the Kanaya research laboratory at the Nara Institute of Science and Technology, Nara, Japan where I was a visiting researcher. By sharing not only your research, but also your country, you truly provided me with an intellectual and engaging experience.

Finally, I would like to thank my family. To my parents who instilled in me an immense love of knowledge, true wonder and intrigue, and the ability to question, I am grateful to the start in life they gave me. To my brother, Scott, who early on explored the world with me and now continues to be a source of friendship and strength, I am thankful. Their support and encouragement has been fundamental to my success.

This work was supported by a Career Development Award from the International Human Frontiers Science Program Organization, the David and Lucile Packard Foundation, and a Biotechnology Training Grant from the National Institute of Health (Grant No. 5T32GM008347-20). In addition, I would like to acknowledge support at the University of Minnesota from a Graduate Student Fellowship and computing resources at the Minnesota Supercomputing Institute.

Abstract

Since the discovery of DNA, researchers have been attempting to decode the detailed structure, properties, and abilities of this molecule. At first approximation, DNA can be thought of as a long, regular, double-stranded helix encoding the genomic information of life. However, on closer analysis DNA has been found to take on a wide variety of complex shapes and functions both *in vitro* and *in vivo*. DNA can be single-, double-, triple-, and even quadruple-stranded in nature and can bind in both the Watson–Crick conformations and also in a variety of non-canonical configurations that add to its inherent flexibility, structure, and activity. Elucidating the varied structures and behaviors of DNA has historically been an experimental endeavor, due in large part to the difficulties in capturing nucleic acid's complex motions and function in a tractable computational model. However, as the applications of DNA expand and computation power increases, simulation models are playing an increasingly important role in DNA understanding and engineering. In this thesis, we simulate short DNA and RNA (less than 100 nucleotides) and examine their complex structures. In particular, we will (i) experimentally evaluate previous DNA coarse grained models for their ability to capture complex nucleic acid structures, and (ii) develop a new model that can better capture both canonical and non-canonical interactions and show its utility in the study of several known structures. Further, we will use our understanding of the intricate interactions of short oligonucleotides to unravel a hereto experimentally inaccessible mechanistic pathway for a catalytically active DNA molecule. The model developed and the importance of non-canonical interactions in nucleic acid systems will be useful in the continued understanding and engineering of DNA and RNA molecules for nanotechnology, genetic engineering, and therapeutic applications.

Contents

List of Tables

List of Figures

1 Primary Structure of Nucleic Acids

We also now appreciate that molecular biology is not a trivial aspect of biological systems. It is at the heart of the matter. Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself. All approaches at a higher level are suspect until confirmed at the molecular level.

Francis Crick, *What Mad Pursuit*, 1988 [1]

1.1 Introduction to Nucleic Acids

Nucleic acids are molecules which play a central role in the transmission, expression, and conservation of genetic information. The most common nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). First discovered by Friedrich Miescher in 1869 as a phosphate-rich chemical found in cells, nucleic acids have been the subject of study ever since [2]. Recently, it has been discovered that nucleic acids can play even wider and more unexpected functions than that of a simple genetic library.

The role of DNA as the carrier of genetic information has been amply demonstrated beginning with the classic experiments of Avery *et. al.* in 1944 [3] and Hershey and Chase in 1952 [4]. These experiments marked the opening of the contemporary era of genetics, the bridge connecting 'chromatin' to classical Mendelian thought, and introducing the 'DNA only' view of inheritance [5]. With this understanding the search began to connect the matter of life with its unique and individual functions.

The classic example of how biological function follows from biomolecular structure comes from the elucidation by Watson and Crick in 1953 of the structure of DNA as a double helix [6, 7], using the X-ray diffraction patterns generated by Franklin [8], Wilkins [9], and their associates, and the chemical evidence of base complementarity of Chargaff [10] from the early 1950s. In their seminal paper, Watson and Crick explicitly suggest the relationship between form and function when they state that "it has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material" [6]. From this start of understanding biological purpose and biomolecular structure, the study of nucleic acids has grown to encompass entire scientific disciplines; in fact, this extremely versatile molecule continues to be found to have new forms and novel functions.

We begin our examination of the interplay of structure and function by examining nucleic acids at the molecular level, as suggested by Francis Crick in the introductory quotation of this chapter from his book *What Mad Pursuit* [1]. This chapter discusses the physical and chemical properties of the monomeric building blocks of nucleic acid structure: the bases, nucleosides, and nucleotides; the components of its primary structure. Seemingly recondite monomer structural features such as the electron distributions and bond conformations are included and will be built upon in this and later chapters as key factors of base stacking, hydrogen bonding, and helix geometries. As we move through this treatise, we will explore DNA's other levels of complexity, moving from the sequence of its bases (primary structure) to base pairing (secondary structure) to its three-dimensional shape (tertiary structure).

1.2 Composition of Nucleotides

Nucleotides have many functions in living organisms; they participate as essential intermediates in virtually all aspects of cellular metabolism [11]. Serving an even more central biological purpose are nucleic acid molecules; they are the elements of heredity, the emissaries of genetic information transfer, and the agents of catalytic change. Nucleic acids are linear polymers of nucleotides in which the order of their subunits, or primary structure, promote additional information, structure, and function. A nucleotide consists of three molecular fragments: a heterocyclic nitrogen group, a sugar group, and a phosphate group.

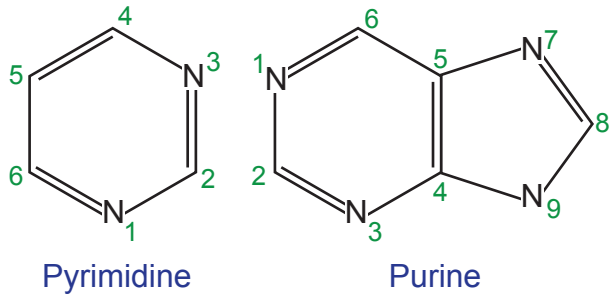

Figure 1.1: A schematic representation of the pyrimidine and purine ring structures with the conventional numbering shown.

1.2.1 Nitrogenous Bases

The heterocyclic nitrogenous bases of nucleotides are derivatives of either pyrimidines or purines. Pyrimidines are six-membered heterocyclic aromatic rings containing two nitrogen atoms. The atoms are numbered in a counterclockwise fashion, as shown in Figure 1.1 [12]. The purine ring structure is represented by the combination of a pyrimidine ring with a five-membered imidazole ring to yield a fused ring system. The nine atoms in this system are numbered according to the convention shown in Figure 1.1. Each base is essentially planar, and its conformations are limited [12].

The common naturally occurring pyrimidines are cytosine (2-oxy-4-amino pyrimidine), thymine (2-oxy-4-oxy-5-methyl pyrimidine), and uracil (2-oxy-4-oxy pyrimidine). The latter is only found in RNA and replaced by the functionally equivalent thymine in DNA. Adenine (6-amino purine) and guanine (2-amino-6-oxy purine) are the two common purines appearing in nucleic acids, as depicted schematically in Figure 1.2 [11]. Various other pyrimidine and purine derivatives are also present in nucleic acids as minor constituents and will not be further discussed here [13]. Resonance possibilities within the pyrimidine and purine ring systems and the electron-rich nature of their –OH and $–NH_2$ substituents endow them with the capacity to undergo tautomeric shifts [12] and multiple hydrogen bonding incidences [13] as will be discussed in Section 1.3.1.

1.2.2 Sugar Group

Nucleic acids are comprised of a cyclic, furanoside-type sugar: $\beta-$ D-ribose in RNA and $\beta-$ D-deoxyribose in DNA which is schematically depicted in Figure 1.3 [13].

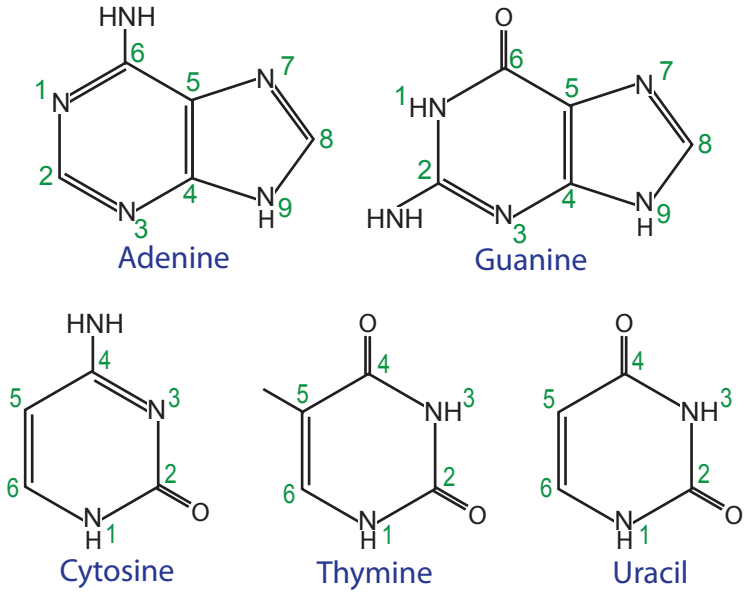

Figure 1.2: A schematic representation of the common pyrimidine and purine nitrogenous bases found in nucleic acids; thymine and uracil are exclusively found in DNA and RNA, respectively. Purines are connected to DNA or RNA sugars at the N_9 position in the imidazole ring. The pyrimidines are connected to DNA or RNA sugars at the N_1 position.

When these ribofuranoses are found in nucleotides, their atoms are numbered as 1', 2', 3', 4', and 5' to distinguish them from the ring atoms of the nitrogenous bases [11]. The presence in the -OH group at the 2'-position of the sugar along with the presence of uracil instead of thymine distinguishes RNA from DNA.

A nucleoside is constructed from a nitrogenous base and a sugar group joined by a β-gylcosyl $C_{1'}-N_{1\text{ or }9}$ linkage. The free bases depicted in Figure 1.2, bear a hydrogen atom in position 9 (purine) and 1 (pyrimidine); in nucleosides, this hydrogen atom is replaced by a single bond connecting it to the sugar moiety [13]. Unlike the planar ring structures of the bases, ribofuranose rings in nucleosides can exist in four different

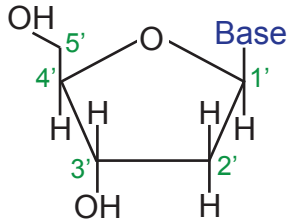

Figure 1.3: A schematic depiction of $\beta-$ D$-$deoxyribose sugar. The nitrogenous base is attached as shown at the $C_{1'}$ position with glycosidic bond.

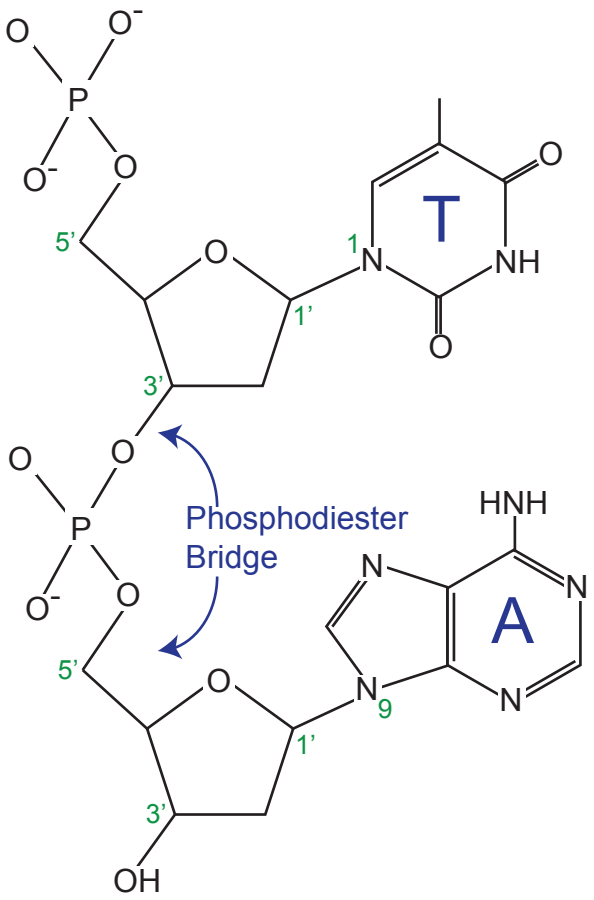

Figure 1.4: A schematic depiction of the 3' to 5' phosphodiester bridge linking nucleotides together to form a polynucleotide chain, in this case single-stranded DNA.

puckered conformations. In all cases, four of the five atoms are in a single plane. The fifth atom ($C_{2'}$ or $C_{3'}$) is on either the same (endo) or the opposite (exo) side of the plane relative to the $C_{5'}$ atom [14].

1.2.3 Phosphate Group

A nucleotide results when phosphoric acid is esterified to a sugar -OH group of a nucleoside. The nucleoside ribose ring has three -OH groups available for esterification, at $C_{2'}$, $C_{3'}$, and $C_{5'}$ (with the first lacking in DNA nucleosides). The $C_{2'}$ and $C_{3'}$ do not occur naturally, but can be generated from nucleic acid hydrolysis; DNA and RNA nucleotides are phosphorylated at the $C_{5'}$ position [13].

Linear chains of nucleotides are formed with 3' to 5' phosphodiester bridges, as depicted in Figure 1.4. They are formed as each nucleotide is successively added to the

3' -OH group of the preceding nucleotide, a process that gives the polymer a directional sense. This polarity is defined by the asymmetry of the nucleotides and the way they are joined. Phosphodiester linkages create the repeating sugar-phosphate backbone of the polynucleotide chain, which is a regular feature of both DNA and RNA. We shall refer to the strand that forms as either RNA (if comprised of uracil and ribose sugar groups) or single-stranded DNA (ssDNA) (if comprised of thymine and deoxyribose sugar groups). Although RNA polynucleotide chains are typically short ($\ll 100$ nucleotides), DNA chains can contain hundreds of millions of nucleotide units [11].

1.2.4 Nucleic Acid Nomenclature

The only significant variation in the chemical structure of nucleic acids is the nature of the nitrogenous base at each nucleotide position. These bases are not part of the sugar-phosphate backbone but instead serve as distinctive side chains. They give each polymer a unique identity, property, and structure. The convention in notation of nucleic acid structure is to read the polynucleotide chain from the 5'-end of the polymer to the 3'-end. Note that this reading direction actually passes through each phosphodiester bond from 3' to 5' as seen in Figure 1.4 [11].

1.3 Nucleotide Interactions

Base-base interactions are of two kinds: (i) those in the plane of the bases (horizontal) due to hydrogen bonding and (ii) those perpendicular to the base planes (vertical) stabilized mainly by London dispersion forces and hydrophobic effects and constituting base stacking [13]. Hydrogen bonding is most pronounced in non-polar solvents where base stacking is negligible; base stacking dominates in water where base-base hydrogen bonding is greatly suppressed due to competition of binding sites by water molecules [12]. The charge densities of the atoms in each nucleotide explain most of their interactive behaviors with other molecules.

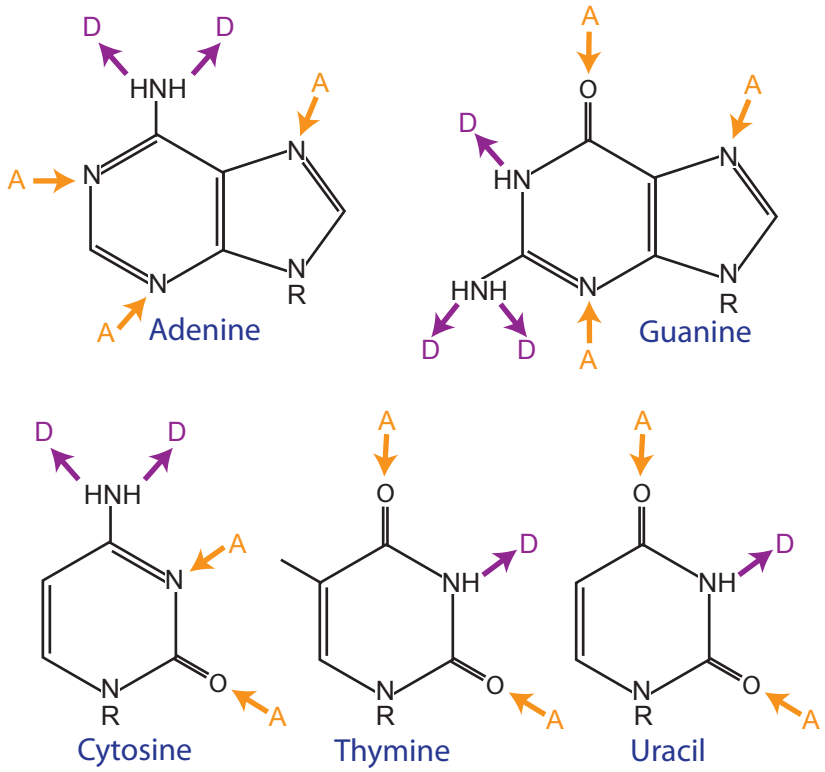

Figure 1.5: A schematic representation of the potential hydrogen bonding sites in the bases of nucleosides where the R group is the respective sugar moiety. Hydrogen bonding donor sites (D) (amino or imino protons) are labeled with dark (purple) arrows; hydrogen bonding acceptor sites (A) (carbonyl oxygens or aromatic nitrogens) are labeled with light (orange) arrows.

1.3.1 Hydrogen Bonding

Hydrogen bonds are mainly electrostatic in character. They play a key role in nucleic acid secondary structure and are fundamental to the biological functions of DNA and RNA. In general, a hydrogen bond X−H $\cdots$ Y is formed if a hydrogen atom H connects two atoms X and Y of higher electronegativity. Hydrogen bonds are "soft" and only weakly directional. Compared to covalent bonds with their well-defined length, strength, and orientation, hydrogen bonds are about 20 to 30 times weaker; as an example, the energy required to lengthen a C−C covalent bond by 0.1 Å is 3.25 kcal/mol while a similar extension to a O−H $\cdots$ O hydrogen bond only necessitates 0.1 kcal/mol [12]. Similarly, other base-base hydrogen bond geometries favor maximum distances of 3.15 Å (for N−H $\cdots$ N) and angles up to 25°, as compared to a covalent C−C bond distance of 1.54 Å and in-line rigidity [12]. Therefore, systems relying on hydrogen bonding, such as those prevalent in nucleic acids, are more susceptible to

bending and stretching, and thus result in many variable geometries.

The oxygen and nitrogen atoms of each nitrogenous base can act as hydrogen bond donors or acceptors; the sites are illustrated in Figure 1.5 [13]. Each base has donor and acceptor sites, so if there are no constraints on the geometry of the interaction each base can pair with itself or any other base through hydrogen bonding [15]. Even when the nucleotides are bound together with phosphodiester bridges, the flexibility and bonding potential of the resulting nucleic acid can lead to a wide variety of overall structures [13]. These structures will be discussed further in relation to RNA and ssDNA in Section 2.3. The bases can also hydrogen bond to polar amino acids in nucleic acid–protein interactions [14]. Nucleosides and nucleotides provide further hydrogen bonding opportunities through the 2'-OH group and the phosphate group [13].

Through a series of electrotitrimetric studies, Gullarnd first determined that bases were linked by hydrogen bonding [13]. Since 1947, hydrogen bonding interactions between like and different bases have been often observed experimentally in crystal structure analysis of individual bases, nucleosides, and nucleotides [12]. Under the assumption that at least two hydrogen bonds must form between any two bases in order to produce a stable base pair (bp), the four bases (since uracil and thymine similarly bond) can be arranged in 28 different configurations [12, 13]. In a series of solutions of nucleotides, all possible two hydrogen bond — but no one hydrogen bond — configurations have been measured [12, 16]. However, when additional geometric restraints are included, the number of probable base-pairs is significantly reduced. Here we will discuss in detail three types of hydrogen bonds that are vital in nucleic acids and measured experimentally: Watson-Crick, Hoogsteen, and wobble base pairs.

Watson–Crick Base Pairs

Hypothesized in their seminal 1953 paper, Watson and Crick introduced the idea of specificity in DNA hydrogen bonding; that is, that the bases in the most common form of DNA maintain two particular pairings: adenine with thymine ($A \cdot T$) and cytosine with guanine ($C \cdot G$) [6, 7], as shown in Figure 1.6. This specific set of pairings was inferred from two sets of data, (i) when a pyrimidine was paired with a purine the distance between their C_1, sugar atoms was essentially constant and (ii) Chargaff had measured that the concentrations of A and T and C and G were always the same

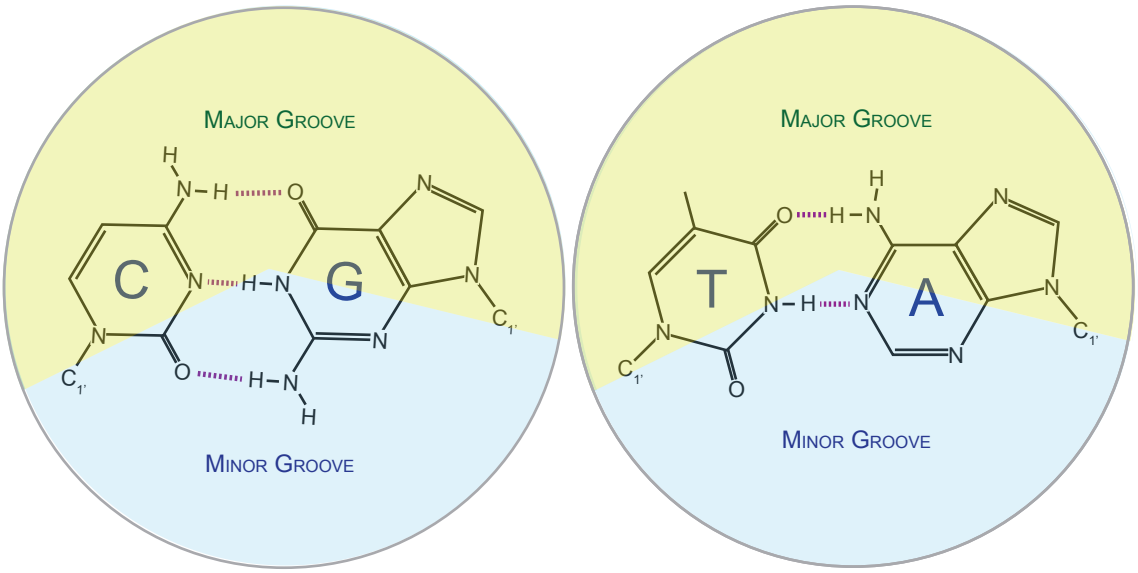

Figure 1.6: Watson-Crick base pairs found in the regular structure of double-stranded DNA. The A · T and C · G base pairs have the same distance between the $C_{1'}$ atoms of their sugars and can form a regular helix of any sequence. Each nucleic acid double helix has a major and minor groove; the minor groove is on the side of the base pair where the sugars are attached. Only the hydrogen atoms explicitly involved in hydrogen bonding are depicted.

throughout the genomes of any organism ($\frac{[A]}{[T]} = \frac{[C]}{[G]} = 1$) [6–10]. In addition, Watson and Crick also hypothesized that the specificity of these base pairs could explain how the genes in DNA could be duplicated (for stable inheritance) upon cell division [6, 7]. The characteristic geometries and features of two nucleic acid strands enjoined by Watson–Crick hydrogen bonds will be further discussed in Section 2.2.

Hoogsteen Base Pairs

If we examine the hydrogen acceptor and donor sites labeled in Figure 1.5, we can see that there are still a large number of configurations available. Indeed, if we simply examine the adenine and thymine bases, we can see that there are at least three other configurations for the A · T base pair with two hydrogen bonds, as depicted in Figure 1.7. These additional possible configurations were used as an explanation by Karst Hoogsteen in 1959 and 1963 for the anomalies he recognized between the electron-density projection measured by the Weissenberg photographs he took and the predicted Watson–Crick results [17, 18]. From these differences, Hoogsteen deduced that nucleotides could arrange themselves in ways other than the newly defined canonical Watson–Crick formation. Further base combinations were found that ex-

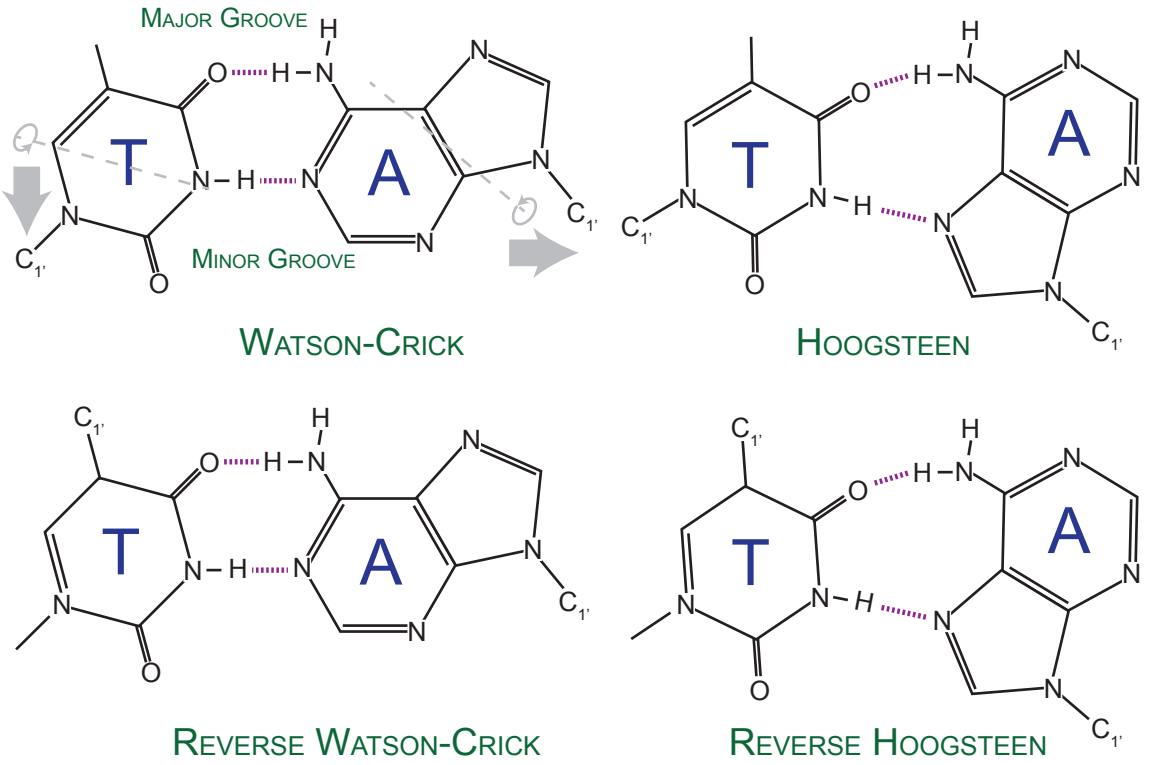

Figure 1.7: The configurations of Watson–Crick and Hoogsteen type bonding between adenine and thymine. Reverse base pairing occurs when one base is rotated 180° with respect to the other. It should also be noted that the two bases in the Watson-Crick hydrogen bonding pair are not coplanar. Rather they are twisted about the hydrogen bonds at approximately 12° in the A · T pair. In contrast, the A · T Hoogsteen base pair can either be in a perfectly coplanar arrangement or rotated 9° [12]. The bases will be schematically drawn in a planar fashion regardless of any twist about the hydrogen bonds. The dashed (gray) line, rotation arrow, and pointer highlight the rotation axis necessary to form the different configurations.

panded the variety of Hoogsteen base pairs; however, a generalized definition of a Hoogsteen type bond is one where the hydrogen bonding occurs with atoms on the major groove side (if it were in a Watson–Crick base pair configuration, see Figure 1.6 for clarification of groove positions) of the nitrogenous base. As an example, Figure 1.7 depicts both the Watson–Crick and Hoogsteen hydrogen bonding configurations for the A · T base pair. It should be noted that a reversed base pair occurs when one base is rotated 180° with respect to the other, and will not be distinguished further in this text.

Due to the variety of additional configurations allowed through Hoogsteen type hydrogen bonding, many more nucleic acid structures can be created. Other Hoogsteen type hydrogen bonding pairs include A · A, C · A, G · A, C^+ · G, G · G, and T · A, as illustrated in Figure 1.8. Hoogsteen hydrogen bonding has been experimentally

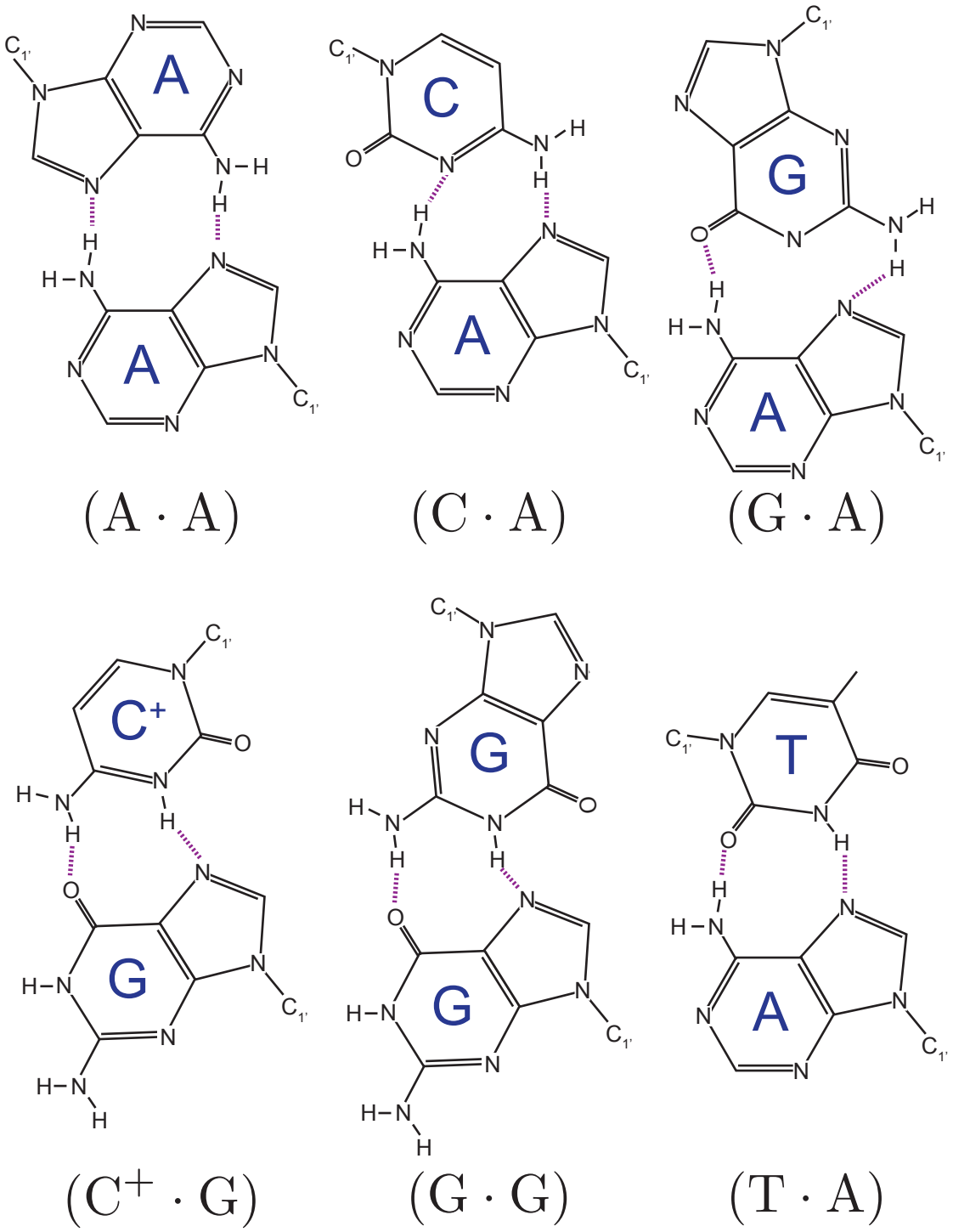

Figure 1.8: Some of the possible configurations of the Hoogsteen-type base pairings with the DNA nitrogenous bases. Similar configurations with uracil instead of thymine are possible in RNA molecules as none of the presented configurations rely on the thymine methyl group. All base-base interactions allow at least two hydrogen bonds.

determined in a variety of single-stranded DNA and RNA, for example it is found in rRNA [19], tRNA [12], and other oligonucleotides [20]. Hoogsteen hydrogen bond configurations have also been measured in canonical, double-stranded, B-DNA helices [21], and they are the primary method of triple-stranded DNA formation. The role of Hoogsteen bond formation in single-stranded (ssDNA), double-stranded (dsDNA) and triple-stranded (tsDNA) nucleic acids will be further discussed in Sections 2.3, 2.2, and 2.4.

Wobble Base Pairs

Additional nucleotide combinations are also possible; wobble base pairs were first put forth by Crick in 1966 to help explain the codon/anticodon amino acid coding system [22]. These additional base pairings include $A \cdot A$, $A \cdot G$, $C \cdot A$, $C \cdot T$, $T \cdot G$ and $T \cdot T$, as illustrated in Figure 1.9 [22]. Ionization of bases can provide further opportunities for hydrogen bonding between nucleotides including $A \cdot A^+$, $A^+ \cdot C$, $C \cdot C^+$, and $C^+ \cdot G$; the $C \cdot C^+$ base pairing is included in Figure 1.9 [13]. Wobble pairs are believed to be a critical element for higher order RNA folding and ssDNA interactions (like hairpins, to be discussed in Section 2.3) because they produce a unique local helical conformation and present a distinctive array of functional groups in the major and minor grooves of the duplex [23]. Evidence for the wobble base pair comes not only from the translation process that Crick first espoused, but has also been directly verified in many other systems [24, 25]. The $G \cdot T$ and $A \cdot G$ wobble base pairs are the most common non-Watson–Crick pairs in large single-stranded nucleic acids and complex triple-stranded structures [26].

The non-Watson–Crick pairs can occur singularly within a helical stem or as stacks of tandem or more base pairs forming intricate and recurrent motifs. As will be discussed in later sections, the flexibility, orientation, and stability of noncanonical hydrogen bonding can produce complex three-dimensional structures with unique functions comprised of DNA and RNA.

1.3.2 Base Stacking

In addition to hydrogen bonding, base stacking interactions occur between nitrogenous bases. In aqueous solutions, interaction of the flat planes of the bases (without the base-base hydrogen bonding described in Section 1.3.1) is favored over coplanar

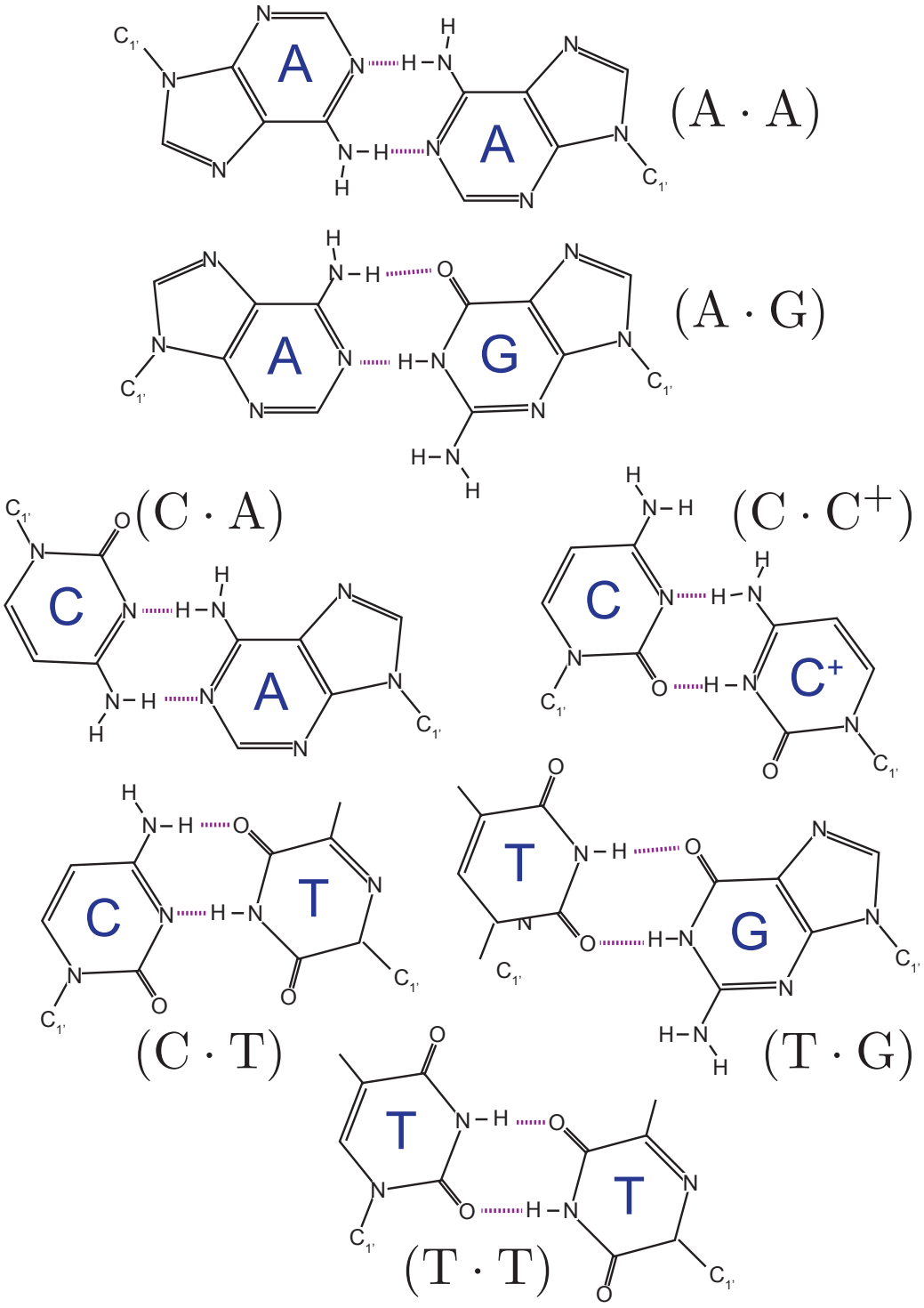

Figure 1.9: Some of the possible configurations of the wobble-type base pairings with the DNA nitrogenous bases. Similar configurations with uracil instead of thymine are possible in RNA molecules as none of the presented configurations rely on the thymine methyl group. All base-base interactions allow at least two hydrogen bonds.

interaction of the edges (with base-base hydrogen bonding). In non-aqueous solutions the reverse is true [13]. With base stacking as the dominant force in polar solutions, the stacking of two bases often contributes more than half of the free energy of the total base pair. This causes the bases to aggregate and form columns of nitrogenous rings [12, 13, 15].

Base stacking is a complex interaction that depends on several non-covalent forces, though the exact nature of these interactions are not well understood [13]. The forces stabilizing DNA stacking include: (i) induced dipole attractions, (ii) permanent electrostatic effects, and (iii) solvation effects.

Van der Waals dispersive forces (dipole induced dipole and induced dipole induced dipole attractions) certainly stabilize the stacking orientation of nucleic acids. Due to the large planar aromatic rings found in the pyrimidine and purine rings, van der Waals dispersive forces play a large role in the formation of DNA secondary structures. The dipole-induced dipole and induced-dipole-induced dipole forces are maximized when the large planar rings of the nitrogenous bases aggregate into stacked configurations. This kind of aromatic stacking commonly refers both to the forces that favor this geometry of the face-to-face juxtaposition of two aromatic molecules. In the large majority of all known DNA structures, the bases are in face-to-face contact [13].

Second, permanent electrostatic effects of interacting dipoles also undoubtedly influence stacking stability; this favorable or unfavorable contribution depends on the bond dipoles of the nucleotides. Electrostatic effects in DNA have received the greatest research focus [13]. This is likely because electrostatic effects are considered easier to model computationally than dispersive forces or solvation-driven effects. On average, it appears that permanent electrostatic effects are significant in influencing variations in stability for different base pairs and structures. This largely explains why stacking efficiency of DNA bases varies considerably depending on the neighboring bases [12]. However, on average, it appears that the electrostatic effects make a very small contribution in both random sequence and genomic sized nucleic acids. Thus, while permanent electrostatic effects can locally stabilize or destabilize a given base stacking interaction, and they affect the preferred geometries, it appears that on average the electrostatic effect is small in relation to the other components of base stacking [12, 13].

Finally, solvation effects also contribute to the stacking energy for nucleic acids.

This interaction depends on whether a nucleotide (most relevantly, its flat π-system surface) is better solvated by water or by a neighboring base's π-system surface [12, 13, 15]. In polar solutions, free nucleotides (those not bound with the phosphodiester bridge between phosphate and sugar groups in the backbone) will aggregate into columns to maximize the overlap of their π-systems. However, the degree of this contribution relative to van der Waals effects is still uncertain. Overall, much more study is needed to gain further insight of how the three forces — dispersive, electrostatic, and solvation-driven — cooperate to stabilize stacking for the nitrogenous bases of nucleotides.

Despite the many unknowns in the details of the base stacking interaction, the generalized behavior and overall geometry of nucleotides in polar solutions shows distinctive attributes. The bases are generally not directly aligned (to maximize surface area of contact and thus the overlapping π orbital systems) but are rather offset [12, 13]. This offset orientation may be favored by the preferred conformation of the backbone (if the nucleotides are joined by a phosphodiester bridge between the sugar and phosphate groups) and/or it may be favored by electronic effects in the bases themselves. Some support is seen that highly polar substituents such as NH_2, N, or O, of one base superimposed over the aromatic system of the adjacent base [12] may give rise to the observed offset. However, in general the distance between two aromatic planes is the van der Waals distance, ≈ 3.4 Å , independent of the structure or presence of a nucleotide backbone [12].

Intra-Chain Base Stacking

In the large majority of all known nucleic acid structures, the bases are in face-to-face contact with the contiguous bases on the same chain. In order to maximize surface area contact along a single DNA or RNA polynucleotide while allowing the backbone sugar and phosphate groups to take a favorable configuration, the nitrogenous bases stack atop one another with an offset helix orientation. This can be seen in the right-handed helical direction of the stacked guanine bases in the single-strand DNA of Figure 1.10.

As discussed previously, the permanent electrostatic effects (while overall impact may be small) are significant in influencing variations in stability for different base pairs and local structures. This largely explains why stacking efficiencies of nucleotides vary considerably depending on the neighboring base [12, 13, 15]. Often such pri-

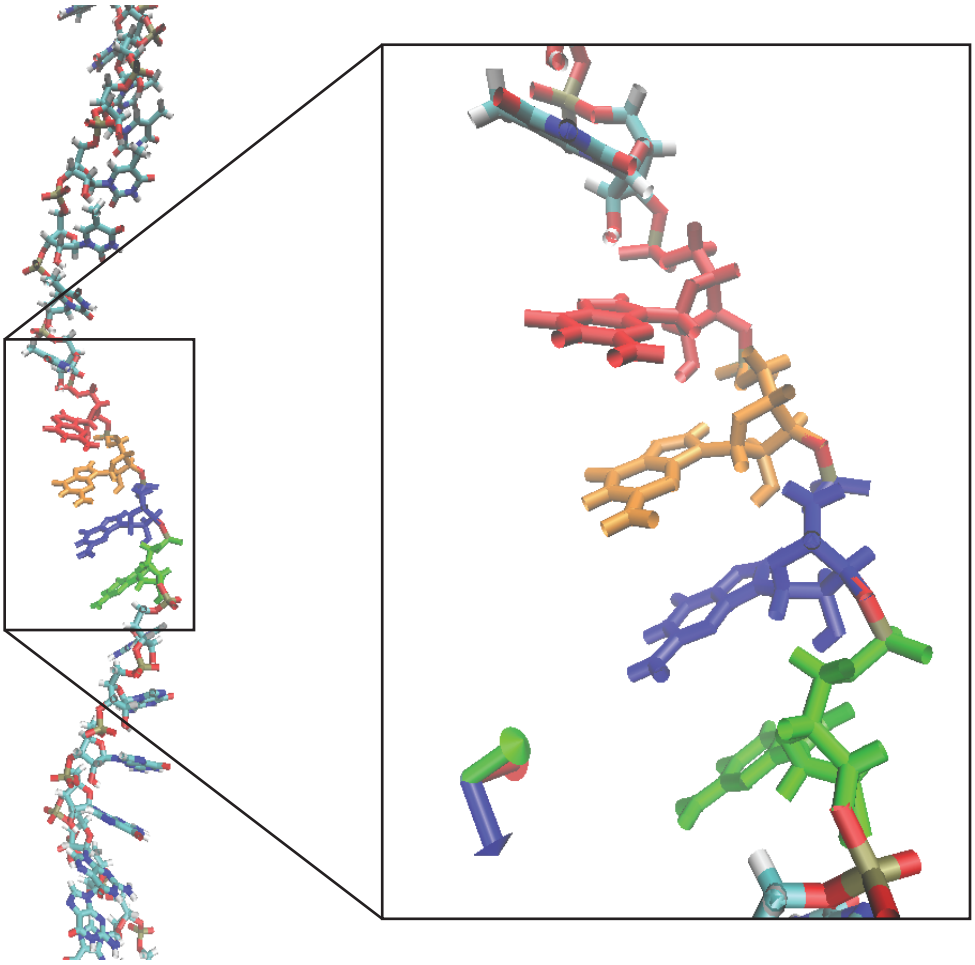

Figure 1.10: A single-stranded nucleic acid depicting intra-chain base stacking interactions. The single-stranded DNA molecule has formed a right-handed helix. Base stacking causes the planar rings of the purine base structures to self-organize in a parallel fashion through free rotation of the $C_{1'}-N_1$ single covalent bond. The inset shows four guanine nitrogenous bases with parallel purine rings; the four nucleotides are arbitrarily colored to facilitate understanding of the base stacking geometry.

mary structure (sequence) effects are neglected when considering chromosomal length nucleic acids, however, for short RNAs and DNAs, the local structure effects may produce considerable changes in the overall form and function of the molecule.

Recently a comprehensive set of data was generated for the intra-chain stacking of all four natural bases on both the 5' and the 3' sides of a nucleic acid polymer and with all four possible neighboring bases in order to understand the sequence dependent effect that stacking can contribute to local structure. These energies will be discussed in more detail in Section 5.1.3. In general, the purines stack more strongly than the pyrimidines due to the two electronic effects: London dispersion and permanent dipoles. These combine and lead to appreciable effects which are more pronounced in purines than in pyrimidine bases with the averaged trend [12]:

$$\text{purine} - \text{purine} > \text{pyrimidine} - \text{purine} > \text{pyrimidine} - \text{pyrimidine}.$$

In addition, a nucleotide's neighboring base can have a significant influence on the stacking energetics. This effect may be due to the differing electrostatic interactions between varied pairings of the two bases directly involved in the stacking interaction and because of variable polarizabilities of those neighboring bases. In aqueous solutions the thermodynamics of stacking of nucleosides has been measured by vapor phase osmometry [27], ultracentrifugation [28], calorimetry [29], and NMR [30]. For DNA, the stacking of the base on the 5' side is more favorable than on the 3' side most likely because the geometry of overlap is more favorable at that position. Interestingly, for RNA the reverse is true; the 3' stacking interaction is the more favorable [12].

Cross-Chain Base Stacking

Similarly to the base stacking interactions that can occur between contiguous bases on the same nucleic acid strand, cross-chain base stacking occurs between noncontiguous bases. The noncontiguous bases can either occur on the same strand (as in folded hairpin structures, see Section 2.3), or among different strands that are being held together with hydrogen bonding (such as double- or triple-stranded nucleic acid conformations). As with all stacking energies, cross stacking interactions approximately decay with the sixth power of the distances between any two nitrogenous bases [12], and therefore are only substantial between nearby base ring structures. Due to the shape and size of the individual bases (with the purines A and G having larger planar

surface areas), cross stacking is most significant between purines due to the possibility of significant nitrogenous base ring structure overlap in this configuration. However, due to geometric and electrostatic constraints, it is estimated that cross stacking will only have an average maximum energy of 10-15% of the similar base-base contiguous stacking energy [12, 13].

The physical features of nucleotides and nucleic acids discussed thus far help explain some of the characteristic properties and the fundamental, functional importance of these molecules in biological systems. The components, construction, and charge densities of each nucleotide along with the associative forces between bases, such as hydrogen bonding and stacking interactions, are greatly responsible for the formation of nucleic acid structures. The vast array of possible DNA and RNA formations will be discussed in the following sections.

2

Secondary and Tertiary Structures of Nucleic Acids

The century of biology upon which we are now well embarked is not a matter of trivialities. It is a movement of really heroic dimensions, one of the great episodes in man's intellectual history. The scientists who are carrying the movement forward talk in terms of nuclei-proteins, of biochemical genetics, of molecular morphology. But do not be fooled, this is the dependable way to seek a solution of the cancer and polio problems, the problem of rheumatism and of the heart, all the problems of the population. This is the understanding of life.

W. Weaver, *Science and Complexity*, 1948 [31]

2.1 Introduction to Nucleic Acid Secondary Structures

As the "century of biology" began, and after years of being diverted by the war effort, scientists in the late 1940s and early 1950s were once again able to address more fundamental scientific questions and focus on problems such as those affecting the nature of life. It was from this desire to tackle the great "problems of the population", as Weaver describes it [31], that spurred the research into the structure of DNA. By 1953 when Watson and Crick [6, 7] began their foray into DNA model building, there was already quite an assemblage of information definitively known or at least probably inferred about the molecule.

As was introduced in Chapter 1, the majority of the data concerning DNA was generated by Franklin [8], Wilkins [9], and their associates at King's College. Under the newly improved methods of Signer for extracting long unbroken molecules of DNA from cells [32], Wilkins began the process of drawing uniform fibers from a viscous solution of DNA. Under polarized light these molecules appeared well ordered, and thus, characteristic of long molecules oriented parallel to one another. The X-ray diffraction photographs taken of these first filaments showed hazy patterns that were later understood to be indicating helical features [32, 33]. This was the first direct evidence of the true nature of the DNA molecule; a long fibril wound in a helical shape. At the beginning of 1951 Rosalind Franklin joined the King's College researchers to continue the exploration of the DNA molecule by X-ray diffraction analysis and within the year she transformed the state of the field [32, 33]. By drawing thinner fibers and developing a micro focus X-ray tube, she was able to enhance both the alignment of the DNA molecules within the specimen and the quality of the diffraction patterns. However, it was not until she made a systematic study of DNA, by controlling the relative humidity, and thus the water content of the samples, that she was able to explain the variable features of the DNA samples previously examined. In fact, in what Franklin would later name A and B, she found that two forms of the DNA molecule existed with a transition from the "crystalline" to "wet" states at 75% relative humidity [32, 33]. These X-ray patterns of the A- and B-forms (the latter the so named B51 photograph) of DNA revolutionized the understanding of the molecule [8, 9]. The individual features and characteristics of several different forms of similar DNAs will be discussed further in Section 2.2.

In addition, from the X-ray crystallography studies conducted by Franklin [8, 33], several key structural features were deduced for both the A- and B-forms of DNA. From the reflection on the meridian it was understood that the molecule consisted of regular stacking of the nucleotide bases on top of each other and the number of units per turn could be calculated, even without knowing the details of the helix itself. In fact, the X-ray diffraction photographs were detailed enough to not only show the pattern for the nucleotide units but also depicted the secondary fans emanating from the two 0.34 nm meridional reflections, top and bottom, and running obliquely to the equator. These shapes are characteristic of a discontinuous helix, as is to be expected from the discrete moieties in a phosphate-sugar chain. Finally, from a series of density measurements, Franklin deduced that there were two or three chains of DNA per lattice point. The packing was determined to be pseudo-hexagonal, which implied that the molecules had an approximate cylindrical shape with a diameter

of about two nanometers [8, 9, 33]. With a detailed Patterson map of the molecule constructed in January 1953, Franklin determined that the phosphate groups of the backbone lay on the outside of the two, co-axial helical strands, arranged antiparallel to each other, with the bases arranged on the inside [8, 32–34].

With (some of) this structural information, Watson and Crick, along with Pauling, began to build models of DNA. Pauling proposed a three-chain structure with a central phosphate-sugar backbone and the bases extended radially on the outside of the structure [35, 36]. Unbeknownst to Pauling at the time, but understood by Watson and Crick, Franklin's measurements had already determined the two strand, phosphates on the outside, characteristics of DNA and thus Watson and Crick were able to dismiss Pauling's model as implausible based on Franklin's Patterson map analysis [8, 32, 33]. Watson and Crick then embarked on redesigning their original DNA model (one with three strands, but the bases on the inside) with all of the structural data determined by Franklin. With the chemical evidence of base complementarity of Chargaff [10] and determining the correct tautomeric forms of the nucleotide bases, Watson determined that when adenine and thymine and cytosine and guanine were paired the geometry of each was almost identical. Moreover each base pair could fit either way between the two chains while maintaining the symmetry found in the molecule by Franklin. With the model such defined, Watson and Crick had determined the most prolific structure of double-stranded DNA [6, 7].

In the nearly three quarters of a century since the discovery of the structure of the double helix, this simple description of the genetic material that regulates all life remains true and has not had to be appreciably altered to accommodate new findings. Nevertheless, we have come to realize that the structure of DNA is not quite as uniform as was first thought. Although the phosphate-sugar backbone is regular along the double helix, the different permutations of nitrogenous bases in the DNA sequence lead to local orientation differences and structural effects. Some DNA sequences even permit the double helix to twist in the left-handed sense [11, 13, 37, 38], as opposed to the right-handed sense of Watson and Crick's model and the majority of natural DNAs. In addition, not all genetic material in organisms is double-stranded as supposed by Watson and Crick; some small viruses have a single-stranded chromosome [11, 13] and both single- and triple-stranded structures are known as regulatory devices in several organisms [38–41]. DNA quadraplexes, for example in the G-quartet, are important structural features found in chromosomal telomeres and other promotional regions of a genome [11, 13, 42, 43]. Still additional complexity comes from

interactions of DNA molecules with each other, RNAs, and proteins [11, 13, 15].

Likewise to the structural discoveries of DNA, we now realize that RNA, which at first glance appears to be very similar to DNA, has its own distinctive structural features. It is principally found as a single-stranded molecule [11, 13, 15]. Yet by means of intra-strand base pairing, RNA exhibits extensive double-helical character and is capable of folding into a plethora of diverse tertiary structures [44–47]. These structures are full of surprises, such as non-classical base pairs, base-backbone interactions, and knot-like configurations. Most remarkable of all, and of profound evolutionary significance, some RNA molecules are enzymes that carry out reactions that are at the core of information transfer from nucleic acid to protein [48–51]. In recent years this enzymatic ability of RNA has been found to also exist within DNA [52–56].

Clearly, the structures of DNA and RNA are richer and more intricate than was first appreciated. Indeed, there is no one generic structure for nucleic acids and this helps explain the vastness of the characteristic properties and the fundamental, functional importance of these molecules in biological systems. As we shall see in this chapter, there are in fact, variations on common themes of structure that arise from the unique physical, chemical, and topological properties of the polynucleotide chain. We will continue our examination of the interplay of structure and function by examining the great complexities of nucleic acid secondary and tertiary structures: double-stranded nucleic acids (dsDNA and dsRNA) including a primer of dsDNA types, triple-stranded DNA (tsDNA), quadruple-stranded DNA (qsDNA), and single-stranded nucleic acids (ssDNA and RNA).

2.2 Double-Stranded Nucleic Acid Structure

Double-stranded nucleic acid molecules can assume a variety of secondary and tertiary structures. However, fundamentally, double-stranded nucleic acids consist of: (i) a regular two-chain structure, comprised of nucleotides joined with the phosphodiester bridges detailed in Section 1.2.3; (ii) hydrogen bonds formed between opposing bases of the two chains, illustrated in Section 1.3.1; and/or (iii) base stacking interactions occurring amongst contiguous bases, as described in Section 1.3.2. The double helical structure of a nucleic acid complex arises as a consequence of its secondary structure, and is a fundamental component in determining the tertiary structure of the molecule at large.

In nature, DNA is predominantly found as a duplex of two single-stranded DNA molecules. A number of factors account for the stability and preponderance of the double-helical structure of DNA. First, both internal and external hydrogen bonds stabilize the double helix. The two strands of DNA are held together by hydrogen bonds between complementary bases. In long range DNA double helices, these are predominantly the Watson–Crick base pairs of A · T and C · G as described in Section 1.3.1. The polar atoms in the sugar-phosphate backbone form external hydrogen bonds with surrounding solvent (such as water) molecules. Second, the negatively charged phosphate groups are all situated on the exterior surface of the helix in such a way that they have minimal effect on one another and are free to interact electrostatically with cations in the solution. Third, as detailed in Section 1.3.2, the core of the helix consists of the base pairs, which, in addition to being hydrogen bonded, stack together.

The two chains of the double helix are usually arranged in an antiparallel fashion (that is one chain runs 5' to 3' while the complementary strand is 3' to 5') to each other, and this arrangement is due to the stereochemical consequence that the sugars in the respective nucleotides have opposite orientations [11, 12]. The polar sugar-phosphate backbone of the two chains are on the outside of the double helix with the bases stacked in the interior. As is discussed in Section 1.3.2, the heterocyclic bases stack and are located on the inside of the complex as a consequence of their π-electron clouds, hydrophobicity of their planar rings, and their geometry [12, 13]. The interactions of the charged phosphates along the backbone, the hydrogen bonding and stacking between the bases, and the steric effects of all atoms in the duplex affect the axial stiffness or persistence length of the molecule. Since dsDNA in solution does not take a rigid structure but is continually changing conformation due to thermal fluctuations and collisions with solvent molecules, and with elastic motions on the time scale of nanoseconds, the classical measures of rigidity are ill-suited to this application. Instead, the persistence length, or the length of DNA over which the time-averaged correlations in the orientation of the polymer are lost, is used to describe DNA stiffness [57, 58]. Both the primary structure and overall secondary structure can alter the persistence length of a double helix significantly. Local sequence effects can change rotational angles of the bonds comprising the polynucleotide backbone due primarily to the sequence dependent variations in the base pair stacking. The consequence of these local interactions can be both gentle and sharp bends in the dsDNA fiber. However, when these local variations are summed over the great length of a DNA molecule, the net result is a double helix, randomly coiled in a spherical shape. A

general estimate of the stiffness of dsDNA is 46 - 50 nm or 140 -150 base pairs [13].

When duplex DNA molecules are subjected to conditions of temperature, pH, or ionic strength that disrupt their hydrogen bonds, the strands are no longer held together. That is, the double helix is denatured and the two strands separate to form single-stranded DNAs. If temperature is the denaturing agent, the double helix is said to melt. Denatured DNA will renature to re-form the duplex structure if the denaturing conditions are removed (that is, if the solution is cooled, the pH returned to neutrality, or the denaturants are diluted out). Renaturing requires reassociation of the DNA strands into a double helix, a processed termed reannealing. For this to occur, the strands must realign so that their complementary bases are once again within hydrogen bonding distance of each other [11, 13, 14, 59].

Wound into a spiral shape, the DNA double helix can be characterized by some select geometries: the major and minor groove sizing, the diameter, the bases per turn, rise, and handedness. First, a fundamental attribute of a double helix is the size of its major groove and minor groove, with the major groove usually being wider than the minor groove. The two grooves, or spatially accessible regions, can also be thought of as void helices that wind around the two nucleotide chains. In a top down examination of each Watson–Crick base pair, shown in Figure 1.6, the glycosidic bonds holding the bases in each base pair are not directly across the helix from each other, this causes the sugar-phosphate backbones of the helix to be dissimilarly spaced along the helix axis and thus the grooves are unequal in size. Instead, the intertwined chains create a major groove and minor groove with the minor groove between the C_1,-glycosidic bonds of each base pair. The major groove lies at 180° from the minor groove and contains more donor and acceptor constituents (and thus functional groups) than the minor groove [11, 13, 15]. The electron donor and acceptor locations of the bases, as illustrated in Figure 1.5, accessible to the major groove, along with its relative greater size, give it unique potential for additional interactions [11, 13–15]. One example of these additional interactions will be detailed in Section 2.4.

Second, the diameter of the DNA double helix is a measure across the cross section of the DNA fiber. For Watson–Crick hydrogen bonding between a purine and a pyrimidine, the diameter is practically constant regardless of which nitrogenous base is on each strand [6, 7, 11, 13, 14]. This leads to a regular diameter of the long double helix with little local sequence effect. However, depending on the secondary structure of the duplex, the double helix's average diameter can vary greatly. Further, for non-

canonical base pairings, the primary structure (sequence) can have a large effect on the diameter of the molecule.

Additional geometric features can be used to describe dsDNA. The number of bases in a complete rotation of the spiral shape, or bases per helical turn, along with the rise, or distance between nucleotides, describe the packing density of the nucleotides in the helix. Finally, the handedness, either a right-handed or left-handed helix, of the DNA chain is vital in describing its shape. Although the geometry of a nucleotide or base pair can be completely characterized by six coordinates: rise, shift and slide (for diameter), tilt and roll (for bases per turn), and twist (for handedness), as these values precisely define the location and orientation in space of every base or base pair in a nucleic acid molecule relative to its predecessor along the axis of the helix [12, 13, 15]. We will use instead, the previously defined characteristics for a generalized description of each helical structure which encompass the specific descriptors in a more intuitive manner.

At least three dsDNA conformations are believed to be found in nature: A-DNA (Section 2.2.2), B-DNA (Section 2.2.1), and Z-DNA (Section 2.2.3) [6–8, 12–15, 37, 59]. However, other DNA conformations can be generated with particular environmental factors or stresses; the sugar-phosphate groupings that constitute the backbone are inherently flexible and the nitrogenous bases are rife with hydrogen bonding sites to promote stabilizing interactions allowing numerous other configurations. To continue the nomenclature that Franklin [8] began by naming her first two structures of DNA, A-DNA and B-DNA, almost all of the alphabet has been used to describe different types of DNA secondary and tertiary structures. Only the letters, F, Q, U, V, and Y are now available to describe any new DNA structures that may appear in the future. Most of these additional forms have been created synthetically and have not been observed in naturally occurring biological systems [12, 13, 60].

Of the naturally occurring dsDNA secondary structures, the B-DNA form is believed to predominate in the cellular environment [6–9, 12, 13]. A-DNA and Z-DNA differ significantly in their geometries and dimensions to B-DNA, although still form the standard helical structures that we have previously described. The A-DNA form appears likely to occur in dehydrated samples of DNA, the hybrid pairings of DNA and RNA strands, and when a double helix is formed between two RNA molecules [8, 11–14]. Segments of DNA that have been methylated for regulatory purposes, along with particular pyrimidine-purine repeating patterns and protein-DNA complexes have been found to adopt the Z geometry [11–14, 37]. The dimensions of each of

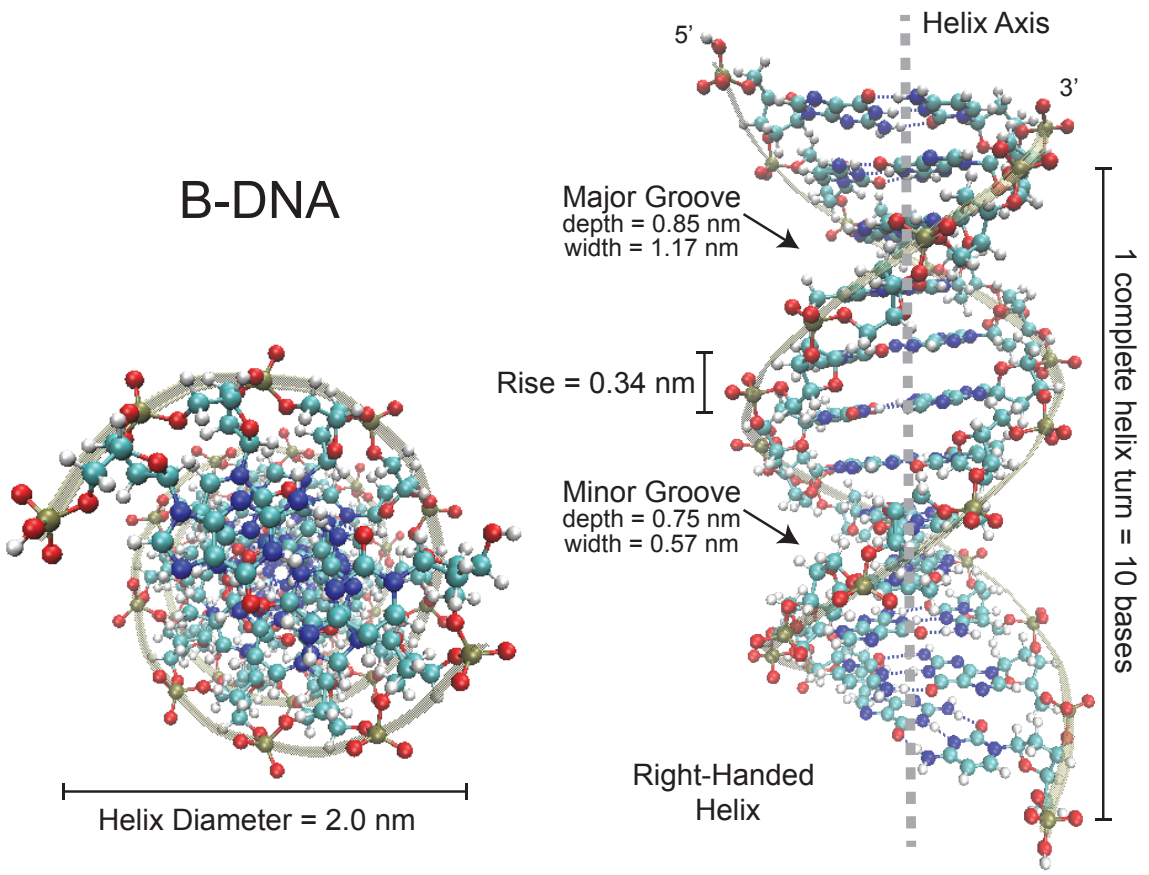

Figure 2.1: The B-form of dsDNA has several distinct features. The end projection, with the helix axis into the page (left illustration), shows that the molecule has a tightly packed core of stacked nitrogenous bases with an outer ring of sugars and finally the phosphate groups comprising the surface. The side projection (right illustration) depicts the right-handedness of the complex. Note that the bases are nearly perpendicular to the helix axis (drawn as (gray) dashed line to aid reader). In addition, both the major and minor grooves along with the rise and helical turn are shown. The 5'- and 3'-ends of the duplex are labeled. The atoms are colored as follows: phosphates (gold), oxygens (red), carbons (teal), nitrogens (blue), and hydrogens (white). A (gold) ribbon connects the phosphate groups along each backbone and is included to aid the reader.

the aforementioned characteristics of DNA are detailed in Sections 2.2.1, 2.2.2, and 2.2.3. It is important to note that primary structure, or local sequence, can have a significant effect on many dsDNA secondary sequences; some of these effects are described in Sections 2.2.6.

2.2.1 B-Form DNA

Watson and Crick described the B-DNA structure in their 1953 model [6, 7]. As can be seen in Figure 2.1, it is characterized as a right-handed double helix with between

10 and 10.6 bases per turn, depending on local sequence. One complete rotation around the spiral (pitch) occurs every 3.4 nm and the distance between neighboring base pairs (rise) is 0.34 nm. The nitrogenous bases are nearly perpendicular to the helical axis. The diameter is 2.0 nm; a more detailed cross-sectional measurement is the distance of the phosphate atom from the helical axis, which measures 0.94 nm. The major groove is approximately 50% wider than the minor groove and is the primary location of B-DNA interaction with other DNAs, RNAs, and proteins. The major groove depth and width is 0.85 and 1.17 nm, respectively. The minor groove depth and width is 0.75 and 0.57 nm, respectively [12, 13]. The geometric descriptors of B-form DNA are summarized in Table 2.1 to ease comparison with dsDNA's other structures.

2.2.2 A-Form DNA

An alternate form of the right-handed double helix is shown in Figure 2.2 as A-form DNA. A-DNA molecules differ in a number of ways from the B-form DNA described in Section 2.2.1. In general, the B-form of DNA is longer and thinner than the short, squat, and more tightly wound A-DNA form. The pitch, or distance required to complete one helical turn has shrunk from the 3.4 nm in B-DNA to 2.46 nm, a 25% macroscopic shrinkage in the length of the fiber. One turn on A-DNA requires 11 base pairs to complete. In A-DNA, the base pairs are no longer nearly perpendicular to the helix axis, but instead are tilted 19° with respect to the centerline axis, depicted in Figure 2.2. Successive base pairs occur every 0.29 nm along the axis as opposed to 0.34 nm in B-DNA [11–13].

Although relatively dehydrated DNA fibers (Franklin found a 75% relative humidity needed to transition from B- to A-form DNA [8, 33]) can be shown to adopt the A-conformation under physiological conditions, it is unclear whether DNA ever assumes this form *in vivo* [13]. However, double-helical DNA:RNA hybrids (dsDNA:RNA) and double-stranded RNA (dsRNA) exhibit an A-like conformation, see Section 2.2.7. The 2'- OH group in RNA sterically prevents the double helical regions of RNA chains from adopting the B-form helical arrangement. Consequently, double stranded regions involving RNA chains assume an A-like structure with their bases strongly tilted with respect to the helix axis. It should be noted that right-handed dsDNA in solution has 10.5 base pairs per turn and a structure that lies between A- and B-DNA forms, but closer to B-DNA [12–15]. The geometric descriptors of A-form DNA are summarized

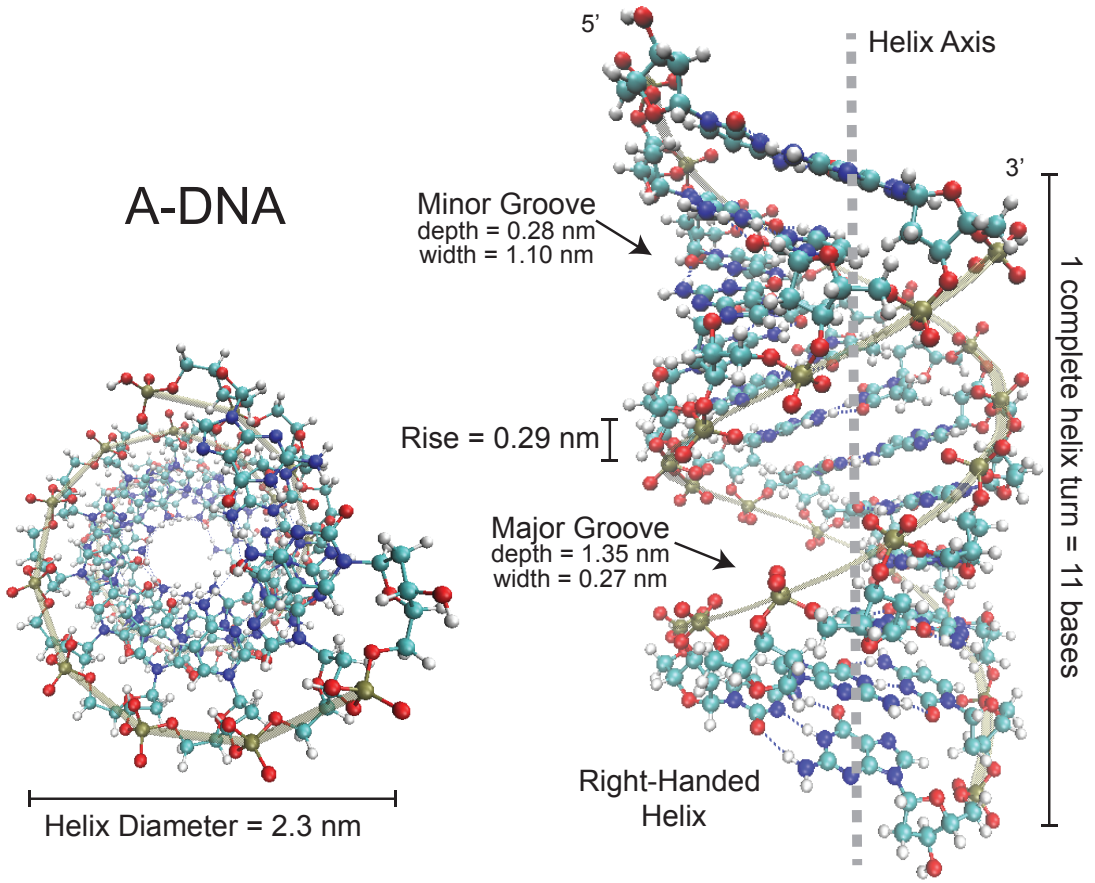

Figure 2.2: The A-form of dsDNA has several distinct features. The end projection, with the helix axis into the page (left illustration), shows that the molecule has a open core with the nitrogenous bases forming an inner ring, surrounded by the sugar and phosphate groups intermixed in an outer ring. The side projection (right illustration) depicts the right-handedness of the complex. Notice that in A-DNA the base planar rings are strongly tilted towards the helix axis, depicted as the (gray) dashed line. The characteristic geometries comprising the major and minor grooves, rise, and bases per turn are shown. The 5'- and 3'-ends of the duplex are labeled. The atoms are colored as follows: phosphates (gold), oxygens (red), carbons (teal), nitrogens (blue), and hydrogens (white). A (gold) ribbon connects the phosphate groups along each backbone and is included to aid the reader.

in Table 2.1 to ease comparison with dsDNA's other structures.

2.2.3 Z-Form DNA

Named for its characteristic "zig-zag" backbone, as can be seen in Figure 2.3, Z-form DNA is distinctively different from either A- or B-DNA as described in Sections 2.2.1 and 2.2.2. Most notably, Z-DNA is left-handed and has a structure that repeats every two base pairs, instead of repeating every one base pair as in A- and B-DNA

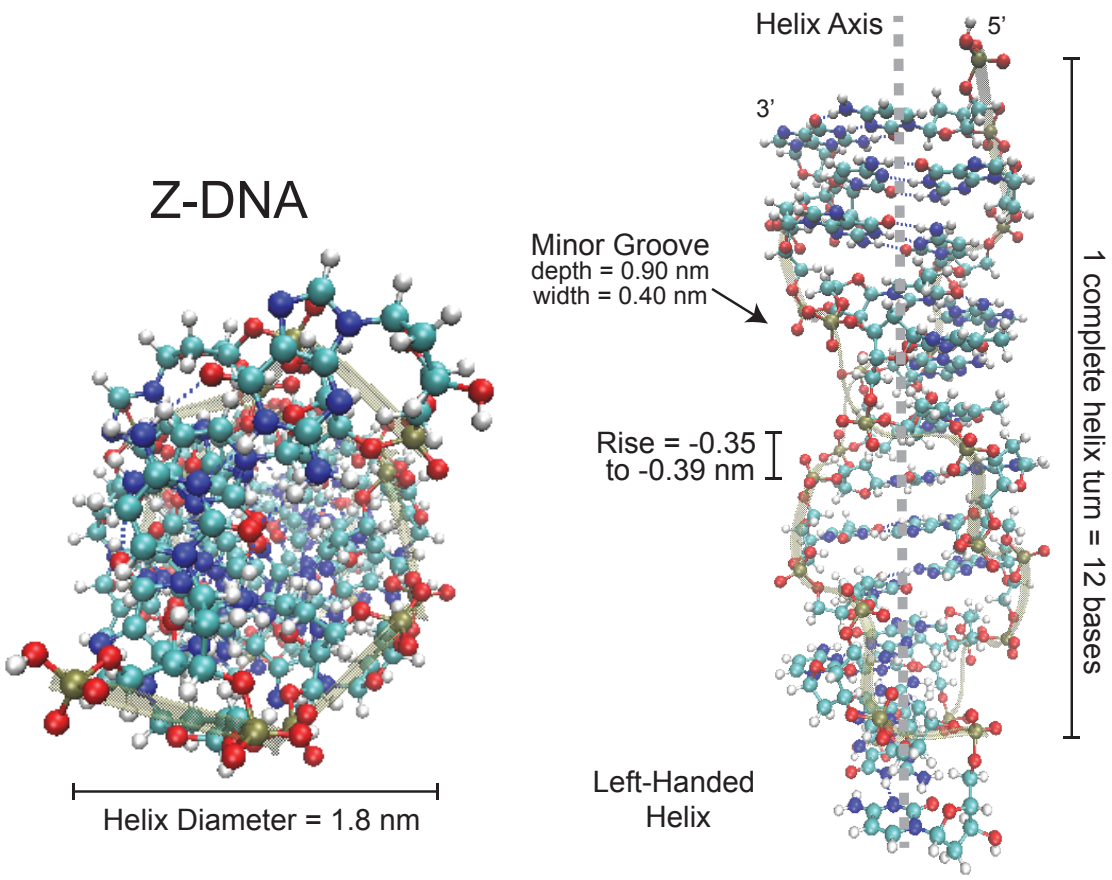

Figure 2.3: The Z-form of dsDNA is considerably different from the right-handed A- and B-form DNAs. The end projection, with the helix axis into the page (left illustration), shows that the molecule is left-handed with a close packed interior core. The side projection (right illustration) demonstrates the "zig-zag" nature of the backbone (from which the structure gets its name) and is traced in a (gold) ribbon to aid the reader. The base planes are nearly perpendicular (-6.2° off perpendicular) to the helix axis sketched as the (gray) dashed line. Z-DNA does not have a major groove; the major groove is filled with the cytosine C_5 and guanine N_7 and C_8 atoms. The characteristic geometries comprising the minor groove, rise, and bases per turn are shown. The 5'- and 3'-ends of the duplex are labeled. The atoms are colored as follows: phosphates (gold), oxygens (red), carbons (teal), nitrogens (blue), and hydrogens (white).

[11–15]. Although Z-DNA is thought to occur naturally, formation of this structure is generally unfavorable, however certain conditions can promote it. As examples, an alternating purine-pyrimidine sequence (especially consisting of only guanine and cytosine), negative supercoiling, or high salt and some cations (all at physiological temperature and pH) can induce a transition from B- to Z-form. The transformation from right-handed B-DNA to left-handed Z-DNA has been experimentally measured with an alternating guanine-cytosine (poly(dGdC)), but has not been observed thus far with a similar adenine-thymine nucleotides [12]. Although the Z-DNA conforma-

tion has been difficult to study because it does not exist as a stable feature of the double helix, it was the first discovered by single crystal X-ray diffraction methods and subsequently rediscovered from fiber diffraction studies [12, 13]. In addition, Z-RNA has been demonstrated by NMR [12]. It is believed that Z-DNA is a transient structure that is occasionally induced by biological activity and then quickly disappears *in vivo* [12, 13].

The alternating pyrimidine-purine sequence of this oligonucleotide is the key to its unusual properties. The N-gylcosyl bonds on the guanine residues in this alternating copolymer are rotated 180° with respect to their conformation in B-DNA, so that the purine ring is over the deoxyribose ring (*syn*- conformation) rather than the usually favored *anti*-conformation. Due to the fact that the guanine nitrogenous ring is flipped, the cytosine planar ring must also flip to maintain normal Watson–Crick base pairing. However, pyrimidine nucleotides do not readily adopt the *syn*-conformation because it creates steric interference between the pyrimidine C_2-oxy substituent and the pentose ring. Since the pyrimidine ring does not rotate relative to the pentose structure, the entire cytosine nucleoside (base and sugar) must rotate 180°. The alternating *anti*-pyrmidine/*syn*-purine arrangement in each strand favors a conformation transition that realigns the sugar-phosphate backbone along the zig-zag course. It is topologically possible for the guanine to go *syn* and the cytosine nucleoside to undergo the 180° orientation without breaking and reforming the three G · C hydrogen bonds; the transition from right-handed to left-handed can occur without disruption of bonding relationships among the atoms involved. The proposed method of transition from B- to Z-DNA involves flipping the bases over one at a time, while maintaining Watson–Crick pairing, in a cavity produced by longitudinal breathing (local stretching). The cavity propagates down the helix after base pair flipping; the cooperatitively of the experimental transition is explained [12, 13, 61]. Sections of Z-DNA can form within longer B-DNA double helices with a pair of B-Z junctions on either end of the transient regions. In a B-Z junction box, a base pair is extruded (flipped out) from the cylindrical double helix (no intra- or inter-chain hydrogen bonding or stacking) which may act as a biological marker and functionalized target. It is also believed that Z-DNA regions within longer, chromosomal, B-DNA, double helices provide torsional strain relief while DNA transcription occurs [11–14].

In general, Z-DNA is more elongated and slimmer than B-DNA. Z-DNA has two chains arranged antiparallel; one complete helical turn spans 4.56 nm and comprises 12 base pairs. The diameter of the molecule is 1.8 nm with a phosphate to helix axis

distance between 0.62 and 0.77 nm [12, 13]. The geometric descriptors of Z-form DNA are summarized in Table 2.1 to ease comparison with dsDNA's other structures. There is another form of Z-DNA, called Z(WC)-DNA with many of the same characteristics of Z-DNA such as being left-handed and having a zig-zag backbone, but also having Watson–Crick backbone directions. Energetically this intermediate structure may explain many of the open questions concerning the B- to Z-DNA structural transition; however, it has not yet been verified by crystallographic studies. Nevertheless, Z-DNA stands as another important and unique structure possible in double-stranded DNA.

2.2.4 P-Form DNA

As was discussed earlier in this chapter, there is a vast palette of possible dsDNA structures, so many that only five letters remain (with the current nomenclature classification) to describe any newly discovered conformations [60]. We will not endeavor to outline all of the possible duplex configurations, but wish to include a few additional (and some non-natural) DNA arrangements. One member of this DNA alphabet soup is Pauling or P-DNA. Here, we do not refer to Pauling's original DNA model of three strands with the bases exposed and extended radially on the outside of the molecule [35, 36] (also called P-DNA), but instead to a two oligonucleotide structure with similarly arranged external bases that has been named in homage to Linus Pauling [60, 62, 63]. P-DNA is believed to be produced as a result of relatively weak supercoiling and topological constraints; such that is probably encountered during replication and transcription (where positive supercoiling is produced downstream of the protein complex) or during recombinational repair where the ends are constrained and writhing is surprised [11, 13, 14, 59, 62].

DNA supercoiling, with degree σ, is involved in gene regulation because locally unwound DNA is necessary for transcriptional activation and recombinational repair. Since the proteins are only acting on a small segment of a larger DNA molecule, there are additional restrictions limiting the behavior near these sites: linking number is constant (sum of twist and writhe) and writhe is suppressed by the pulling forces of the proteins (which can be up to 14 pN). As a consequence, pulling on the DNA molecule increases the effective torque applied [59, 62]. The pulling forces in P-DNA do not overstretch the molecule with the transition at ≈ 3 pN for overwound DNA, but instead limit the writhe. As an example, local denaturation of DNA has been ob-

served in plasmids unwound by $\sigma \approx$ -0.07. By stretching the molecule and preventing its writhing, denaturation is already observed at $\sigma \approx$ -0.015 [62].

Twisting a DNA molecule, which is unable to writhe, can occur in one of the two directions: (i) unwound molecules, $\sigma < 0$ and (ii) overwound molecules, $\sigma > 0$ [59]. Here we will consider starting with the B-form of DNA, see Section 2.2.1 for geometric details, and will make all comparisons to this structure. For unwound molecules, with $-1 < \sigma < -0.015$, the torque is relieved by a local denaturation of the DNA: for every helical turn of bases denatured, one turn of unwinding is released. For overwound molecules, with $0.037 < \sigma < 3$, the torque is also relieved by the local formation of a new DNA structure: for every helical turn converted to the new structure, three turns of overwinding are released. This new structure, P-DNA, has approximately 2.62 bases per helical turn and an extension 75% larger than B-DNA [62].

As twisting increases on a B-DNA structure, stretching of the phosphodiester bridges forces the stacked base pairs against one another. Beyond $\sigma \approx 1$, the backbones resist further extension, the Watson–Crick hydrogen bonds are broken, and the bases are expelled from the double helix. This allows the backbones to move to the center of the molecule. Twisting can continue until $\sigma \approx 4$, although the energy cost is quite high past a supercoiling degree of 3.5. Around $\sigma = 3$, the optimal structure is only $\approx$ 60% longer than the original B-DNA structure, but the length can be modified easily between $\approx$ 25 and 80% with some variations as a function of base sequence. The persistence length for P-DNA is approximated at 19 nm [62]. It is important to note that the resulting P-DNA conformation can be reached by starting from either B- or A-form DNA.

In this conformation, the phosphate groups are diametrically opposed around the helical axis, and their anion oxygens point outwards and are fully accessible for stabilizing interactions with counter ions. There are also important inter-strand phosphate-sugar stabilizing interactions. The expelled bases are relatively free to rotate. Depending on the sequence, both stacking and hydrogen bonding between adjacent bases can occur (with some purines taking a *syn* conformation to favor such interactions). It should also be noted that very similar left-handed P-DNA can be created by strong negative twisting; these structures are reached after passing through a completely unwound state and strongly resemble their right-handed analogues. The geometric details of P-DNA are included in Table 2.1 to aid comparison to the other types.

Table 2.1: A summary of geometric descriptors of B-, A-, Z-, P-, and S-form dsDNA.

Property	B-DNA	A-DNA	Z-DNA	P-DNA	S-DNA
Helix Handedness	right-handed	right-handed	left-handed	both (depending on sign of supercoiling - values given for right)	right-handed
Base Pair per Repeating Unit	1	1	2	1	1
Base Pair per Helix Turn	10 to 10.6	11	12	2.62	37.5
Rise per Base Pair	0.34 nm	0.29 nm	-0.35 to -0.39 nm	0.585 nm	0.40 nm
Base Pair Inclination (off perpendicular to helix axis)	2.4°	19°	-6.2°	free rotation	highly inclined
Diameter	2.37 nm	2.55 nm	1.8 nm	0.69 to 1.6 nm	1.4 nm
P Distance from Helix Axis	0.94 nm	0.95 nm	0.62 to 0.77 nm	0.55 to 0.60 nm	
Glycosidic Bond Orientation	*anti*	*anti*	*anti* for CG step *syn* for GC step	both	
Major Groove Depth	0.85 nm	1.35 nm	convex	-	shallow
Major Groove Width	1.17 nm	0.27 nm		-	
Minor Groove Depth	0.75 nm	0.28 nm	0.90 nm	-	narrow
Minor Groove Width	0.57 nm	1.10 nm	0.40 nm	-	

2.2.5 S-Form DNA

When the pulling forces greatly exceed those used in the formation of P-DNA and are no longer simply eliminating writhing in the molecule, but instead are inducing stretching, an additional conformation the overstretched, or S-DNA, conformation can be reached. Extensional experiments on the double-stranded B-DNA have shown that the overstretching transition occurs when the molecule is subjected to stretching forces of 65 pN or more [62–64]. The DNA molecule thereby increases in length by a factor of 1.8 times the normal contour length. Above 150 pN (for a random sequence) S-DNA denatures into characteristic ssDNA molecules.

S-DNA can take two conformations depending on whether the ends are allowed to freely rotate. If the ends of the molecule are unrestricted, it will be "ladder like" and can be considered an unwound helix. This unwinding leads to an elevation of the S-DNA compared to B-DNA and may allow easier access to the base pairs for transcription. If, instead, the ends are not free to rotate (rather it is a section of dsDNA in a longer molecule, as in P-DNA), then the molecule undergoes stretching due to the high force acting on it. The elongated DNA is characterized by a strong base pair inclination, a narrow minor groove, and a diameter roughly 30% less than that of B-DNA. The base pairs, which are exposed on the major groove side of the double helix are still bound by a single hydrogen bond, and strong inter-strand stacking between the bases is seen. The conformational change occurs progressively and cooperatively during stretching [63, 64]. As pulling force is applied to a B-form dsDNA, regions of B-, P-, and S-DNA can all form [65].

Extensions of dsDNA molecules, with both the P- and S-forms, are important aspects of the winding and subsequent unwinding of the B-form DNA that must occur during transcription and replication. Genetic recombination, which requires the protein RecA and triplex formation, discussed in Section 2.4, is thought to extend and unwind dsDNA as an intermediate step. RecA induces pulling forces and a 1.5 times extension of B-DNA to aid in this process [64, 65]. Although the conformation of S-DNA is not as well understood as the others, its geometric characteristics are included in Table 2.1 to aid in comparison.

2.2.6 Other Structures in dsDNA

The previous conformations of dsDNA discussed are all variations sharing a common secondary structural theme, the double helix. In this context, the DNA is assumed to be in a regular, linear form. However, DNA can also adopt regular structures of higher complexity (and outside helical geometries). By relaxing the rules governing its local structure in small pockets along a longer, regular double helix, DNA can adopt a much wider variety of structures.

Palindromic sequences within DNA are DNA base sequences that are inverted repeats of each other and have the potential to form a tertiary structure known as a cruciform. Cruciform configurations exist when the normal inter-strand hydrogen bond base pairing is replaced by intra-strand hydrogen bonding. In effect, each DNA strand folds back on itself in a hairpin structure to align the palindromic sequences. The structural features of a DNA hairpin and single-stranded cruciform structures are discussed in Section 2.3. Such cruciforms are never as stable as normal DNA duplexes due to the unpaired segment in the loop region. However, negative supercoiling can cause a localized disruption of hydrogen bonding between base pairs in DNA and may promote the formation of cruciform loops. Of biological significance, cruciform structures have a two-fold rotational symmetry about their centers and can potentially create distinctive recognition sites for specific DNA binding proteins [12–15, 59].

Local sequence effects are not limited to additional hydrogen bonding conformations, but can also include structural changes within a standard B-DNA form. For example, the poly-A tract (when there are more than four contiguous adenines) induces local bending due to the presence of several, contiguous adenosine residues. The adenosine nitrogenous rings stack well and cause each base to tilt with respect to the helical axis. Due to this extreme tilting of the bases, the junction of the poly-A tract and the regular, random sequence, B-DNA is not smooth and yields a global curvature to the DNA with a larger angle at the 3' than the 5'-end of the A tract. The overall bend for a poly-A tract is approximately 20°. Chains of poly-A tracts are found in phase with each other down the length of the double helix (that is, lengths of poly A are equally spaced along the double helix with units of 10 nucleotides between them, such as A_5N_{10} where N is any nucleotide) which allows quite drastic global bend angles to form [12, 13, 66, 67]. The poly-A tract, as with the cruciform structures, is believed to be involved in transcription regulation and site recognition [11, 13, 15].

Table 2.2: A summary of geometric descriptors of B-DNA, A-RNA, and A'-RNA.

Property	B-DNA	A-RNA	A'-RNA
Helix Handedness	right-handed	right-handed	right-handed
Base Pair per Repeating Unit	1	1	1
Base Pair per Helix Turn	10 to 10.6	11	12
Rise per Base Pair	0.34 nm	0.273 to 0.281 nm	0.30 nm
Base Pair Inclination (off perpendicular to helix axis)	2.4°	16 to 19°	10°
Diameter	2.0 nm	1.9 nm	2.0 nm
P Distance from Helix Axis	0.94 nm	0.87 nm	0.93 nm
Glycosidic Bond Orientation	*anti*	*anti*	*anti*
Major Groove Depth	0.85 nm	very deep	very deep
Major Groove Width	1.17 nm	and narrow	and narrow
Minor Groove Depth	0.75 nm	wide and	wide and
Minor Groove Width	0.57 nm	shallow	shallow

2.2.7 Forms of dsRNA

Depending on their biological function, naturally occurring double-stranded RNAs either display long, double-helical structures or they are globular, with short double-helical domains connected by single-stranded stretches. Double-helical domains can, in many cases, be predicted from the primary nucleotide sequence and special computer algorithms have been developed for this purpose [68]. Short double helices are found in tRNA, in ribosomal RNAs, in globin mRNA, and in many of the genes of bacteriophages. In viruses, the RNA structures can include pronounced, well-developed double helices [12, 13].

RNA double helices display two major, structurally similar conformations, depending on the salt concentration of the environment, much as was seen in the A-DNA and B-DNA structural transition at 75% relative humidity. At low ionic strength, the A-RNA double helix, with 11 bases per helix turn predominates [12, 69]. If the salt concentration is raised in excess of 20%, A-RNA is transformed into A'-RNA with a 12-fold helix. Both A- and A'-RNA structures exhibit features typical of Watson–Crick base pairs [13]. The polynucleotide chains are arranged antiparallel and form a right-handed double helix. Because the base pairs are displaced 0.44 nm from the helix axis, a very deep major groove and a rather shallow minor groove are created. This extremely skewed helix spacing aids in the formation of additional structures [12, 69]. The principle difference between A- and A'-RNA are in the pitch heights, about 3.0 nm for A-RNA but 3.6 nm for A'-RNA. The axial rise per residue in A-RNA,

0.273 to 0.281 nm, is smaller than for A'-RNA, 0.30 nm, a difference that is reflected in the base pair tilt angle of 16° to 19° in A-RNA and 10° in A'-RNA. Otherwise, the nucleotide conformation in both A- and A'-RNA are the same [12, 13]. The geometric descriptors previously used are listed in Table 2.2 for A-RNA and A'-RNA and compared with the standard B-DNA values for a double helix. When a hybrid molecule of DNA and RNA is formed, the A-RNA structure is observed [12, 69].

Double-stranded nucleic acids, in their common natural forms, are uniquely suited to be able to reliably transmit, express, and conserve the vast genetic information needed by an organism. Due to the long-range, regular, and sequence independent features, along with the inherent stability of the duplex, the canonical B-form DNA and A-form RNA genomic molecules can be billions of bases long and yet be accessible, organized, and compact enough to be stored in each and every cell of an organism. The additional secondary and tertiary structures found in these long dsDNA and dsRNA molecules (such as the Z-form, P-form, S-form, cruciforms, and poly-A tracts) are transitory in nature, responding to the local structure or stresses on the double helix and providing unique markers in an otherwise regular fiber. However, for shorter, more functionalized, and more active nucleotide molecules, many other configurations are possible. These include single-, triple-, and quadruple-stranded nucleic acids.

2.3 Single-Stranded Nucleic Acid Structure

Nucleic acid secondary structure is generally divided into helices, as was previously described in Section 2.2, where the base pairs are numerous and contiguous, and various kinds of loops where there are stretches of unpaired nucleotides. The double helix is an important tertiary structure in nucleic acid molecules which is intimately connected with the molecule's secondary, or base pairing, structure. When the duplex is denatured into its single-stranded components, there are a wide variety of other configurations possible based on its primary and bonded secondary structures.

When a single-stranded nucleic acid folds back on itself and is able to base pair, it forms a stem-loop or hairpin configuration. The bonded bases comprise the stem

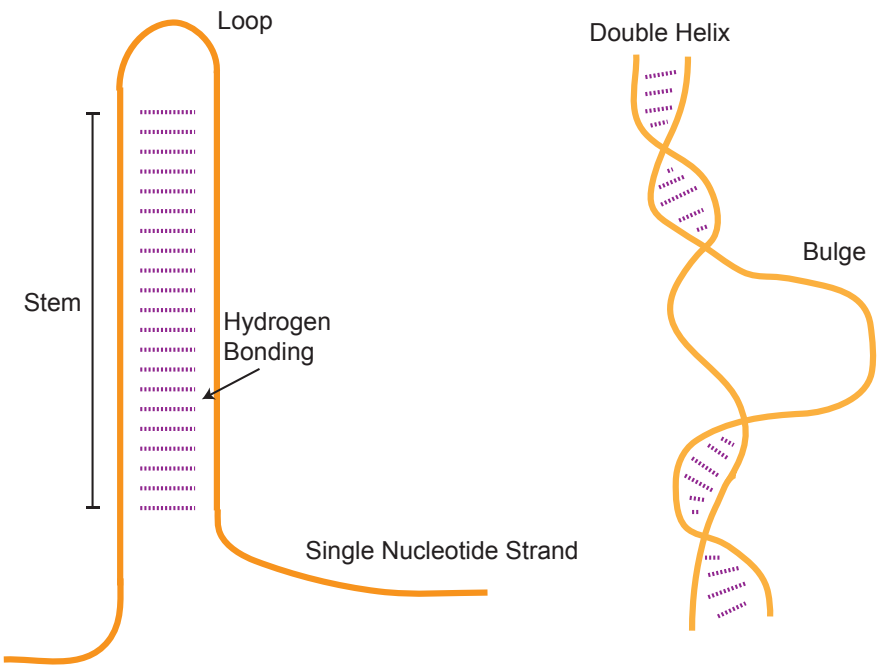

Figure 2.4: The single-stranded nucleic acid structural components of a stem-loop and bulge. A stem-loop configuration occurs when a single nucleic acid chain folds back on itself forming a hydrogen bonded stem region and an unbounded loop region. A bulge interrupts a regular double helix when there are unpaired nitrogenous bases on one strand that do not correspond to any nitrogenous bases on the other strand of the duplex. The backbone of each ssDNA is represented by the (orange) line and the hydrogen bonding is shown as the (purple) dashed lines.

of the molecule while the unpaired bases fold into the loop structure. The stem-loop structure is extremely common and is a building block for larger structural motifs. Internal loops (short series of unpaired bases in a larger paired helix) and bulges (regions in which one strand of the helix has additional inserted bases with no counterparts in the opposite strand) are also frequent structural components [11, 13, 15]. These structural elements are illustrated in Figure 2.4.

In fact, it has been newly proposed that ssDNA secondary structure formation of hairpins may cause significant issues for a variety of utilizations including sequencers, sensors, and other antisense applications [70, 71]. The next generation Illumina/Solexa type sequencers use a ensemble of ssDNA probes ligated to a chip surface and then free, fluorescent dyed, nucleotide terminators are added and the fluorescence signals are read and recorded. The ssDNA probes are short fragments of genomic DNA; it is possible for reverse complementary sequences of exist within these ssDNA probes [72]. As was discussed previously, such sequences have the ability to fold back on themselves and form stem-loop or hairpin structures. Moreover, the secondary structures

of single-stranded nucleic acid sequences do not necessarily consist of a single stretch of Watson–Crick complementary sequences. Some non-Watson–Crick base pairs may form, and the secondary structure may even include unpaired regions, yet these sequences may still provide large overall free energy stabilization. In a next generation sequencer such secondary structures would stop the ssDNA probe from fluorescing, and thus the original fragment of genomic DNA from being properly read. Not only would such systematic error bias the sequencing data for recognizing single nucleotide polymorphisms (SNPs), but also significant errors throughout a genome could accumulate [71].

Similar to the persistence length described in Section 2.2 for double-stranded DNA, the persistence length for single-stranded DNA and RNA can be measured. Experiments on the more stable molecule, DNA, have found values of 0.75 nm via mechanical stretching [73], 1.3 nm utilizing atomic force microscopy [74], 1.4 nm from thermal melting profiles [75], 1.76 nm and 1.82 nm with sedimentation experiments [76], 1.5 - 3.0 nm with fluorescence spectroscopy [77], 2.0 - 3.0 nm via transient electrical birefringence [78], and 3.1 - 5.2 nm with fluorescence recovery after photobleaching [79]. The persistence length of ssDNA seems to vary widely due to a variety of factors including the length of the sequences examined i.e., long ($\gg$ 100 nucleotides) and short ($<$ 100 nucleotides), the model used to examine the data (freely jointed chain or wormlike chain), and the concentration and type of buffer used in each experiment. It is believed that RNA exhibits similar flexibility, though direct experimental measurements are rare. Although no one value is agreed upon, the flexibility is much greater in ssDNA than in dsDNA.

Single-stranded nucleic acids lack of prominent double-stranded (helical) structure along with its high degree of flexibility and ability to bond with a wide variety of hydrogen bonding patterns, as described previously and in Section 1.3.1, give these molecules the unique ability to fold into a wealth of diverse formats. Due to the absence of consistently base-paired nucleotides, base stacking interactions, as described previously in Section 1.3.2, are especially important in single-stranded secondary and tertiary structure formation. In fact, the medley of structures is so vast that ssDNA and RNA are ideal candidates as functional molecules for *in vivo* and *in vitro* applications.

2.3.1 Aptamers

Aptamers are functional molecules comprised of short strands of nucleotides that can, as Crick described, do "a very neat job in a small space" [80]. Fundamental to their function, aptamers exhibit strong binding to a specific target (other nucleic acids, proteins, small organic compounds, or even entire organisms) [50]. It is the ability of these molecules to bind to a specific target molecule with a high degree of specificity and selectivity that makes aptamers quite useful as biological agents.

Aptamers range in size from approximately 10 to 100 nitrogenous bases and sometimes have complex three-dimensional structures, produced by a combination of Watson–Crick and non-canonical intramolecular interactions [81]. They bind to their targets with a dissociation constant, K_D, typically in the low nano-molar range. Aptamers can distinguish enantiomers of small molecules or minor sequence variants of macromolecules frequently with K_D ratios spanning several orders of magnitude [50, 82]. In a striking example of specificity, an aptamer to the small molecule theophylline (1,3-dimethylxanthine) binds with 10,000-fold lower affinity to caffeine (1,3,7-trimethylxanthine) that differs from theophylline by a single methyl group [50]. They are typically composed of RNA, single-stranded DNA, or a combination of these with non-natural nucleotides [82].

Theoretically it is possible to select aptamers virtually against any molecular target; aptamers have been selected for small molecules, peptides, proteins, as well as, viruses and bacteria [50, 82, 83]. Aptamers are isolated from extremely complex libraries of nucleic acids, generated by combinatorial chemistry, by an iterative process of adsorption, recovery, and re-amplification. These libraries are usually comprised of a pool of oligonucleotides (usually of chains less than 100 bases) of 10^{10} to 10^{20} sequences [84, 85]. Additional sequence variation can be introduced at each cycle and the process becomes an *in vitro* paradigm of Darwinian evolution. This protocol, called systematic evolution by exponential enrichment (SELEX), is generally used with modification and variations for the selection of specific aptamers. After sufficient enrichments, aptamers can be cloned and studied as homogenous sequence populations [50, 82, 84–86]. Using this process, it is possible to develop new aptamers quickly and for unique targets. Aptamers are usually created synthetically, but natural aptamers also exist [50, 82, 87].

In addition to the genetic information (primary structure) encoded by the nucleic acids, aptamers also function as highly specific affinity ligands by molecular inter-

action based on their three dimensional folding pattern [50, 81, 82, 88]. The three dimensional complex shape of a single-stranded oligonucleotide is primarily due to the base-composition led intra-molecular hybridization that initiates folding to a particular molecular shape. This molecular shape assists in binding through shape specific recognition to its targets leading to considerable three dimensional structure stability and thus the high degree of affinity [50, 82, 89].

Aptamers can be used to analyze the natural processes of nucleic acid – protein recognition, to generate inhibitors of enzymes, hormones and toxins with potential pharmacological uses, to detect the presence of target molecules in complex mixtures and to generate lead compounds for medicinal chemistry [50, 81–83, 87, 90–92]. Their advantages over alternative approaches include the relatively simple techniques and apparatus required for their isolation, the number of alternative molecules than can be screened, and their chemical simplicity. Disadvantages of aptamers include their pleomorphism, their high molecular mass, and the restricted range of target sites that appear to be suitable [50, 82].

The very first paper describing directed *in vitro* evolution of nucleic acids described the isolation of a totally new enzyme, a nuclease composed of DNA [93–95]. The likelihood was thereby opened up that it was possible to develop any form of useful new catalyst that could be conceived and for which a selection procedure could be devised. To an extent, this has been realized, with the isolation of both DNA and RNA enzymes that can cleave and ligate DNA or RNA [53, 56, 96–102], can form and cleave amide bonds and alkylate halogenated peptides [103], have oxidative and peroxidase activity [50, 82], and can have a carbon–carbon bond forming activity [50, 82]. These enzymatic nucleic acids, called RNAzymes and DNAzymes, will be discussed further in Sections 2.3.2 and 2.3.3.

2.3.2 RNAzymes

Due to its enormous structural and dynamic flexibility, single-stranded RNA molecules can exhibit many of the structural features of a classical enzyme, such as an active site, a binding site for a substrate, and a binding site for a cofactor, such as a metal ion [49, 52, 93, 96]. Such RNA molecules are known as RNA enzymes, catalytic RNA, ribozymes, or RNAzymes.

Like the traditional protein based enzymes, RNAzymes act in similar ways. As with

all catalysts, enzymes work by lowering the activation energy for a reaction and thus dramatically accelerating the rate of the reaction [59]. An enzyme's characteristics (complementary shape, hydrophilic/hydrophobic interactions, stereospecificity, regioselectivity, and chemoselectivity) specify the reactions that it catalyzes. Since enzymes are extremely selective for their substrates and speed up only a few reactions from among many possibilities, the set of enzymes present determines which metabolic pathways occur [11, 14]. Proteins, which dominate the enzyme field and were once considered the sole biological enzymes, are evolutionarily suited to act as catalysts in nature; almost all processes *in vivo* need enzymes in order to occur at biologically relevant time scales [11, 14, 49].

In the early 1980s, the perception of protein-only enzymes was challenged by the revelation that certain natural RNAs could function as enzymatic machines to mediate biological catalysis [48, 49, 52, 104]. The discovery of ribozymes has profoundly altered the view of how life might have evolved. For example, in the RNA World hypothesis, RNA once functioned as both the genetic material and the enzymatic machines of life [49, 104]. According to this theory, the genetic coding ability of DNA and the catalytic ability of proteins have evolved from a RNA genome and RNA enzymes, respectively. The natural RNA enzymes discovered are believed to be the remainder of this evolutionary process. The RNA World theory has gained support as more natural ribozymes are discovered that perform the most basic of functions *in vivo*. Other natural RNAzymes have been shown to possess significant capacity for catalyzing various other chemical reactions, prompting researchers to take advantage of the potential structural complexity of RNA to generate novel RNA species that have specific desirable enzymatic properties for *in vivo* and *in vitro* applications [49, 104].

Before long, natural ribozymes were being supplemented by novel artificial ribozymes that catalyzed an even broader range of chemical transformations [53]. With the success of both natural and novel ribozymes, questions concerning DNA's innate ability to perform as a catalyst arose. However, at present, no naturally occurring enzymes have been found to be composed of DNA.

2.3.3 DNAzymes

As noted above, the general impression of DNA's potential as an enzyme was shaped by its natural role as a double-helical molecule used to store genetic information.

The repetitive structure of double-helical DNA restricts its catalytic potential by prohibiting the formation of more complex secondary and tertiary structures. Structural sophistication is a prerequisite for catalytic function, as is commonly observed with ribozymes and protein enzymes. However, DNA can be synthesized as a single-stranded polymer and, therefore, like single-stranded RNA, has the freedom to form higher-ordered and elaborate structures.

Preliminary evidence showed single-stranded DNA possessed adequate flexibility for complex tertiary structure formation [50]. In addition, demonstrations were conducted showing that a large fraction of a particular RNAzyme, the hammerhead ribozyme, could be replaced with deoxyribonucleotides without significant loss of catalytic activity [50, 105]. However, without a natural deoxyribozyme to study, little effort in the mid 1990's had been spent on establishing DNA's ability to form structures or to enhance chemistry relative to either RNA or proteins. This question was ultimately answered in 1994 when the first DNA enzyme was created by *in vitro* selection techniques [52]. The SELEX protocol, as described previously in Section 2.3.1 is a powerful and yet simple technique that has been routinely used to isolate extremely rare DNA or RNA sequences with a function of interest from extraordinarily large populations of single-stranded DNA or RNA molecules [84–86].

The primary structure, or the sequence of bases, comprises the configuration of a DNAzyme. The configuration is constant for a deoxyribozyme but the conformation depends on the environment (such as temperature, pH, salt concentration, or interactions with other molecules). It is vital that the conformation is dynamic; deoxyribozymes and their substrates fold and bend in order to induce fit and catalyze reactions. The secondary structure of the DNAzyme is comprised of two major components: a catalytically active loop and complementary base pairing substrate binding arms, as illustrated in Figure 2.5.

The catalytic core consists of a sequence of nucleotides that aids the desired reaction, such as cleavage or ligation, on the substrate molecule. Usually forming a loop of some shape, the catalytic function occurs at the central unpaired region of the substrate. Deoxyribozymes, like anti-sense oligonucleotides, can be designed to bind to any nucleotide molecule simply by following the rules of Watson–Crick base pairing. In the molecule, two binding arms are attached to the catalytic loop. These arms are designed in such a way so that they complementarily bind to the substrate nucleotides. Both the length and sequence of these chains can be used to exactly match the desired catalytic site. The successful fusion of the catalytic activities of the core loops with

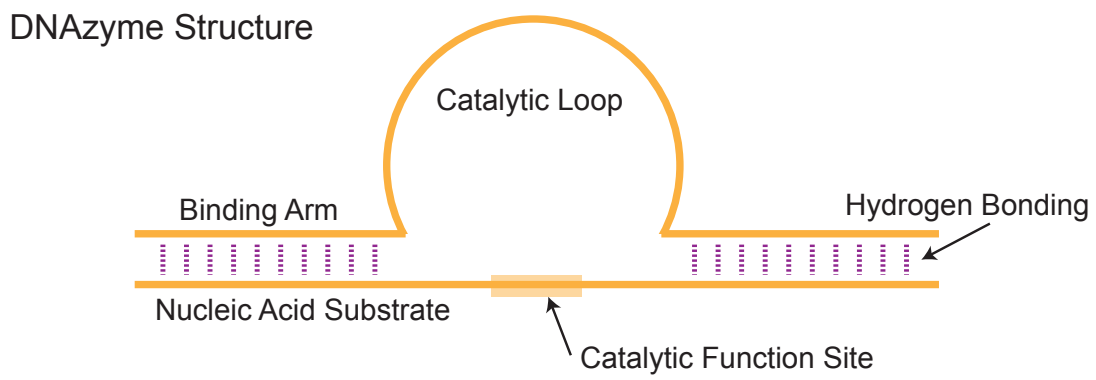

Figure 2.5: The DNAzyme secondary structure is comprised of a catalytic loop or core of the molecule and two substrate binding arms. The binding arms hydrogen bond (typically with Watson–Crick hydrogen bonding) with a nucleic acid substrate. The binding arms are typically 7 to 12 nucleotides (here 10 bases) in length and can be tuned for the specific substrate of interest. The catalytic function site is located in the unpaired base region of the nucleic acid substrate. The backbone of each single-stranded nucleic acid is represented by the (orange) line and the hydrogen bonding is shown as the (purple) dashed lines.

the target recognition capabilities of the arms produces deoxyribozymes whose utility can be customized for uses from basic biotechnology applications to advanced forms of therapeutics [106–111].

Thousands of enzymatic nucleic acids have been found through the forced evolution process and examined in laboratory settings. Of all of these molecules, the 10-23 DNAzyme is probably the most studied. Named for its origin as the 23^{rd} clone of the 10^{th} cycle of *in vitro* selection, it can cleave almost any RNA molecule at a purine – pyrimidine junction in a biological cofactor range (i.e. Mg^{2+}). The cleavage occurs by increasing the rate of the background hydrolytic degradation reaction that makes RNA inherently unstable [96, 112]. It cleaves A−U and G−U sites with very high proficiency and A−C and G−C sites with reduced efficiencies [55, 56]. The ability of the 10-23 DNAzyme to cleave at such positions means that the AUG start codon of any gene can be used as a target [113–115]. Due to this versatility, the 10-23 DNAzyme has been investigated as a potential tool for *in vitro*, cell culture, and *in vivo* applications [51]. The particular and far reaching applications will be discussed in Chapter 7.

The core of the 10-23 deoxyribozyme is composed of only 15 nucleotides, flanked on each side by a substrate-binding arm of seven to ten nucleotides that bind to the RNA target via Watson–Crick base pairing, as shown in Figure 2.6. This simple structure permits easy alteration of the substrate specificity to generate precise cleavage agents

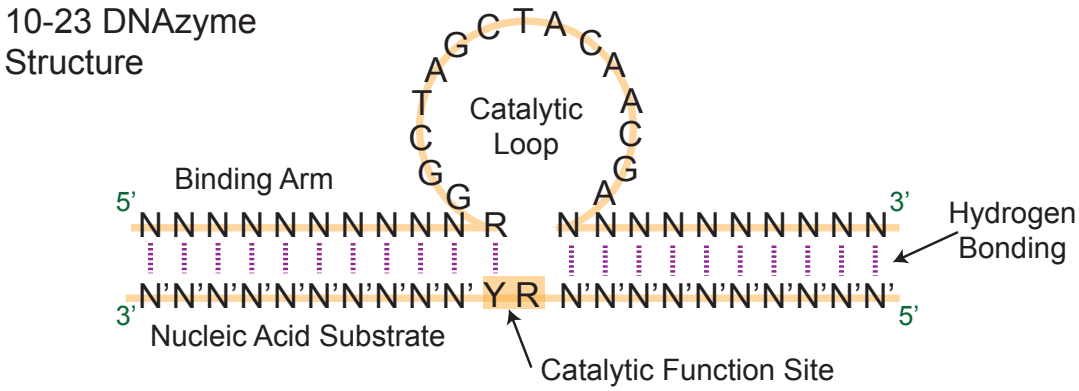

Figure 2.6: The 10-23 DNAzyme secondary structure is comprised of a catalytic loop or core of the molecule and two substrate binding arms. The binding arms hydrogen bond with a nucleic acid substrate. The catalytic core is 15 nucleotides and is a semi conserved sequence across variants. The binding arms are typically seven to ten nucleotides (N) in length and can be tuned for the specific substrate of interest. The catalytic function site is located in the unpaired base region of the nucleic acid substrate and cleavage occurs between a pyrimidine (Y) and purine (R). The backbone of each single-stranded nucleic acid is represented by the (orange) line and the hydrogen bonding is shown as the (purple) dashed lines.

for almost any RNA sequence. The chemical stability, high catalytic proficiency, mismatch discrimination, and the ease of synthesis of DNA have made the 10-23 DNAzyme an attractive alternative to ribozymes for site-specific cleavage of biological RNA targets [113–115]. Such advantages to the 10-23 DNAzyme have made it a much studied molecule both by experimental and simulation approaches; further discussion of this aptamer is detailed in Chapter 7.

With the discovery of the first deoxyribozyme, DNA joined RNA and proteins as the building blocks of powerful biocatalysts [52–54, 95, 96, 98–100, 116, 117]. Although it had been excluded from natural evolution, the ability of DNAzymes to be specifically designed and created in the laboratory opened an entirely new field of biological molecular function. This function is inherently rooted in the structure and thus sequence of the single-stranded nucleic acids comprising aptamers. Further complex structures involving multiple nucleic acid strands are described next in Sections 2.4 and 2.5.

2.4 Triple-Stranded DNA Structure

The generalized double helix can, under certain conditions, accommodate a third strand in its major groove. This can occur with both DNA and RNA molecules, but we will concentrate on DNA [12, 13, 69]. A DNA triplex is formed when pyrimidine or purine bases occupy the major groove of the DNA double helix forming Hoogsteen, reverse Hoogsteen, wobble, or reverse wobble hydrogen bonds, as described in Section 1.3.1 and Section 1.3.1, with the Watson-Crick bonded nitrogenous bases comprising the regular double helix. Examples of common triple-bonded nitrogenous base geometries and bonding sites are illustrated in Figure 2.7. Triple-stranded helices can form with three individual oligonucleotides bound with inter-strand hydrogen bonding, between two strands (where one is in an intra-strand hairpin configuration), or with the one nucleic acid exhibiting intra-strand hydrogen bonding with itself twice. Two of the segments are parallel with each other while the third is antiparallel to the first two. Triplex structures have been found to play important roles in gene regulation, DNA repair, and site specific modification or cleavage [13].

The possible base triples are not limited to those depicted in Figure 2.7, and experimentally pH, and thus protonation, have greatly altered the possible combinations [118]. The greater wealth of base hydrogen bonding combinations, each with their own local structure due to steric effects, significantly increases the complexity of the base stacking and hydration interactions that further stabilize the helical structure. Although there is little empirical and experimental base stacking data available for tsDNA, these interactions are believed to be strong and vital to triple-stranded DNA stability [12, 13].

However, even when complementary base pairing and stabilizing stacking interactions exist, there are still many other factors that effect the feasibility of triplex formation beyond the basic considerations, outlined previously in Section 2.2, for duplex formation. The structure of a duplex or a tract within a duplex determines the feasibility of the acceptance of a triplex forming oligonucleotide into the major groove [12, 13, 69]. A deep major groove along with a slightly unwound (more base pairs per helical turn) structure, as exists in A-RNA, A'-RNA, A-DNA and B-DNA, make suitable candidates for triplex formation. In addition, divalent counter ions are also important in the charge screening of the phosphate groups and are probably involved in site-specific interactions to counterbalance the high linear charge density formed in triplexes. Finally, for intramolecular triplexes, such as H-DNA structures discussed

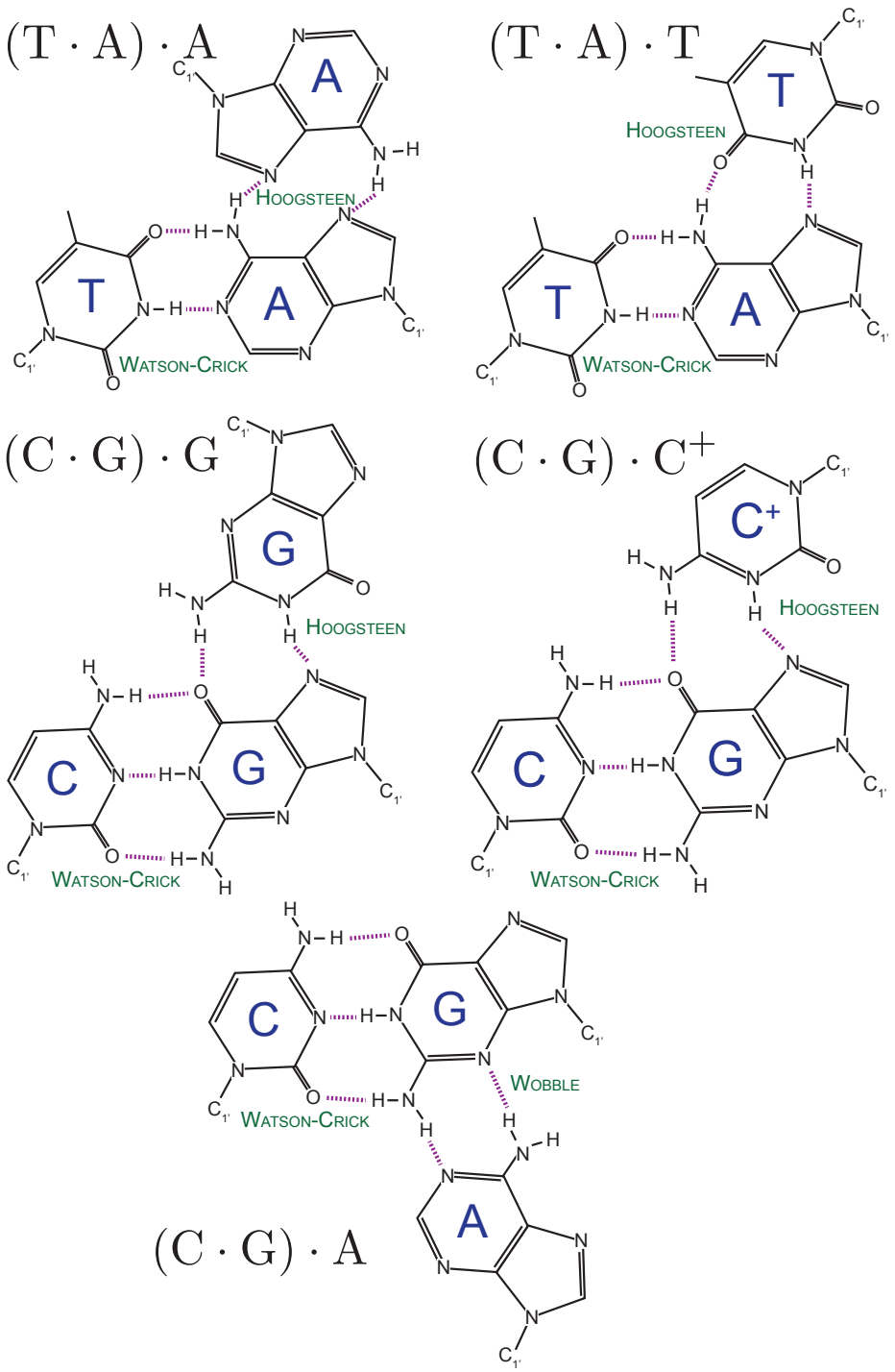

Figure 2.7: Schematic illustration of some of the possible base configurations comprising triple-stranded DNA. All base-base interactions allow at least two hydrogen bonds. The bases stabilized by Watson–Crick hydrogen bonding interactions are denoted with the parentheses with the additional Hoogsteen or wobble hydrogen bonding base indicated after this pair, as labeled above. The hydrogens not explicitly involved in the hydrogen bonding have not been included for clarity's sake. Additional triples are possible; protonation of bases, such as the $(C \cdot G) \cdot C^+$ triple, greatly increases the possible hydrogen bonding sites.

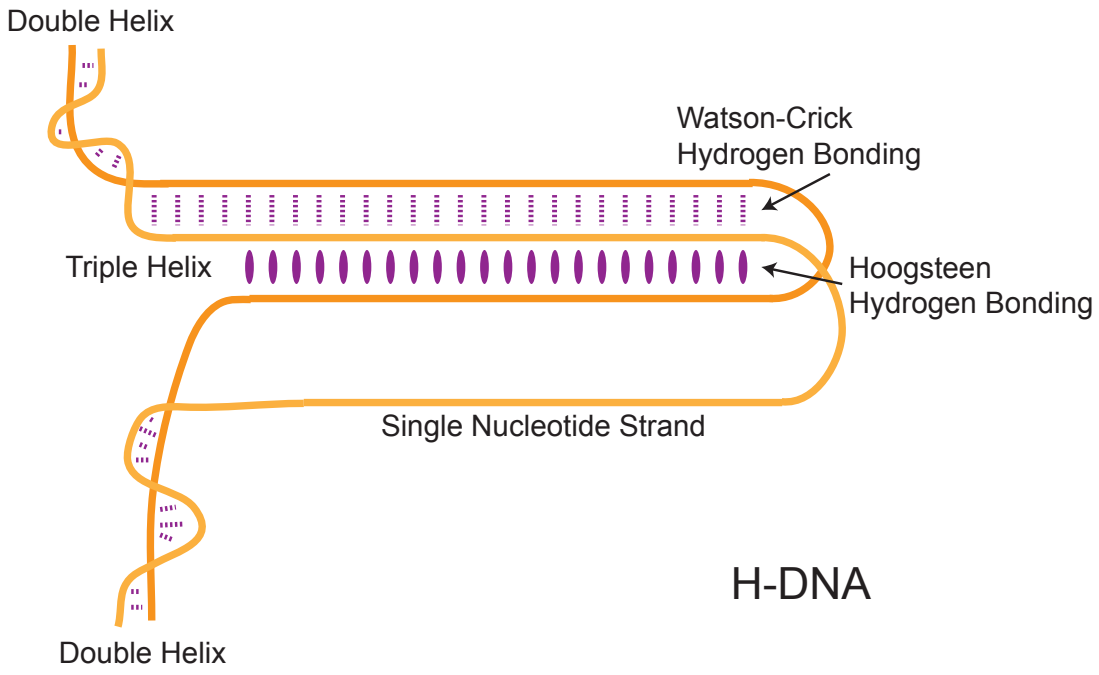

Figure 2.8: The schematic structure for H-DNA. A section of the double helix denatures, breaking the Watson–Crick hydrogen bonds. One strand of the denatured section remains as a single nucleotide strand, as labeled above. The other strand has a palindromic sequence to that of the nearby double helix and winds its way into the major groove. A triplex formation arises as Hoogsteen hydrogen bonds are formed between the Watson–Crick duplex and the invading strand. The Watson–Crick and Hoogsteen hydrogen bonds are shown as (purple) dashed lines and (purple) ellipses, respectively. The two solid (dark orange and light orange) lines outline the backbone of the two nucleotide strands.

below, suitable sequence and negative superhelical density (slightly unwound helical structure) are critical for the complex formation.

2.4.1 H-DNA

Although the canonical Watson–Crick double helix is the most stable DNA conformation for an arbitrary sequence under usual conditions, some sequences within duplex DNA are capable of adopting structures quite different from the canonical B-form. H-DNA, named for the hydrogen ions that stabilize it, is formed when negative superhelical stress is applied on specific sequences contained in larger dsDNA [41]. In order to relieve the negative supercoiling stress on the molecule, the duplex tract kinks and folds in the middle while twisting and unpairing several of the base pairs. This structure will quickly re-anneal into its regular dsDNA structure unless sufficient hydrogen bonding can occur between one of the now single-strands and a nearby region of the

dsDNA duplex can occur. Thus, the overall H-DNA structure is contained within a stretch of dsDNA but has both a triple-stranded region and a single-stranded region, as illustrated in Figure 2.8.

2.4.2 Strand Invasion for Repair

Organisms have developed several DNA repair mechanisms to cope with DNA damage. While some types of DNA damage are primarily repaired by the action of a single, specific repair pathway, most types of damages are repaired by more than one pathway. One such example is how triplex formation plays an important role in the repair of a stalled replication fork or a break in dsDNA [11–13]. For *in vivo* systems, the process is quite complex since the strand invasion is aided by proteins, such as RecA or Rad51 [11, 13, 14, 59, 119]. However, we can consider the simplest structure of this phenomenon as the process consisting of a dsDNA molecule and an invading ssDNA strand, where the ssDNA possesses the same sequence as one of the strands of the dsDNA. At the end of the process, the ssDNA is wrapped inside the major groove of the dsDNA and the complex is stabilized by a combination of Watson–Crick and Hoogsteen bonds [11–14, 59, 119, 120]. Large sections of error prone DNA can be corrected when the third strand replaces the erroneous section of double helix through a series of protein catalyzed reactions, binds to the complementary strand, and ligates to the nicked ends of the strand.

2.5 Quadruple-Stranded DNA Structure

Tetrameric DNA structures consist of four nucleic acid strands wound together. This can occur in one of two ways: (i) with four individual strands twisted as if an additional chain has been added to a triple-stranded DNA molecule and the base pairs are fully intercalated, or (ii) as two base-paired parallel-stranded duplexes are intimately associated with their base pairs fully intercalated. In the first case, the chains are all antiparallel to their nearest neighbors, in the second case the relative orientation of the duplexes is antiparallel [121]. The first type of tetrameric nucleic acid structure is characteristic of the G-quartet.

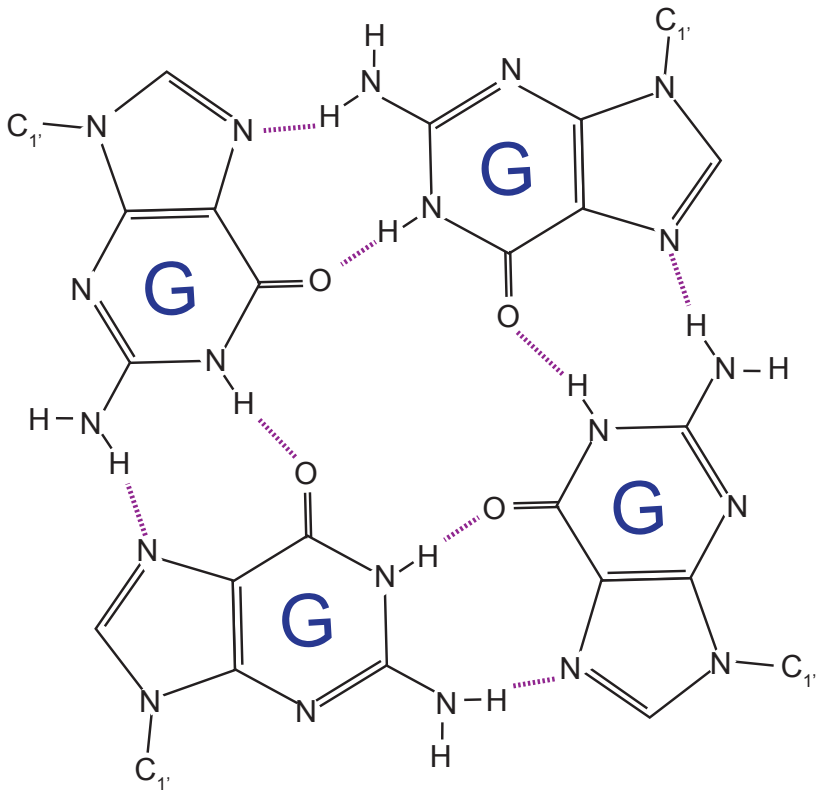

Figure 2.9: The guanine tetrad, or G-quartet, is formed when four guanine nucleotides bond in a planar fashion via Hoogsteen bonding. G-quartets are often stacked on top of one another to form additional tertiary structures.

2.5.1 G-Quartet

G-quartets (also known as G-quadruplexes and G-tetrads) are nucleic acid sequences that are rich in guanine and are capable of forming a quadruple-stranded structure. Four guanine bases can associate through Hoogsteen hydrogen bonding, see Section 1.3.1, to form a square planar structure called a guanine tetrad, as shown in Figure 2.9. Hydrogen bonded G quartets are known to occur in the telomeric regions at the ends of chromosomes, and G-tetraplexes can form from guanine monomers, guanine rich DNA and RNA oligomers, and polymers [13]. Two or more guanine quartets can also stack on top of each other to form a stacked G-quadruplex. The quartet structure is further stabilized by the presence of a cation (K^+ is favored over Na^+) which sits in the central channel between each pair of tetrads. They can be found in both DNA and RNA and may be intramolecular, bimolecular, or tetra molecular. Depending on the direction of the strands or parts of a strand that form the quaduplexes, the structures may be described as parallel or antiparallel. Depending on the sequence and the number of nucleic acids involved, there is a wide variety of tertiary structures

that can be formed from G-quartets.

Telomeric repeats in a variety of organisms are rich in G-quartets. The human telomeric repeat consists of many repeats of the sequence d(GGTTAG), and the tetrads formed by this structure have been well studied by NMR and X-ray crystallography. G-quartets, with their higher degree of stabilization, decrease the activity of many enzymes. These tetrads also appear in or near the promoter regions of genes. In fact, genome wide surveys have been conducted and find that there are thousands of potential G-quartet sites on each chromosome. There are several possible models for how quadruplexes could control gene activity, by either up- or down-regulation. Up-regulation could be achieved by forming in the non-coding DNA strand and helping to maintain an open conformation of the coding DNA strand to enhance transcription and expression. Down-regulation could be achieved by a G-quartet forming in or near a promoter sequence and blocking transcription [12, 13]. In addition to chromosome stability and gene transcription regulation, guanine tetrads are implicated in recombination, viral integration, and cellular senescence.

2.5.2 I-Motif

On the complementary strand of a Watson–Crick double helix that can form G-quartets, the cytosine rich strand can also fold into a complex and unique secondary structure. Called the i-motif, the tetrad is comprised of intercalated $C \cdot C^+$ base pairs. Although only two cytosines are bound to each other, the i-motif arises due to the intercalated nature of the folded structure. The four sections of the nucleic acid are 90° from each other. The molecules are arranged so that they are intercalated; that is, the two nucleic acid strands directly across from each other have the base rings planar, but the other two nucleotides are not in the same plane. This allows a zipper-like configuration to arise, with each cytosine bound to only the one directly across from it, but the four strands arranged in a stable tetrad complex. The stability of the i-motif depends on the pH (since the cytosines need to be protonated), and C-rich telomeric repeats have been measured folded in stable i-motifs at a slightly acidic pH. As there was a diverse set of possible G-quartets, there are a myriad of possible i-motif structures each with different intercalation and looping topologies [42].

Although the Watson–Crick double helix is the predominant form under physiological conditions, the G-quadruplex and i-motif may be formed under different conditions such as high temperature or low pH. *In vivo* these structures can occur anywhere

along chromosomal DNA (but particularly in the telomeric regions), permanently or temporarily, either naturally or in a pharmacological context. Stability at neutral or alkaline pH can be enhanced by inter-molecular interactions (proteins-DNA, RNA-DNA, DNA-DNA, or drugs-DNA), or by superhelical stress on duplex DNA. The formation of an intramolecular G-quartet and i-motif can occur separately or together, at the same position or at different locations on a duplex [42].

The secondary and tertiary structures of nucleic acids are vital to the tasks that these molecules perform, whether it is the long, regular, organized double-helix for the storage of genomic data or the complex three-dimensional shapes essential to aptamer affinity, selectivity, and function. In order to better understand these biological molecules and their abilities, a multitude of experimental methods have been devised. In addition to laboratory based examinations, simulation has arisen as an additional method to understand nucleic acids at every length and time scale. In the next chapter we will describe several nucleic acid models and simulation methods that span the relevant length and time scales for the features of DNA and RNA discussed in Chapters 1 and 2.

3 Nucleic Acid Simulation Models and Methods

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

George E. P. Box, *Empirical Modeling-Building and Response Surfaces*, 1987 [122]

3.1 Introduction to Nucleic Acid Simulation Models and Methods

Nucleic acids can take a wide array of complex shapes as their single, double, triple, and even quadruple nucleotide interactions permit multifaceted conformations to form. From the very beginning of nucleic acid study with the physical models of Watson and Crick [6, 7], abstraction and modeling have played a significant role in the understanding of DNA and RNA. These physical models gave rise to mathematical concepts and techniques to study complex configurations. DNA simulation models span many length scales, from atomistic models describing only a few nucleotides, to dumbbell or cylinder DNA models that encompass entire chromosomes. Just as these models describe DNA molecules ranging from the micro to the macroscale, the methods used in their simulation are also suited for an array of length and time scales.

As with the sentiment of Box in *Empirical Model-Building and Response Surfaces*, it must be remembered that nucleic acid models are only approximations; each model therefore, must be evaluated to determine to which biological systems it is best suited

before it becomes "not useful" [122]. Careful consideration must be utilized when choosing a nucleic acid model and method; in this chapter we will outline some of the possible models and methods for DNA and RNA investigations. The simulation of nucleic acids is now a mature field [123] and modeling techniques are powerful tools for the understanding and description of their forms and functions.

3.2 Nucleic Acid Simulation Models

Many models of DNA are available in literature. Although a complete description in terms of all atomic coordinates would at first glance appear desirable, as more chemical detail is included in a model not only do the computational requirements associated with its solution increase significantly, but such detailed models may garner little additional physical insight. If the property of interest are not defined on the Angstrom length scale or femto to picosecond time range, then the calculations will simply examine the emergent phenomena. The additional burden on the system to evaluate unnecessary detail severely restricts the length and time scales amenable to study. The overarching challenge in simulating nucleic acids is therefore to include just enough detail in a model to capture the physics that are responsible for DNA's relevance in a particular application [124]. We will continue our discussion of nucleic acid models by examining three model length scales: atomistic, coarse-grained, and continuum models in Sections 3.2.1, 3.2.3, and 3.2.2, respectively.

3.2.1 Atomistic Scale of Nucleic Acid Models

At the most detailed level, atomistic simulations provide the most intimate embodiment of DNA. In these models, most, if not all, atoms are explicitly depicted as shown in Figure 3.1. This allows nucleic acid models at this scale to capture sequence dependent characteristics of oligonucleotides. The interactions between groups are defined by the force field, or the set of parameters and mathematical functions used to describe the potential energy. Force field functions and parameter sets are derived from both experimental work and quantum mechanical calculations [125–132]. These quantum mechanical calculations are often used to further refine particular characteristics in a nucleic acid model that may be of particular importance to the system examined, such as the hydrogen bonding and stacking interactions between nitrogenous bases [12, 13, 131, 132]. In fact, the *ab initio* quantum chemical calculations

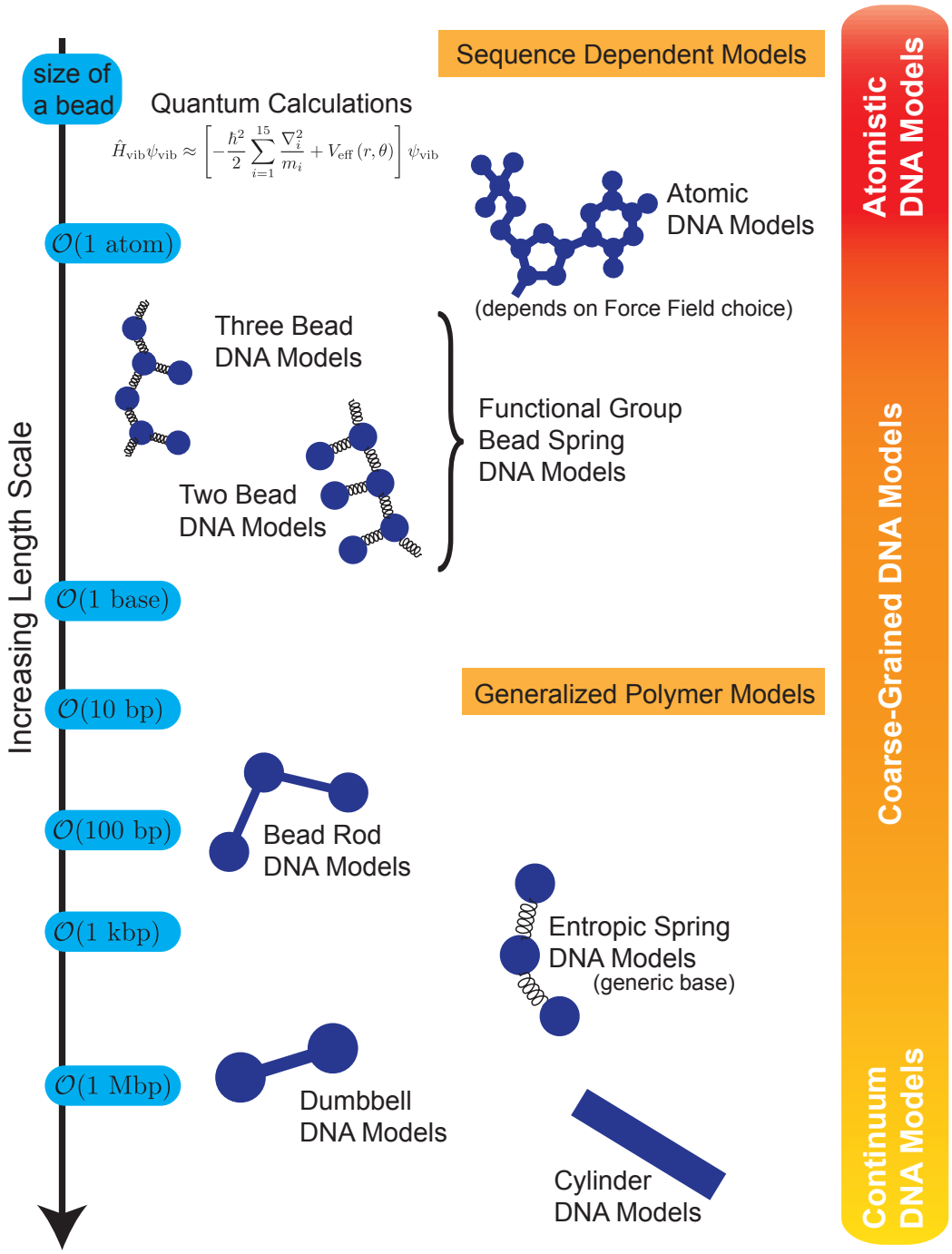

Figure 3.1: A schematic representation of nucleic acid models over many orders of magnitude. The models are divided into three broad categories: atomistic models where every or nearly every atom is included, coarse-grained models where groups of atoms, nucleotides, or other substructures are combined, and continuum models where the DNA is represented as cylinders or fibers. These models can be used to capture sequence dependent features and generalized polymer features of a system. On the left side of the figure is a qualitative length scale (not to scale) of the depiction of a single bead (when applicable) ranging from a bead representing one atom to a bead representing millions of base pairs (or thousands of persistence lengths).

with inclusion of electron correlation made since 1994 (such reliable calculations were not feasible before then) significantly modified our view on interactions of nucleic acid bases [13, 127].

All-atom force fields provide parameters for every type of atom in a system, including hydrogen, while united-atom force fields treat the hydrogen and carbon atoms in methyl and methylene groups as a single interaction center. These two types of atomistic (or near atomistic) nucleic acid models can be used separately, or in some situations, together in a hybrid approach that gives the most detail to particular aspects of the system and thus greater accuracy for these components. In biological systems, the CHARMM [129, 133] and AMBER [134, 135] force fields commonly provide the interaction descriptions necessary. Atomistic simulations of biological systems often include explicit representation of every atom in the system, solute and solvent alike, however it can also be limited to explicit rendering of the biological molecule and implicit representation of the solvent [46, 123, 125, 126, 135–151]. Further groupings and approximations of the atoms and thus the force fields used to describe biological molecules have been completed; these simplifications are essentially coarse-graining the system and will be discussed in Section 3.2.3.

Despite the excellent agreements between the atomistic simulations of nucleic acids and experimental results, the scale of oligonucleotides that can be examined with such a detailed model is quite restricted. Typically less than 50 nucleotides are simulated for at most of a few microseconds; most studies are of shorter nanosecond length simulation times. Comprehensive reviews and recent calculations show that extensive studies of long DNA molecules continue to be well beyond the reach of fully atomistic representations [124].

3.2.2 Continuum Scale of Nucleic Acid Models

At the other end of the spectrum of length scales, and for the study of full genomic DNA, continuum models of DNA treat the double helix as a uniform medium [152, 153]. Several continuum models have been used to shed considerable light onto the mechanical behaviors, such as bending, of DNA in various environments. For example, one of the simplest models of DNA deformability treats DNA as an ideal elastic rod, that is a thin elastic body that is inextensible, intrinsically straight, transversely isotropic and homogeneous [154]. When these models are used, deformation of nucleic acids in any direction of the space or any part of the rod are equally probable. These

features make it possible, for instance, to represent large pieces of DNA (at the kilo base pair to mega base pair range) or even entire ribosomal RNAs (rRNA), but only with low resolution. While these approaches can investigate macroscopic properties where DNA can be considered as a continuum, they are by definition, unable to deal directly with complex structural formations, sequence dependent features, or environmental effects.

3.2.3 Coarse-Grained Scales of Nucleic Acid Models

For many applications of interest, however, the approaches mentioned above are either superfluous (atomistic) or inadequate (continuum). As with many of the configurations described in Chapters 1 and 2, the complex structures require both numerous nucleotides and sequence dependent features to form; atomistic models typically cannot reach the number of nucleotides necessary while continuum models fail to capture any sequence dependent (and thus local) features. In order to capture these features, a variety of coarse-grained models of nucleic acids have been developed that span several orders of magnitude (from a couple dozen nucleotides to millions of base pairs), as shown in Figure 3.1. Such models have been successful in reproducing distinct features of nucleic acids but do not provide a comprehensive description of both local features (such as hybridization or melting) and global features (such as mechanical properties) within one representation. Therefore, coarse-grained nucleic acid models are chosen such that they provide the relevant information at the length and time scales inherent in their design. In the remainder of this section we shall endeavor to give a brief overview of the different levels of coarse-grained models available and to which systems they have been applied. This discussion will continue in Chapter 4 for a two bead nucleic acid model, Chapter 5 for a three bead nucleic acid model, and in Chapter 7 for a multi-scale model approach.

Continuing with the model simplifications described in Section 3.2.1, additional atoms are grouped together in each nucleotide. Transitioning from atomistic models with between 32 and 36 atoms (also called interactions sites) per nucleotide, coarse-grained models typically begin with five to eight interaction sites or beads per nucleotide. In these models the beads are representing nucleotide structures of groupings of atoms; the springs are used to keep these groupings at desired relative positions. This level of coarse graining allows for many of the features of the nucleotide to be included (for example, the steric hinderance of the nitrogenous base planes) without having to

include all or even most of the atoms. Further, sequence dependent coarse-graining allows for functional groups to be captured and have on the order of two or three beads per nucleotide. As nucleic acid models continue to be less refined, we transition from primary structure dependent models to generalized polymer models. In these, the base identity can no longer be captured and/or the model represents not ssDNA but instead generic double-stranded DNA. Here a single bead can represent anywhere from a single base (or one base pair) to thousands of base pairs (kbp). Finally, dumbbell models can model an entire chromosome (on the order of millions of base pairs (Mbp)) with only two interaction sites and are typically used for polymer and mechanical property studies. These models fully span the atomistic to continuum scale of nucleic acid modeling.

Sequence Dependent Models

Beyond the atomistic length scale, a series of increasingly less refined nucleic acid models have been developed in which multiple beads represent each nucleotide. These beads are typically connected with springs, and thus, are deemed bead-spring models. These models most often encode primary structure though not necessarily, and often have a varying nature of what each bead represents. Fundamentally, each bead represents a functional group in the nucleotide, the division of beads can be defined by steric considerations or chemical considerations. For example, one system with eight beads per nucleotide [155] does not include specific base interactions (and thus no primary structure information), but rather focuses on the steric properties of each nucleotide in their Watson–Crick bound state (secondary structure information) for particular study of the spontaneous helical tertiary structure of DNA. A system with seven beads per nucleotide [156] is parameterized particularly for RNA and can encode base identity. Further, in another system, nucleotides are represented by six interaction sites [157] and can encode base identity and thus particular chemical interactions between functional groups. Finally, a slightly less refined system of four or five beads per nucleotide [158] has been developed to examine some secondary structures of nucleic acids. Depending on the computational resources and methods utilized, these models may be able to reach the simulation time (on the order of seconds) and number of nucleotides necessary to examine some of the secondary and tertiary structures described in Chapter 2. However, it should be noted that none of these models encode non-canonical hydrogen bonding, therefore many of the formations detailed in Chapter 2 cannot be examined.

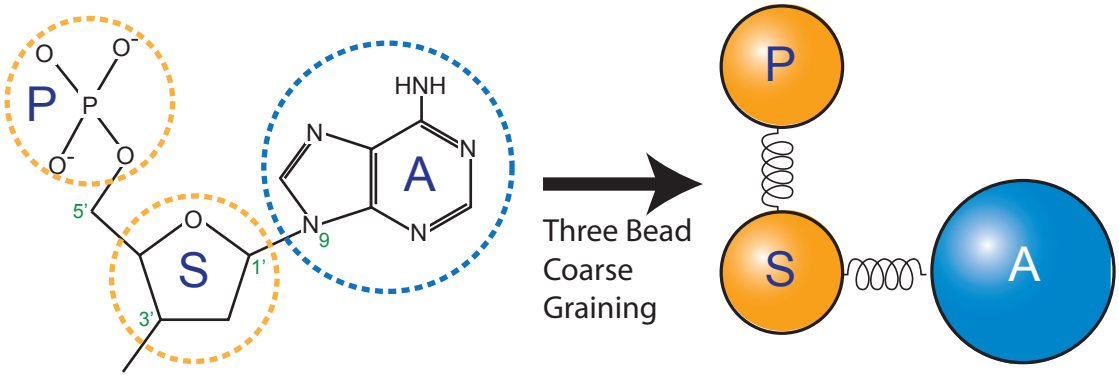

Figure 3.2: A schematic illustration of the coarse-graining process of a nucleotide to a three bead model. A nucleotide containing adenine is represented in both depictions; the phosphate group (P), the deoxyribose sugar group (S), and the nitrogenous base group (A) that are circled are then embodied three beads. Each bead in the coarse-grained effort is then given different characteristic interactions.

Further coarse-graining depicts three beads per nucleotide (also called a three bead model) and is schematically represented in Figure 3.2 [63, 159–167]. These systems have been used for a wide variety of investigations, including everything from hybridization and melting to bending and packing studies. For three bead models, the three functional groups: the phosphate group; the sugar group; and the nitrogenous base group are rendered with three separate beads. Each of these can have its own interaction specificity. As depicted in Figure 5.1 and described in detail for one such system [166] in Chapter 5, a three bead representation allows for chemical considerations to be included with base identity preserved. In addition, fundamental steric considerations are also included in the model through the bead sizes or angular potentials. The three bead per nucleotide model can capture several detailed aspects of nucleic acids that were described in Chapter 1 such as chirality (left- or right-handed helices), major and minor grooves, base stacking, and hydrogen bonding characteristics. Depending on the model, parameters, and other considerations, the ability to capture some or all of these features allows many of the structures described in Chapter 2 to be examined.

Two interaction sites per nucleotide models [168–177] are also widely utilized to examine oligonucleotide systems. These models have also been used to examine hybridization and melting, in addition to helical structures, aptamers, quadraplex structures

(Holliday junctions), and nanotweezers. These models have varying abilities to encode primary structure; some have generic bases, some encode only hydrogen bonding base identities, and some can capture many features of each adenine, cytosine, guanine, and thymine base. As illustrated in Figure 4.1 and described in Chapter 4 [172, 174], a two bead representation has a backbone bead encompassing the sugar and phosphate groups and a base bead for each nitrogenous base. This level of coarse-graining allows partial steric considerations to be included in the model by characterizing the size of the beads or the angles between them, along with chemical factors, such as base stacking and hydrogen bonding, These features allow the capture of some of the secondary and tertiary structures described in Chapter 2.

Generalized Polymer Models

As we continue up the length scale, we move from functional group bead-spring models (where the beads are fundamentally different to capture sequence specificity) presented previously to a system where the model divides the polynucleotide into beads that usually are uniform in type and characteristics. The size of one bead-spring unit can vary depending on the type of problem under analysis: from a single base pair to tens of thousands of base pairs. These will be discussed below.

At the next level of bead-spring models, and in the transition from sequence dependent features to generic nucleic acid polymer models, are the one bead per nucleotide or base pair systems. In these, either each nucleotide becomes a single interaction site [178–183] or each base pair, that is two nucleotides enjoined by hydrogen bonding (and almost always Watson–Crick hydrogen bonding) is represented as a single bead [147]. Since there is no longer nucleotide specificity, base identity is lost in these models and each bead acts as a generic base with features not dependent on whether it is an adenine, cytosine, guanine, or thymine nucleotide. Simulating tens to hundreds of nucleotides, these models have been used successfully to examine helical structure, flexibility, hybridization, and melting.

Further polymer models that are defined at the sub-persistence length consist of entropic spring DNA models. In these systems, we can think of the entropic springs now representing the nucleic acid strand with the beads mainly providing reference points along the chain for long range interactions; opposed to how the functional group bead spring DNA models, described above, represent the components of each nucleotide with the beads and the springs are used for position regulation. These

models range from a bead-spring unit representing three base pairs [184, 185] to a bead-spring unit representing anywhere from eight to 150 base pairs [186]. Often sub-persistence length (on the order of 150 base pairs) models can be tuned through the use of bending, electrostatic, and other potentials to embody any number of base pairs in this range. These bead spring models have often been used to examine twisting, bending, supercoiling, ejection of DNA from a phage, packing into a nucleus or phage, and band translocation (as through a nanopore).

At the persistence length of DNA (RNA molecules do not typically reach these lengths due to their inherent instabilities), the springs connecting the backbone beads, are often replaced with extremely stiff springs or rods (and thus named a bead-rod model) and used to simulate thousands of base pairs [58, 187–190]. Bead-rod models typically investigate static and dynamic properties of DNA including stress, relaxation, diffusivity studies along with solvent effects on the three-dimensional conformation of the DNA polymer. It should be noted that bead rod models are not restricted to this scale (though they are often used on the persistence length scale) and can be used at other length scales [170].

Beyond the persistence length of DNA, entropic spring models are again typically used [188, 191–198]. A repeating unit of the polymer can thus represent anywhere from a Kuhn length (or twice the persistence length) to hundreds of Kuhn lengths. The level of coarse-graining is therefore usually decided by the computational resources and the amount of information detail needed to represent a particular dsDNA strand. One of the more commonly modeled DNAs is referred to as λ-DNA and is an approximately 48,000 bp digest of the λ bacteriophage [199, 200]. Other digests are available and can create dsDNA fragments that range from a few hundred base pairs to tens of thousands of base pairs in length. Extremely coarse (where a bead is thousands of base pairs) entropic spring models are often used to capture the polymer type behaviors of dsDNA in various confinements (for example, nanoslits, wells, and post arrays) and reagent environments. In spite of their low resolution, they yield results in excellent agreement with experimental data for diffusion, structural relaxation, and behavior under different flow fields for bulk and confined DNAs [124, 201].

Finally, entropic spring models can evolve to the most simplistic of systems: the trumbbell (three beads connected with two springs) and the dumbbell (two beads connected with one spring) models can be used to represent double-stranded DNA [189, 193]. These models define the broadest of coarse-grained length scales where each bead and spring unit represents tens of thousands to millions of base pairs. It

should be noted that trumbbell and dumbbell models can be used for shorter DNA and RNA molecules.

3.2.4 Treatment of the Solvent

Nearly all biological systems are studied in a solvent, either a simplistic solvent comprised typically of water and indicative of many simplified *in vitro* experiments or complicated solvents containing many salt and metal ions and characteristic of often necessary *in vivo* conditions. In this work we will regard the treatment of the solvent as a simulation methodology decision instead of a simulation modeling consideration. Therefore, as will be described in Section 3.3.1 the solvent can either be represented explicitly with every (or nearly every) water, salt, and metal atom or ion individually represented, implicitly as a generalized force acting on the nucleic acid polymer as detailed in Section 3.3.4, or the fluid can be treated in a manner somewhere in between as in Sections 3.3.2 and 3.3.3.

3.2.5 Multi-Scale Models of Nucleic Acids

Frequently, a combination of models focusing on different length scales will be used to examine a particular nucleic acid system. Such multi-scale approaches can take one of two forms: (i) a particular system is examined at multiple length scales with differently resolved models in each simulation, but the entire nucleic acid is represented by only one scale model at a time, or (ii) part of the system has more or less detail than other components, giving rise to increased information only for particular sections of the nucleic acid. The first form is often utilized in coarse-grain model development; parameters are often found by using a "bottom-up" approach [157, 164, 182, 183, 202, 203]. However, as will be described in detail in Chapter 7, the entire system can be represented with different models to gain a more detailed understanding of particular phenomena by varying the length and time scales for particular observations. Secondly, multi-scale models are often mixed in a particular system in order to increase the resolution of a particular part of the nucleic acid without having to fully represent the entire system at the same length scale [204–207]. Because of the simplifications to the model and the reduction in the number of interaction sites needed, longer simulation times can be reached with the same computational resources. Further discussions in Section 3.3.6 will outline how multi-scale

modeling methods can also be implemented, on both components of a nucleic acid polymer and the solvent, in order to be able to gain more resolution of particular aspects in a biological system.

3.3 Nucleic Acid Simulation Methods

The application of simulation methods has proven to be an increasingly powerful tool for the study of molecules of biochemical and biological interest. In particular, nucleic acid simulation applications have employed different methods to computer dynamic and/or static properties of interest. In this chapter we will briefly describe some of the more common methods of nucleic acid simulations including the temporal methods of molecular dynamics (MD), dissipative particle dynamics (DPD), multi-particle collision dynamics (MPC), and Langevin/Brownian dynamics (LD/BD) techniques. In addition, the Monte Carlo (MC) configurational sampling methods will also be outlined. This list is limited in nature and the details are intended only to give a broad overview and do not include the intricacies vital to each of the simulation methods. In addition, new simulation methods are proposed every year, the development of new methodologies will continue to improve nucleic acid understanding via simulation methods [124, 201]. Finally, there are a variety of multi-method approaches that have also been utilized for nucleic acid studies. Like the nucleic acid models discussed in Section 3.2, many different simulation methods are associated with certain ranges of length and time scales and must be carefully chosen for each particular nucleic acid system.

3.3.1 Molecular Dynamics Simulation Method

Molecular dynamics (MD) is a simulation technique that allows the prediction of the time evolution of a system of interacting particles (such as atoms, molecules, or coarser interaction sites or beads) by numerically integrating the classical equations of motion and can estimate relevant physical properties [201, 208–211]. Although MD is often used to simulate systems at the molecular level, it is also suitable for modeling coarser systems. Specifically, it generates such information as positions, velocities, and forces from which the macroscopic properties (such as pressure, energy, and heat capacities) can be derived by means of statistical mechanics. MD simulation consists of three constituents: (i) a set of initial conditions (initial positions and velocities

of all particles in the system); (ii) the interaction potentials to represent the forces among all the particles; and (iii) the evolution of the system in time by solving a set of classical Newtonian equations of motion for all particles in the system. The equation of motion is generally given by

$$\mathbf{F}_i(t) = m_i \frac{\mathrm{d}^2 \mathbf{r}_i}{\mathrm{d}\, t^2} \tag{3.1}$$

where $\mathbf{F}_i$ is the force acting on the ith particle at time t which is obtained as the negative gradient of the interaction potential U, m_i is the mass, and $\mathbf{r}_i$ the position. A physically relevant molecular dynamics simulation involves not only the Newtonian equation of motion for all particles in the system, but also the proper selection of interaction potentials, numerical integration, boundary conditions, and the controls of pressure, temperature, particle number, and volume to mimic physically meaningful thermodynamic ensembles.

The interaction potentials together with their parameters, constitute the so-called force field, and describe in detail how the particles in the system interact with each other; the potential energy of the system is dictated by the particle coordinates. Such a force field may be obtained by a variety of methods, and several different atomistic nucleic acid force fields were described previously in Section 3.2.1. A typical interaction potential U may consist of a number of bonded and non-bonded interaction terms:

$$\begin{aligned} U(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \;=\; & \sum_{i_{\mathrm{bond}}}^{N_{\mathrm{bond}}} U_{\mathrm{bond}}(i_{\mathrm{bond}}, \mathbf{r}_a, \mathbf{r}_b) + \sum_{i_{\mathrm{angle}}}^{N_{\mathrm{angle}}} U_{\mathrm{angle}}(i_{\mathrm{angle}}, \mathbf{r}_a, \mathbf{r}_b, \mathbf{r}_c) \\ & + \sum_{i_{\mathrm{torsion}}}^{N_{\mathrm{torsion}}} U_{\mathrm{torsion}}(i_{\mathrm{torsion}}, \mathbf{r}_a, \mathbf{r}_b, \mathbf{r}_c, \mathbf{r}_d) \\ & + \sum_{i=1}^{N-1} \sum_{j>i}^{N} U_{\mathrm{vdw}}(i, j, \mathbf{r}_a, \mathbf{r}_b) \\ & + \sum_{i=1}^{N-1} \sum_{j>1}^{N} U_{\mathrm{electrostatic}}(i, j, \mathbf{r}_a, \mathbf{r}_b). \end{aligned} \tag{3.2}$$

The first four terms represent bonded interactions: bond stretching U_{bond}, bond-angle bend U_{angle}, and dihedral angle torsion U_{torsion}. The last two terms are non-bonded interactions: van der Waals energy U_{vdw} and electrostatic energy $U_{\mathrm{electrostatic}}$. In the above equation, $\mathbf{r}_a$, $\mathbf{r}_b$, $\mathbf{r}_c$, and $\mathbf{r}_d$ are the positions of the atoms or particles

specifically involved in a given interaction; N_{bond}, N_{angle}, and N_{torsion} stand for the total numbers of these respective interactions in the simulated system; i_{bond}, i_{angle}, and i_{torsion} uniquely specify an individual interaction of each type; i and j in the van der Waals and electrostatic terms indicate the atoms involved in the interaction.

There are many algorithms for integrating the equation of motion. The algorithms of Verlet, velocity Verlet, leap-frog, and Beeman, are commonly used in MD simulations [201, 208**?** –212]. MD simulations can be performed in many different ensembles, such as grand canonical (μVT), micro canonical (NVE), canonical (NVT), and isothermal-isobaric (NPT). The constant temperature and pressure can be controlled by adding an appropriate thermostat and barostat, respectively [124, 201, 213].

In addition, it is most common within atomistic MD systems to represent the solvent explicitly; many differing models have been developed that explicitly describe a water molecule [201]. By including each individual water molecule in the simulation, the number of interaction sites is drastically increased and thus there is a great reduction in the accessible time and length scales [183, 201]. There are several advantages to using an explicit fluid instead of a coarser, bead representation of the solvent or an implicit representation with a generalized force field, such as local solvation interactions (such as sugar hydration effects) and the preservation of long range hydrodynamic interactions; however most of the calculation time in these systems will be spent on the fluid and not the nucleic acid of interest. As will be seen in Sections 3.3.2 and 3.3.3, other methods have been developed (by varying degrees of coarse graining) that represent the solvent in the systems without the computational costs of MD. The Langevin and Brownian dynamics methods described in Section 3.3.4 continue the trend by eliminating the need to explicitly represent the fluid at all. Finally, we will conclude our discussion of methods with Section 3.3.5 which gives a broad overview of Monte Carlo techniques.

The molecular dynamics simulation method can be used for a variety of nucleic acid models ranging from the atomistic scales to much coarser systems [46, 123, 125, 126, 129, 130, 133, 135–150, 155–165, 171, 181, 204]. Molecular dynamics, with the proper atomistic level model choice (as described in Section 3.2.1), allow access to the finest length and time scales and is often used in atomistic and microscopic systems. As will be seen in Chapter 7, atomistic models coupled with molecular dynamics can access features on the Angstrom length scale and femtosecond time scale; however, due to limited computational resources, these types of simulations rarely exceed microseconds worth of simulation time [124].

3.3.2 MD - Dissipative Particle Dynamics Simulation Method

Like molecular dynamics, dissipative particle dynamics (DPD) is a particle-based method [201]. The solute particles continue to be governed by molecular dynamics, however, the fluid is modeled by large particles interacting via soft potentials [192, 214–218]. Each of these particles represents a cluster of solvent molecules moving together in a coherent manner. The computing time is reduced by coarse graining of the solvent and thus reducing the number of interaction sites at the expense of capturing hydration effects.

DPD particles are defined by their mass m_i, position $\mathbf{r}_i$, and momentum $\mathbf{p}_i$. The interaction force between two DPD particles i and j can be described by a sum of conservative $\mathbf{F}_{ij}^C$, dissipative $\mathbf{F}_{ij}^D$, and random forces $\mathbf{F}_{ij}^R$:

$$\mathbf{F}_{ij} = \mathbf{F}_{ij}^C + \mathbf{F}_{ij}^D + \mathbf{F}_{ij}^R, \tag{3.3}$$

$$\mathbf{F}_{ij}^C = \Pi_0 \omega_C(\mathbf{r}_{ij}) \hat{e}_{ij}, \tag{3.4}$$

$$\mathbf{F}_{ij}^D = -\gamma \omega_D(\mathbf{r}_{ij})(\hat{e}_{ij} \cdot \mathbf{p}_{ij}) \hat{e}_{ij}, \tag{3.5}$$

$$\mathbf{F}_{ij}^R = \sigma \zeta_{ij} \omega_R(\mathbf{r}_{ij}) \hat{e}_{ij}, \tag{3.6}$$

where $\mathbf{r}_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ and $\hat{e}_{ij} = \hat{r}_{ij}/r_{ij}$. The force $\mathbf{F}_{ij}^C$ represents conservative forces that act on the particle, and in the above expression Π_0 is a constant related to the fluid compressibility. The dissipative force is an inter-drag force between a pair of soft fluid particles moving through each other opposing their relative motion ($\mathbf{p}_{ij}$) and dissipating heat where γ is the friction constant between the two clusters. The random noise force is given by $\mathbf{F}_{ij}^R$ with σ a noise amplitude, and ζ_{ij} a random noise term with zero mean (i.e., $< \zeta_{ij} >= 0$) and unit variance. To satisfy the fluctuation dissipation theorem, the dissipative and random forces are interrelated through the weight functions, $\omega_D(\mathbf{r}) = [\omega_R(\mathbf{r})]^2$; $\omega_C(\mathbf{r})$ is the weight function for the conservative force function. While the interaction potentials in molecular dynamics are high-order polynomials of the distance r_{ij} between two particles, in DPD the potentials are softened so as to approximate the effective potential at microscopic, and not atomic, length scales. The form of the conservative force in particular is chosen to decrease linearly with increasing r_{ij}. Beyond a certain cut-off separation, r_C, the weight functions and thus the forces are all zero.

Therefore the total force $\mathbf{F}_i(t)$ acting on particle i at time t is

$$\mathbf{F}_i(t) = \sum_{j \neq i} \mathbf{F}_{ij}^C + \sum_{j \neq i} \mathbf{F}_{ij}^D + \sum_{j \neq i} \mathbf{F}_{ij}^R. \tag{3.7}$$

Because the forces are pairwise and momentum is conserved, the macroscopic behavior directly incorporates Navier–Stokes hydrodynamics. Dissipative particle dynamics can simulate both Newtonian and non-Newtonian fluids, including polymer melts and blends, on microscopic length and time scales. Dissipative particle dynamics are often coupled with a standard molecular dynamics approach so that the fluid and the nucleic acid polymer are simulated with the respective techniques. Nucleic acids in fluid flow environments, be that tethered, in confined geometries, or in separation devices, have been modeled with dissipative particle dynamics schemes [192, 214–218].

3.3.3 MD - Multi-Particle Collision Dynamics Simulation Method

Multi-particle collision dynamics (MPC), also known as stochastic rotation dynamics (SRD), continues the trend of coarse-graining the interactions of the fluid particles [201]. In a DPD system, the solvent beads represent clusters of fluid molecules and all bead-bead interactions are calculated. By contrast, in MPC the collisions between fluid particles are replaced by multi-particle collision events that omit the molecular details and eliminate the need to calculate the forces between the fluid particles. Here the savings of reducing the fluid particles is countered by the loss of capturing hydration effects. In this method, the fluid is modeled as point particles of mass m_i. MPC simulations occur in two steps: (i) a streaming step where the particles move ballistically and (ii) a collision step which transfers momentum between the particles.

In the first step, each particle evolves in time according to Newton's laws of motion. There are no interactions among the particles, but external forces may be present. The positions $\mathbf{r}_i(t)$ are updated in discrete time intervals δt:

$$\mathbf{r}_i(t + \delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\delta t + \frac{1}{2}\mathbf{f}_i \delta t^2, \tag{3.8}$$

where $\mathbf{v}_i$ is the velocity and $\mathbf{f}_i$ is the acceleration due to some external force.

The second step transfers momentum between the particles. The simulation domain is partitioned into cells and each cell has a center of mass velocity, $\mathbf{v}_{CM}$ which corresponds to the local macroscopic velocity. The particles within each cell only interact with other members of the same cell by means of a non-physical scheme constructed to conserve momentum. Multi-particle collisions within each cell are represented by the operation

$$\mathbf{v}_i(t + \delta t) = \mathbf{v}_{CM}(t) + \mathbf{R}(\mathbf{v}_i(t) - \mathbf{v}_{CM}(t)). \tag{3.9}$$

The collision operator, $\mathbf{R}$, consists of a rotation through an angle α about a randomly chosen axis. The collision events are defined to conserve mass, momentum, and energy such that the hydrodynamic equations of motion are obeyed on sufficiently long length and time scales. Similar to the coupling that is often seen between DPD and MD, MPC can be integrated into a standard MD simulation. The nucleic acid polymer, and any other solute molecules in the simulation can be integrated with the momentum of the fluid by including them in the MPC collision step. As an example, multi-particle collision techniques have been used to examine the dynamics of nucleic acids in micro-channels [219–221].

3.3.4 Langevin and Brownian Dynamics Simulation Methods

Beyond the mesoscopic fluid models of DPD and MPC, further coarse-graining can be achieved by avoiding direct simulation of the fluid altogether [201, 222–224]. It is computationally advantageous to coarse-grain out the fine details of the collisions of individual collisions with the solvent molecules but to keep the statistical effects of their frictional drag and Brownian motion. At this scale we can consider the two main effects of the fluid acting on a particle: (i) a frictional force opposing its motion and (ii) random kicks arising from collisions with the solvent. The frictional (or dissipative) force $\mathbf{F}^D(t)$ removes energy from the particle while the fluctuating Brownian force $\mathbf{F}^B(t)$ adds energy to the particle. Hence, at this level of coarse-graining, the fluid is included solely in a statistical manner govern by the fluctuation dissipation theorem. By replacing the explicit fluid with a drag and a Brownian force, the long-range particle-particle interactions mediated by the fluid are lost. However, there are methods of including hydrodynamic interactions in Langevin or Brownian Dynamic systems [225].

By adding the dissipative drag force and the Brownian force to Newton's second law shown in Equation 3.1 (with the additional conservative forces, $\mathbf{F}^C(t)$), we reach

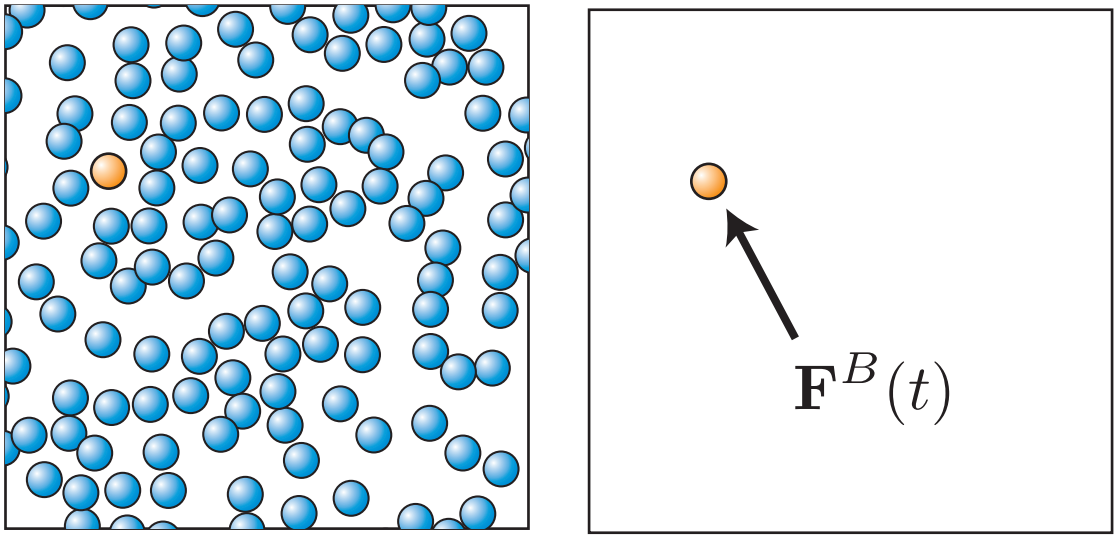

Figure 3.3: To avoid explicitly calculating interactions with a solvent for the particle (orange sphere), the fluid particles (blue spheres) are treated as a viscous medium. To account for the Brownian motion and dissipative losses that occur as a result of collisions with large numbers of solvent particles, a stochastic force, $\mathbf{F}^B(t)$ and a drag force are implemented into the simulation. The result is that larger systems and longer time scales are accessible than in transitional molecular dynamic approaches.

Langevin's equation,

$$\mathbf{F}(t) = m\frac{\mathrm{d}^2\,\mathbf{r}}{\mathrm{d}\,t^2} = \mathbf{F}^C(t) + \mathbf{F}^D(t) + \mathbf{F}^B(t), \tag{3.10}$$

where m is the mass of the particle. The frictional force due to the drag exerted on the particle by the fluid is given by

$$\mathbf{F}^D(t) = -\gamma\mathbf{v}(t) \tag{3.11}$$

where γ is the frictional coefficient [59]. The velocity $\mathbf{v}(t)$ is the velocity of the particle with respect to the local solvent velocity. For a spherical particle the Stokes formula,

$$\gamma = 6\pi\eta R \tag{3.12}$$

where R measures the radius of the particle and η is the viscosity of the solvent, is used.

The Brownian force represents the constant molecular bombardment exerted by the surrounding fluid, as depicted in Figure 3.3. Although $\mathbf{F^B(t)}$ is due to the solvent molecules colliding with the particle, it can only model the net effect of a large number

of collisions. The Brownian force is taken as a centered Gaussian random variable with zero mean and variance $2\gamma k_B T/\Delta t$, where Δt is the integration time step. As with DPD, the fact that the variance is related to the frictional coefficient γ is again a consequence of the fluctuation dissipation theorem. At the time scales of interest, the values of the Brownian force are uncorrelated at different time steps.

It can be shown that the energy transferred to the particle from a single collision with a solvent molecule decays on the viscous time scale m/γ [226]. If this is much smaller than the timescale over which $\mathbf{F}^C(t)$ changes (over damped limit), then $m\mathbf{a}(t) = 0$ in the Langevin equation and obtain the following discretized equation of motion:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \frac{\Delta t}{\gamma}[\mathbf{F}^B(t) + \mathbf{F}^C(t)] \tag{3.13}$$

which defines Brownian dynamics (BD).

BD is particularly useful for systems where there is a large gap in the time scale governing the motion of the solvent and polymer. For example, in DNA or RNA polymer and solvent mixtures, a short time-step is required to resolve the fast motion of the solvent molecules, whereas the evolution of the slower polymer molecules of the system requires a larger time-step. However, if the investigations concern only the nucleic acid and not the solvent molecules then they may be removed from the simulation and their effects on the polymer can be represented by dissipative and random force terms described above.

A Brownian dynamics simulation method has been used across many nucleic acid model length scales [166–169, 172–174, 184, 191–193, 195, 197, 206]. By this method there are a greatly reduced number of interaction sites once the explicit atoms or molecules of the solvent have been replaced by an implicit Brownian solvent. The time and length scales accessible depend on the model chosen to represent the nucleic acid only, and the characteristics of the fluid have little effect. By this method long DNAs (through coarse-grained models, described in Section 3.2.3) can be simulated for seconds of time.

Traditionally LD/BD systems do not include hydrodynamic interactions. This approximation is valid for certain systems, such as when hydrodynamic interactions are screened out, and this consideration must be carefully considered when choosing a simulation method. When necessary, there are methods of incorporating hydrodynamic interactions with LD and BD simulations of nucleic acids [191, 192].

3.3.5 Monte Carlo Simulation Method

While MD, DPD, MPC, and LD/BD simulation methods are all dynamic in nature, allowing the time evolution of the system to be observed; the broad class of Monte Carlo (MC) simulation techniques do not attempt to simulate the dynamics of a system [201, 227]. There are many types of Monte Carlo procedures; fundamentally, however it is a biased random walk in phase space. The nucleic acid polymer can be modeled at any level of detail from atomistic to generic entropic spring representations. With this method, components of the model are translated or rotated in space, the change in the potential is calculated, and then the move is either accepted or rejected. If accepted, the positions of the nucleic acid are updated; with either outcome the method is repeated. At one extreme, MC methods can only be used to study equilibrium properties, which are different from the dynamic methods (Sections 3.3.1, 3.3.2, 3.3.3, and 3.3.4) which give non-equilibrium, as well as, equilibrium properties. An assortment of additional techniques have been added to the general Monte Carlo scheme in order to approximate dynamic properties of interest [201, 227].

Metropolis Monte Carlo is the most widely used Monte Carlo method for nucleic acid simulations. In an NVT ensemble, with N interaction sites, a new configuration is created by arbitrarily moving bead(s) from position $i \rightarrow j$. Due to such model movement, it is possible to compute the change in the potential energy of the system ΔU:

$$\Delta U = U(j) - U(i), \tag{3.14}$$

where $U(i)$ and $U(j)$ are the potentials associated with the original and new configurations, respectively. This new configuration is then accepted or rejected according to the following rules. If $\Delta U < 0$, then the translation or rotation movement would bring the system to a state of lower energy. In one possible manner to sample the Boltzmann distribution, the movement is immediately accepted when the change in energy is negative. If $\Delta U \geq 0$, the move is accepted only with a certain probability $p_{i \rightarrow j}$ which is given by

$$p_{i \rightarrow j} \propto \exp\left(\frac{-\Delta U}{k_B T}\right), \tag{3.15}$$

where k_B is the Boltzmann constant. According to Metropolis *et al.* [228], if a random number ξ with uniform distribution between 0 and 1 is generated, then the

new configuration is accepted according to the following rule:

$$\xi \leq \exp\left(\frac{-\Delta U}{k_B T}\right), \text{the move is accepted,} \tag{3.16}$$

$$\xi > \exp\left(\frac{-\Delta U}{k_B T}\right), \text{the move is not accepted.} \tag{3.17}$$

If the new configuration is rejected, the positions are not updated and the process is repeated for another randomly chosen selection of interaction sites.

Monte Carlo simulations of nucleic acids [229–231] span all length scales, from atomistic [151], to coarse-grained [63, 175–177, 180, 190, 207], to continuum [152, 207]. Often dynamics are approximated in these nucleic acid simulations to gain more insight than mere equilibrium values [170, 196]. Monte Carlo simulations have some distinct advantages in biological systems since they are able to throughly sample the phase space and thus gather information concerning many different energy states. Understanding the complex energy surface for nucleic acid structures can give great insight into the fundamental nature of their structures and functions.

3.3.6 Multi-Scale Methods of Nucleic Acids

Classical approaches to simulations assume that a nucleic acid can be represented using Newton's laws and simple equations (the force field) relating the structure of the system with its energy. The level of accuracy in the representation of the nucleic acids, the solvent, and the force-fields describing them leads to a variety of simulation methods, some of which have been discussed previously. Each methodology is orientated to the study of different aspects of a biological system. At the smallest scale (microscopic), many functional aspects of nucleic acids depend on local, primary structure and structural details, whose analysis requires an atomistic, quasi-atomic, or sequence dependent coarse-grained model level of description. The classical microscopic methods are based on the calculation of the molecular energy or force for a given nuclear configuration using a force field. The difference between the methods are found in (i) the representation of the nucleic acid and solvent, (ii) the force field, and (iii) the post-processing of the energy information derived from the force field.

At the mesoscopic scale, despite the importance of understanding the molecular structure, parts of the system are homogenized with respect to different aspects (for ex-

ample, the fluid) which can be at different scales. The techniques of DPD and MPC, described in Sections 3.3.2 and 3.3.3, are by definition multi-scale in nature. The Langevin and Brownian dynamics techniques of Section 3.3.4 are in spirit also multi-scale as all of the eliminated fluid particles are represented by a single set of forces. Finally, at the macroscopic (or continuum) scale, further averaging and homogenizing to the system is done so that discrete components are ignored. Macroscopic models obey the fundamental laws of continuity, equilibrium, and the conservations of energy and entropy. The laws governing these models are coupled with the appropriate constitutive equations and equations of state. Computational approaches range from simple closed-form analytical expressions to complex structural mechanics [126].

In addition to the mesoscopic methods, combinations of other techniques can be used to better examine a particular nucleic acid system. Such multi-scale approaches typically employ a sequential series of simulation methods to examine particular details of a system. For example, as will be described in Chapter 7, a coupling of Brownian dynamics (with its long time and length scales) and molecular dynamics (with small but highly detailed time and length scales) can be done to gain insights into biological systems at relevant *in vitro* or *in vivo* conditions and experimental times. In addition, other method combinations have been used for nucleic acid study: Monte Carlo and Brownian dynamics routines [187, 205] and molecular dynamics and Lattice Boltzmann techniques [186].

In fact, the challenge in the future will be not simply to develop new and improved simulation techniques at individual time and length scales, but to also integrate the developed methods at a wider range of time and length scales, spanning the entire spectrum. As more multi-scale methodologies are constructed the connections and influences between scales will be better captured and understood; as was shown in Chapter 2, nucleic acid primary structure is fundamental to secondary and tertiary structural formations. Such development of multi-scale methods is crucial in order to achieve the longstanding goal of predicting global features from local effects [212].

3.4 Conclusions

The understanding of DNA molecules and the complex structures that they can form, as described in Chapters 1 and 2, necessitates a comprehensive understanding of nu-

cleic acid phenomena at different time and length scales. In the past decade and a half or so, this need has significantly stimulated the development of computer modeling and simulation, either as a complementary or alternative technique to experimentation. In this connection, many traditional simulation techniques (such as MC, MD, BD, DPD, and MPC) have been employed to study biological systems. These techniques and the models that they utilize indeed represent approaches at various time and length scales from molecular scale (atoms), to the micro/mesoscale (coarse-grained beads and particles), and then to the continuum or macroscale (domains), and have shown success to various degrees in addressing many aspects of DNA configurations.

In order to model the complex nucleic acid structures described in Chapter 2 we need to carefully consider which nucleotide model to choose. Since all of the features detailed in Chapter 2 are sequence dependent, we must choose either an atomistic or low level coarse-grained model, as shown in the upper section of Figure 3.1. Further, due to the fact that some of the structures involve the movement of large fragments of nucleotides we must choose a model that can represent at least 100 nucleotides; atomistic models at this scale are computationally expensive, therefore we must choose one of the functional group coarse-grained nucleic acid models. In addition, since we are interested in understanding the formation and folding pathways, a dynamic method is vital. With the large time scales necessary for large segments of polynucleotides to arrange in complex, three-dimensional formations, mesoscale simulation techniques must be utilized. Finally, although the structures described in Chapter 2 are undoubtedly affected by the solvent, the characteristics of the fluid in these structures is not of the first concern; therefore we chose a Brownian dynamics simulation method and either a two or three bead per nucleotide model for our future investigations.

We will begin, in Chapter 4, with the simpler of the two models, a canonical two bead depiction of each nucleotide and we will attempt to simulate the most simplistic of Chapter 2's structures, a single-stranded nucleic acid hairpin. Before moving on to more complex shapes, we will investigate and evaluate the two bead model with experimental data. It will be seen that the two bead DNA model fails to consistently capture the dynamics of single-stranded DNA hairpin behavior. Therefore, in Chapter 5 we will develop and experimentally validate a slightly more complex three bead per nucleotide model that will allow us to capture not only the sequence dependent features of both canonical (Watson–Crick) and non-canonical (Hoogsteen and wobble) hydrogen bonding and base stacking effects, but also polynucleotide chirality. It

will be shown that this model can not only capture single-stranded characteristics of hairpin melting, but also double-stranded structures such as B-Form DNA. Chapter 6 will demonstrate the full capabilities of this newly developed three bead DNA model with examples of double-stranded structures of P-Form dsDNA (Section 2.2.4), S-Form dsDNA (Section 2.2.5); triple-stranded structures of H-DNA (Section 2.4.1) and strand invasion (Section 2.4.2); and quadruple-stranded structures of G-quartets (Section 2.5.1). Finally, Chapter 7 will provide a multi-scale approach to a particular biological system, the 10-23 DNAzyme (Section 2.3.3), in which a coarse-grained Brownian dynamics simulation is coupled with an atomistic molecular dynamics simulation to understand particular global and local features of mechanistic importance. As will be discussed in detail in Chapter 7, many important processes in biology are inherently multi-scale.

4

Two Bead DNA Model and Verification Process

Observation and theory get on best when they are mixed together, both helping one another in the pursuit of truth. It is a good rule not to put overmuch confidence in a theory until it has been confirmed by observation. [...] it is also a good rule not to put overmuch confidence in the observational results that are put forward until they have been confirmed by theory.

Sir Arthur Stanley Eddington,
New Pathways in Science, 1934 [232]

As was seen in Chapter 3, the last two decades have witnessed a vast accumulation of biological models of nucleic acids. The theories put forth concerning DNA and RNA structure have been built in order to understand the fundamental principles governing nucleic acid behavior through reverse engineering techniques. However, the models have not been well studied with actual, consistent, and reliable experimental evidence. Without verifying these nucleic acid models by confirming them with data, it is difficult to place much importance on their conclusions. Indeed, as Eddington [232] explained when combined "observation and theory" can both be used to come to the aid of "one another in the pursuit of truth".

In this chapter we seek to develop a method of "mixing together" [232] a nucleic acid model with experimental data concerning simple oligonucleotide behavior. This method chooses to compare a system for which the experimental and simulation data can be adequately captured and for which the resultant data is simple yet informative. In particular, the ssDNA open-close melting transition is chosen as the particular study approach due to the fact that the behavior can be well captured by both dynamic and static modeling methods. Although we will only consider a dynamic

method, that of Brownian dynamics, as described in Section 3.3.4, the hairpin melting data can be used for a variety of simulation models and methods. We will focus on a relatively utilitarian nucleic acid model that is able to span a wide range of both length and time scales, as described in Chapter 3.

4.1 Required Features of the Nucleic Acid Model

As was introduced in Chapters 1 and 2, DNA structure possesses several levels of complexity, ranging from the sequence of bases (primary structure) to base-pairing (secondary structure) to its three-dimensional shape (tertiary structure). Models have been built at each of these levels and are able to describe some phenomena, but their ability to span multiple complexity levels is limited. Fundamentally, however, in order to be able to capture the base-scale resolution behavior imbedded in the primary structure of the nucleotides, and thus the base specific secondary and tertiary structural features described in Chapter 2, we must retain some individual base characteristics in the DNA model. As was described in Section 3.2, there are still many DNA model types that preserve this type of base identity.

It is also important to note that in order to fully capture higher order structures, a sufficient length of time must elapse. The time scale necessary to allow for large nucleic acid sections to arrange into complex patterns spans several orders of magnitude; this limits the prospective modeling techniques, as detailed in Section 3.3, that provide tractable models and methods for investigating the advanced structures of nucleic acids.

Coarse-grained modeling is an increasingly widespread method to study the dynamics of nucleic acids. As was discussed in Chapter 3, these approaches include Monte Carlo [63, 151, 170, 176, 229, 230], Brownian dynamics (BD) [168, 169, 172, 184], molecular dynamics (MD) [46, 123, 134, 136, 140, 142–144, 155, 158–160, 164, 171, 233], and lattice Boltzmann [192, 194, 234] methods. As was detailed in Chapter 3, for a given level of computational resources, there always exists a trade-off between the available length and time scales. Coarse-grained models allow access to relatively long time scales, albeit at a correspondingly coarse length scale [45, 160]. Yet, if we properly chose the length scale necessary for the particular application, we can find a reasonable balance between these two considerations.

4.2 Two Bead Nucleic Acid Model

For the intricate secondary and tertiary structures of short ($\ll$ 100 nucleotides) nucleic acids, it is reasonable to begin with a two bead descriptor of a nucleotide, as discussed in Section 3.2.3. This approach allows for the base identity and properties to be preserved (as one of the beads represents the nitrogenous base) along with the generic features of a oligonucleotide strand. In addition, we utilized a standard Brownian dynamics algorithm, discussed in Section 3.3.4 with this base-backbone single-stranded DNA model in order to be able to reach the long time scales necessary for complex structure conformations. The exact simulation conditions will be further detailed in Section 4.2.4.

For the basic structural unit of the simulated DNA, we adopted the model depicted schematically in Figure 4.1. The nucleic acid's backbone groups, the deoxyribose sugar and phosphate groups, are represented in a single bead, the backbone bead. Contiguous backbone beads are connected to one another. The nitrogenous bases A, C, G, and T are represented by a separate bead and connected to the backbone beads as schematically depicted in Figure 4.1.

The model, developed by previous group work [172], is written in terms of a dimensionless length σ and energy ϵ. The value of σ is fixed as the size of the beads; all beads are the same size. The degree of freedom embodied in ϵ is used to map the simulation reduced temperature T to the experimental temperature; this will be further discussed in Section 4.3.7. Each of the potentials depicted in Figure 4.1 and governing the interactions of the model can be sorted into three categories: nonspecific interactions occurring between all beads, backbone-backbone interactions, and base-base interactions. Throughout the experimental testing and model validation process described in this chapter, the general forms of the potentials will not change, however, specific parameters are often altered to understand overall model sensitivity and importance for experimental matching. For this reason, the potential equations are given in their more general forms with particular parameter values listed below each equation. These parameters are often adjusted individually or in tandem depending on their type, the interactions involved, and the underlying physics.

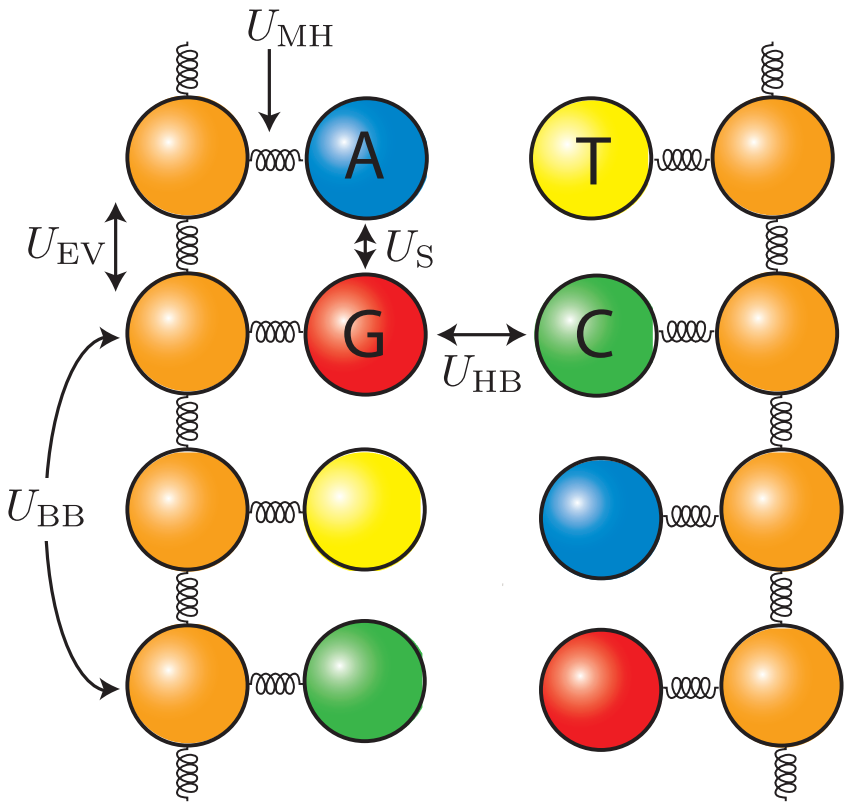

Figure 4.1: A schematic representation of a two bead coarse-grained model of DNA. Each nucleotide is represented by two beads, one for the phosphate and sugar groups and one for the nitrogenous base; A (blue), C (green), G (red), T (yellow)). The backbone beads (orange) are contiguous for each nucleic acid strand. The various interactions are labeled between the beads: all beads have excluded volume ($\mathrm{U_{EV}}$) interactions; connected beads have modified harmonic ($\mathrm{U_{MH}}$) interactions, backbone beads have a bending potential ($\mathrm{U_{BB}}$) between them and the bases have both hydrogen bonding ($\mathrm{U_{HB}}$) and stacking ($\mathrm{U_S}$) interactions.

4.2.1 Nonspecific Interactions

In general, the spacing between the backbone group and the base beads are enforced by a combination of excluded volume interactions and modified harmonic springs, defined by Equations 4.1 and 4.2, respectively; the sum of these potentials creates a relatively deep well that minimizes the fluctuations in these distances [172]. In the model every bead interacts with the other beads by excluded volume interactions in order to provide each bead with a physical size. For each bead i, the interaction with bead j is given by the truncated pairwise Lennard-Jones potential,

$$U_{\mathrm{EV}}(r_{ij}) = \begin{cases} 4\epsilon\left[\left(\dfrac{\sigma}{r_{ij}}\right)^{12} - \left(\dfrac{\sigma}{r_{ij}}\right)^{6}\right] + \varepsilon, & r_{ij} \leq r_c \\ 0, & r_{ij} > r_c \end{cases} \tag{4.1}$$

where r_{ij} is the distance between beads i and$j$ and r_c is the minimum in the potential [172, 213],

The connectivity between the various species (backbone-backbone and base-backbone) is achieved with the use of a modified harmonic potential, which ensures finite extensibility [172, 235]. This potential is given by the so-called finitely extensible nonlinear elastic potential (FENE),

$$U_{\mathrm{MH}}(r_{ij}) = -\frac{\kappa}{2} R_0^2 \ln\left[1 - \left(\frac{r_{ij}}{R_0^2}\right)^2\right]. \tag{4.2}$$

In Equation 4.2, κ is the effective strength of the potential and R_0 is the cutoff of the potential. We use the values $\kappa = 30.0\epsilon/\sigma^2$ and $R_0 = 1.5\sigma$, independent of the tip of bond [172, 235]. Although the chemistry of the bonds connecting backbone-backbone (the phosphodiester bridge described in Section 1.2.3) and the base-backbone (the β-glycosyl C_1,-N linkage described in Section 1.2.2), are rather different, the choice of parameters yield fairly stiff springs. At our level of coarse-graining, stiff springs are sufficient to describe both bonds, as we are mostly concerned about the connectivity between different parts of the backbone, rather than a detailed examination of the particular internal modes of the molecule. The combination of excluded volume and spring forces maintains a relatively constant extension between bonded bead pairs.

4.2.2 Backbone-Backbone Interactions

To provide some additional stiffness to the backbone of the single-stranded DNA, we utilized a bending potential,

$$U_{\mathrm{BB}}(r_{ij}) = \frac{u_b}{2}(\cos\theta + \cos\theta_0)^2 \tag{4.3}$$

applied only between contiguous backbone beads. The parameter θ_0 is the desired bending angle along the backbone, and u_b is a tunable parameter dictating the effective stiffness of the molecule; $u_b = 0$ implies a freely jointed chain while $u_b \gg 0$ implies a stiff molecule [155, 168, 169]. Here we use a straight (equilibrium angle of π) $\theta_0 = 0$ bending angle and a $u_b = 18\epsilon$ [172]. These parameters give a persistence length for the molecules on the order of a few nucleotides. This value, as a rough approximation, is relatively in line with the persistence lengths for single-stranded DNA discussed in Section 2.3.

4.2.3 Base-Base Interactions

The sequence dependent structure is captured by base specific potentials. In this basic model of DNA we seek only to include the most indispensible base identity features of the molecule, mainly its Watson–Crick hydrogen bonding and its base stacking interactions, as discussed in Sections 1.3.1 and 1.3.2, respectively. The sequence dependent interactions have the generic form

$$U_{\mathrm{k}}(r_{ij}) = -\epsilon u_{\mathrm{k}} \delta_{\mathrm{k}}^{ij} \left[\exp\left(20\left(\frac{r_{ij}}{\sigma} - \Gamma\right)\right) + 1\right]^{-1}, \tag{4.4}$$

where k denotes whether it is a hydrogen bonding or a stacking interaction. This particular form of the potential has been used elsewhere [172, 184] to model hydrogen bonding and stacking in DNA. The parameter u_{k} gives the overall strength of the stacking to hydrogen bonding interactions [12, 13, 184] and will be used in the validation process as a tunable parameter. The individual δ_{k}^{ij} gives identity to each base and describes how a base of type i and a use of type j interact with one another through either stacking or Watson–Crick hydrogen bonding interactions. The parameter $\Gamma = 1.5$ is used for both the stacking and hydrogen bonding interactions through this first iteration of the model.

The two bead model uses a general approximation of hydrogen bonding strength for

Table 4.1: The base specific hydrogen bonding parameters, $\delta_{\mathrm{HB}}^{ij}$ for the hydrogen bonding (Watson-Crick only) in the two bead coarse grain model of DNA.

	A	C	G	T
A	0	0	0	2/3
C	0	0	1	0
G	0	1	0	0
T	2/3	0	0	0

the Watson–Crick type bonding. As was explained in Section 1.3.1 and shown in Figure 1.6, a A · T Watson–Crick base pair has two hydrogen bonds while a G · C Watson–Crick base pair has three. As an approximation of the strength of the interaction, even though hydrogen bonding strength is not additive, we simply count the number of hydrogen bonds; a A · T base pair being two thirds the strength of a G · C base pair. Therefore, the effective hydrogen bonding is given by the anti-diagonal matrix of Table 4.1. The hydrogen bonding of some base bead i is computed with all other base beads $j \neq i$, which allows the model to move smoothly between secondary structure in ssDNA and dsDNA.

As can be seen in Table 4.1, only A · T and G · C base bead pairs interact by hydrogen bonding with one another. We exclude hydrogen bonding interactions between contiguous base beads, even if they are an A · T or G · C pair. We also exclude hydrogen bonding interactions between base i and base $j = i \pm 2$ to prevent the formation of three membered rings. This choice is motivated by the heuristic criteria one observes in the M-Fold server [68] where such rings are normally excluded when one employs moderate values of buffer ionic strength. In the present model, the isotropy of the hydrogen bonding interactions can, in principle, lead to small-membered rings through an unrealistic twisting of the backbone chain and interactions through the phosphate backbone or along oblique angles that should not, in reality, lead to significant hydrogen bonding interactions. The specific issues of pseudo-knots and other unrealistic ssDNA formations will be discussed further in this chapter and in Section 5.1.3.

The two bead model also uses a rudimentary approximation of stacking strength. As was explained in Section 1.3.2, there are many aspects that affect the base stacking strength of two nucleotides, but a generalized ranking for the stacking interactions can be considered to be

$$\text{purine - purine} > \text{pyrimidine - purine} > \text{pyrimidine - pyrimidine.}$$

These qualitative guidelines can be used for the rough stacking interactions provided

Table 4.2: The base specific stacking parameters, δ_{S}^{ij} for each stacking combination possible.

	A	C	G	T
A	3/4	1/2	3/4	1/2
C	1/2	3/4	3/4	1/4
G	3/4	3/4	1	1/2
T	1/2	1/4	1/2	1/4

in Table 4.2, with the strongest stacking occurring between two guanines (purines) and the weakest between cytosine and thymine or thymine and thymine (pyrimidines). Since the model only has two beads, it does not have directionality and therefore the stacking table is diagonally symmetric. The two bead model does not included any cross-stacking interactions.

4.2.4 Simulation Algorithm

The potentials detailed above are incorporated into a standard Brownian dynamics algorithm. We scale the length with σ, the energies with ϵ, and the time with $\tau \equiv \xi\sigma^2/\epsilon$, where ξ is the bad friction coefficient. The friction of each bead is identical and there are no hydrodynamic interactions between the beads. Given the positions $\mathbf{x}_i(t)$ of all of the N beads in the system, the position dependent energy of bead i is given by the sum of Equations 4.1, 4.2, 4.3, and 4.4,

$$U_i(\mathbf{x}_i(t)) = U_{\mathrm{EV}} + U_{\mathrm{MH}} + U_{\mathrm{BB}} + U_{\mathrm{k}}. \tag{4.5}$$

where the potentials on the right-hand side refer to the relevant values for bead i. The equations of motion governing the time-dependent positions $\mathbf{x}_i(t)$ of each bead in the simulation are given by the Langevin equation

$$\frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} = -\frac{\partial U}{\partial \mathbf{x}_i} + \sqrt{\frac{2T}{\Delta t}}\mathbf{r}_i \tag{4.6}$$

where $\mathbf{x}_i$ are the dimensionless bead positions, $T = k_B T(K)/\epsilon$ is the dimensionless temperature in terms of Boltzmann's constant, k_B, and the dimensional temperature, $T(K)$, and Δt is the dimensionless time step. The random numbers, $\mathbf{r_i}$, are Gaussian with mean zero and unit variance. The stochastic differential equation is integrated using a predictor-corrector scheme [224]. We only report time-independent data (qualitative trajectories or thermodynamic data), rendering the choice of friction

coefficient one of convenience since it is adsorbed into the time constant. Typical time steps are 0.1τ or 0.01τ and the bead positions are either saved every 100τ or 1000τ, depending on the polymer characteristics examined.

4.3 Validation of Two Bead DNA Model

At the heart of any DNA simulation model are the potential energy functions used to quantify the interactions between different parts of the chain. These functions and their parameters, given above, need to be selected to capture the relevant physical properties of single-stranded DNA. Careful selection and testing must be done to understand the quality of any given DNA model.

At the most generic and macroscopic level, we would expect a robust nucleic acid model to be able to capture thermal denaturation (melting). As a prototypical example, we consider here the case of a DNA hairpin. At the minimum, the model should at least lead to the correct melting-point temperature, i.e., the temperature at which there is a 50% probability of locating the hairpin in the open state. This is an experimentally accessible and relevant descriptor of DNA behavior. The melting point temperature is a function of ionic strength, which provides an approach to tune the parameters for different experimental conditions [160, 176]. At the next level of complexity, we would like the model to mimic the entire dependence of the sigmoidal melting temperature curve from the 100% closed state at low temperature to the 100% open state at high temperatures. In particular, capturing the "shoulders" of the melting temperature curve near the fully closed and fully open states is particularly challenging. A model that passes this test, especially in a biologically relevant buffer, could be used with confidence in more complex *in vivo* scenarios.

4.3.1 DNA Hairpin Melting Experimental Method

In order to provide a stringent test of the model [172–174], we needed a large set of consistent experimental data. Since the melting transition of single-stranded DNA hairpins was chosen as a basic test for the general behavior of the DNA model at the relevant detail scale, we first looked to the literature for experimental hairpin melting curves and temperatures. After exhaustive literature searches, we found that the cited values varied greatly amongst one another (even for very similar sequences)

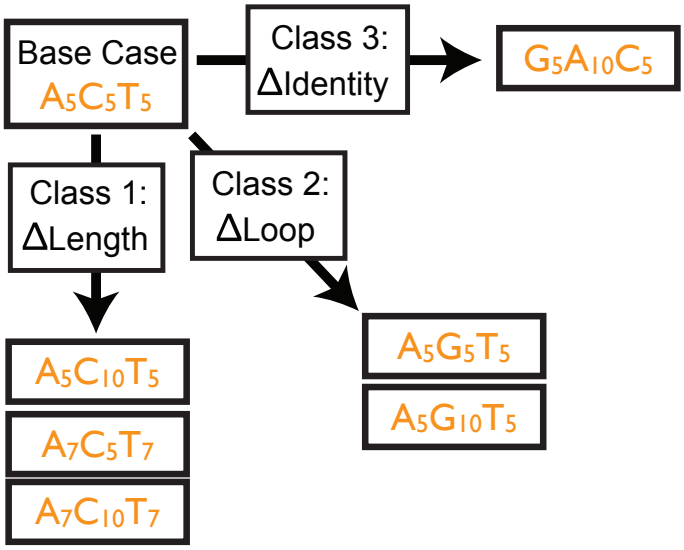

Figure 4.2: The ssDNA hairpin base case sequence of 5'-$A_5C_5T_5$-3' is used in a series of melting experiments. There are three classes of hairpin variants: Class 1 alters the stem and/or loop lengths; Class 2 modifies the loop identity; and Class 3 varies the identity of the bases in the stem and loop. A total of seven hairpins are examined for their melting characteristics.

and speculated that the melting point temperature differences (some as large as 30 K) were due to the specific experimental conditions [12, 13, 236, 237].

Therefore, in order to provide a rigorous test of the model [172–174], we obtained a large set of experimental data under well-controlled conditions. We examined a base case of $A_5C_5T_5$ and select variants, which are grouped in the three classes depicted in Figure 4.2 and listed in Table 4.3: (1) In the first class we retain the nucleotide sequence $A_XC_YT_X$ while varying the values of X and Y; (2) In the second class, the stem identity is conserved but the cytosine bases in the loop are replaced with guanines of different lengths; (3) In the third class, the stem bases were changed to guanine and cytosine with adenine and thymine as the loop bases. To simplify the subsequent discussion of the data analysis we assigned each sequence the value $j = 1 : 7$ appearing in Table 4.3. In addition to the sequences in Table 4.3, we also considered several other sequence variants that fall into these classes ($A_{10}G_{20}T_{10}$, $A_5G_{20}T_5$, $A_{10}C_{20}T_{10}$, $C_{10}T_{20}G_{10}$, $G_5T_{10}C_5$, $A_{10}C_5T_{10}$, $A_{12}C_5T_{12}$, $A_{12}C_{10}T_{12}$) but these were rejected due to high melting point (T_{MP}) temperature or other experimental difficulties (such as low synthesis yields, especially with poly-guanine sequences). Each single-stranded DNA sequence was obtained from IDT (Integrated DNA Technologies) and HPLC purified by the manufacturer prior to use. The lyophilized powder was serially diluted with

Class	Type	Sequence	j	T_{MP}(*mfold*)
	Base Case	$A_5C_5T_5$	1	319.1 K
1	ΔLength	$A_5C_{10}T_5$	2	308.5 K
		$A_7C_5T_7$	3	330.6 K
		$A_7C_{10}T_7$	4	323.1 K
2	ΔLoop	$A_5G_5T_5$	5	331.8 K
		$A_5G_{10}T_5$	6	324.3 K
3	ΔIdentity	$G_5A_{10}C_5$	7	347.1 K

Table 4.3: List of single-stranded DNA sequences. The Δ classes refer to the change in stem length (1), loop (2), or sequence (3) when compared to the base case $A_5C_5T_5$. The index j will be used throughout to denote the DNA hairpin type. The *mfold* predicted melting-point temperatures were found for 1μM monovalent sodium, similar to Buffer A, and are given for the aligned stem bonding configuration.

DI water to a concentration of 1 μM to make a stock solution.

A Quantitative Polymerase Chain Reaction (QPCR) machine (Mx3000P, Stratagene) was used to collect the temperature and fluorescence intensity data for the DNA hairpins in Table 4.3. Although acquiring melting curve data is not the standard use of a QPCR system, the machine's accurate temperature control, ability to excite and detect fluorescent molecules, and its 96 sample capacity make it well suited for our experiments. However, this equipment choice led to some experimental restrictions. For example, the DNA hairpins must have relatively low melting-point temperatures ($T_{MP} < 353$ K) due to the fact that (i) the buffer experiences significant vaporization in the upper temperature range, which can lead to inaccurate fluorescence readings due to condensation on the cap of the sample tube, and (ii) the QPCR system is not designed to capture temperatures greater than 363 K. We only used sequences with *mfold* [68] predicted melting-point temperatures $T_{MP} < 353$ K (1 M monovalent sodium, similar to Buffer A) to ensure a sufficient dynamical range to capture the upper plateau of the melting curve. The *mfold* predicted melting-point temperatures are summarized in Table 4.3. This restriction limits the possible sequences to be examined experimentally.

The experiments were conducted in a biologically relevant buffer, Buffer A [99] at 1X concentration (0.05 M HEPES, 0.5 M NaCl, 0.5 M KCl; Ionic Strength = 1; pH = 7.1 at 25° C). This standard composition of Buffer A is a reasonable model for *in vivo* conditions and has proven useful for *in vitro* applications as well; for example, Buffer A at these specifications was used in the initial evolution of the 10-23 DNAzyme [55, 99]. While appropriate for studying basic physics, a monovalent salt poorly captures the effects of a complex biologically relevant buffer, like Buffer A.

Well Type	Well Contents			Well Number, k
	Buffer A	SYBR Green I Dye	DNA Hairpin	
Control Buffer Well	1X	0	0	$k_{buffer} = 1 - 4$
Control Dye Well	1X	2X	0	$k_{dye} = 5 - 12$
Experimental Signal Well j	1X	2X	2.2 μM stem base pairs	$k_{DNA} = 1 + 12j$

Table 4.4: List of well types randomly loaded on each 96 well plate. One of the seven DNA hairpins ($j = 1 - 7$) under study was loaded into each experimental signal well. The number of replicates of each well type per plate is the number of well numbers, k, corresponding to that experimental condition.

Three different types of wells (Control Buffer Wells, Control Dye Wells, and Experimental Signal Wells), each containing 50 μL of solution, were arranged randomly on each 96 well plate. The contents of each well are summarized in Table 4.4. The control buffer wells each contained only a 1X Buffer A solution and were used to measure the background fluorescence signal of the biological salt solution. The control dye wells contained a 1X Buffer A solution and a 2X SYBR Green I dye solution. The SYBR Green I fluorescence signal when bound to double stranded DNA is 800- to 1000-fold greater than when bound to single stranded DNA (Molecular Probes). This well type was used to measure the background fluorescence signal of the unbound SYBR Green I intercalating dye. In the experimental signal wells, the 1X Buffer A and 2X SYBR Green I dye solutions were combined with 2.2 μM of DNA hairpin stem base pairs. These optimal concentrations of stem base pairs and dye were determined by an independent set of measurements, described in Section 4.3.3, of the melting of the $A_{10}C_5T_{10}$ sequence over a range of SYBR I concentrations (0.02 X to 10 X) and DNA concentrations (0.02 μM to 0.2 μM). The replicates of each well type are enumerated in Table 4.4. Two identical plates were made and each plate was run twice to generate the experimental intensity data.

We tested temperature ramps of $\delta T = 1$ K/5 min and $\delta T = 1$ K/min and did not notice any significant change in the data when the fluorescence versus temperature curves obtained at each ramp rate were overlaid. Most of the results reported here were obtained with $\delta T = 1$ K/min as the temperature transition rate. Fluorescence data were collected over the range of 293 K to 363 K. The fluorescence signal at a single temperature was measured for each well 14 times before raising the temperature by one degree. The specifications of the QPCR system allowed for temperature plateaus to be programmed in increments of 1K while the actual temperature has a precision of $\pm$ 0.1 K. The plate was allowed to equilibrate for 5 minutes once the new temperature

was reached before obtaining fluorescence data. This process was repeated for each temperature in the specified range, creating over 80,000 measurements per plate. The resultant raw data consists of the temperature and corresponding fluorescence intensity (in arbitrary units) of each sample.

Each 96-well plate was loaded in a randomized manner in order to control for the known edge effects in QPCR systems. The wells consist of a series of control buffer wells, control dye wells, and experimental signal wells. All data collected from areas of the plates deemed inconsistent was eliminated from further analysis if it satisfied either of the following criteria: (i) the signal strengths were less than ten percent or greater than 500 percent of the average signal strength at the corresponding temperature, or (ii) the signal variance was greater than ten times the average signal variance at the corresponding temperature.

4.3.2 DNA Hairpin Melting Experimental Data Analysis

For each well, we binned the raw intensity data into temperature increments of 1 K and then averaged the 14 data replicates in each bin to produce an intensity $I_k(T)$ for each of the $k = 1:96$ wells, where T is measured in $\Delta T = 1K$ increments. For the control buffer wells, $k = k_{buffer} = 1:4$ on a single plate, the intensity signals were averaged to create a plate specific background signal corresponding to the specific characteristics of Buffer A. This produced a plate specific background signal,

$$B_{plate}(T) = \langle I_k \rangle_{1:4}, \tag{4.7}$$

such as the one in Figure 4.3a. In the latter and what follows, $\langle\ \rangle_k$ represents an average of the k wells. The background fluorescence signal, $B_{plate}(T)$, is subtracted from the raw data for the control dye wells ($k = k_{dye} = 5{:}12$) and the DNA containing wells ($k = k_{DNA} = 13{:}96$) to create the corrected intensity

$$I'_k(T) = I_k(T) - B_{plate}(T). \tag{4.8}$$

After correcting for the background signal, the data from the control dye wells ($k = k_{dye} = 5{:}12$), were similarly smoothed into an average curve to form the **u**n**b**ound **S**YBR **G**reen I, background corrected, fluorescence corrected signal

$$\langle ubSG(T) \rangle = \langle I'_k \rangle_{5:12} \tag{4.9}$$

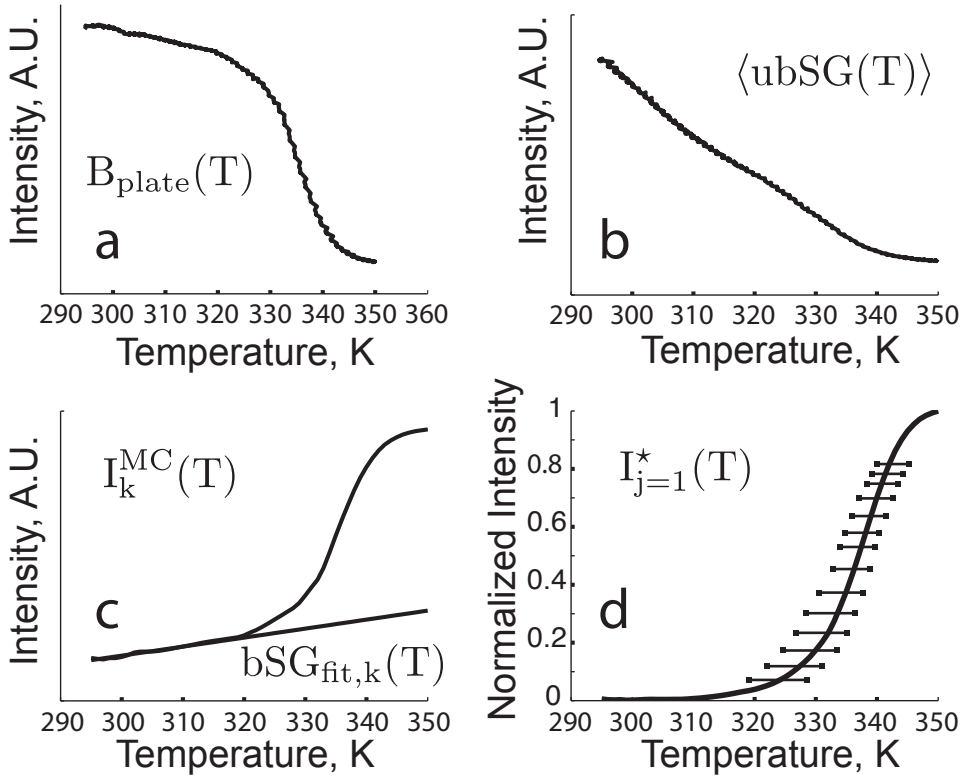

Figure 4.3: Postprocessing of the experimental data for the $A_5C_5T_5$ (j = 1) sequence. (a) Average background fluorescence signal, $B_{plate}(T)$. (b) Average background corrected unbound SYBR Green I temperature dependence, $\langle uSG(T) \rangle$. (c) An example well's melting curve, $I_k^{MC}(T)$; the temperature dependence of the bound SYBR Green I signal is fit from the closed hairpin state, $bSG_{fit,k}(T)$. (d) Normalized and averaged replicates for the j = 1 sequence. The temperature variation range is reported at selected normalized intensity values.

seen in Figure 4.3b. The ubSG(T) curve was fit with a linear temperature relationship, $\mathrm{ubSG_{fit}(T)}$

$$\mathrm{ubSG_{fit}(T) = a_{fit}T + b_{fit}}, \tag{4.10}$$

where $\mathrm{a_{fit}}$ and $\mathrm{b_{fit}}$ are the coefficients of the linear regression for the averaged unbound SYBR I signals. The linear SYBR I background fluorescence, $\mathrm{ubSG_{fit}(T)}$ is subtracted from the raw data for the DNA containing wells $(\mathrm{k = k_{DNA}}) = 1+12j$, where $j = 1:7$ counts the seven DNA hairpin types,

$$\mathrm{I''_k(T) = I'_k(T) - ubSG_{fit}(T)}. \tag{4.11}$$

The value $\mathrm{I''_k(T)}$ thus corrects for the background fluorescence signal of the buffer and the signal due to the excess amounts of intercalating dye.

The maximum of each $\mathrm{I''_k(T)}$ curve, $\mathrm{I_k^{max}}$, was found for each of the $\mathrm{k_{DNA}}$ wells. Applying the melting curve (MC) algorithm,

$$\mathrm{I_k^{MC}(T) = I_k^{max} - I''_k(T)}, \tag{4.12}$$

transforms the data into the standard melting curve format prevalent in literature (though not yet normalized on $[0, 1]$). This melting curve has the familiar sigmoidal shape, with high temperatures corresponding to high intensity values and low temperatures corresponding to low intensity data.

In addition to the unbound dye, we also needed to correct for the temperature dependence of the bound SYBR Green I. To make this correction, we assumed that: (i) the stems are all fully closed between 298 and 303 K, (ii) the bound SYBR I fluorescence depends linearly on temperature, similar to the unbound dependence measured in Figure 4.3b, and (iii) the fluorescence intensity is proportional to the number of bonded base pairs. From the low temperature (and thus closed state) data, an extrapolation of the **b**ound **S**YBR **G**reen I background signal

$$\mathrm{bSG_k(T) = a_kT + b_k} \tag{4.13}$$

was formed, as shown in Figure 4.3c, for each of the DNA containing wells $\mathrm{k = k_{DNA}}$.

We then removed each bound SYBR I contribution from the corresponding fluorescence signal at every temperature using the iterative approach described in Equations

(4.14) - (4.16). We first approximated the fraction of **c**losed **b**ase **p**airs as

$$\mathrm{cbp_k(T) = 1 - I_k^{MC}(T)/I_k^{MCmax}}, \tag{4.14}$$

where $\mathrm{I_k^{MCmax}}$ is the maximum value for the adjusted intensity $\mathrm{I_k^{MC}}$. We then subtracted the scaled extrapolated bound dye intensity, given by

$$\mathrm{I_k^{scale}(T) = cbp_k(T) \times bSG_k(T)}, \tag{4.15}$$

from every data point to arrive at a new intensity value for $\mathrm{I_k}$,

$$\mathrm{I_k^{\star}(T) = I_k^{MC}(T) - I_k^{scale}(T)}. \tag{4.16}$$

In the algorithm, $\mathrm{I_k^{\star}}$ is set equal to $\mathrm{I_k^{MC}}$ in Equation (4.14) and the algorithm of Equations (4.14) - (4.16) is iterated until the difference between $\mathrm{I_k^{\star}(T)}$ and $\mathrm{I_k^{MC}}$ is less than 1×10^{-8}. Once convergence is reached for each $\mathrm{I_k^{\star}(T)}$ value, where $\mathrm{k = k_{DNA} = 1 + 12j}$ for each DNA sequence j, the twelve DNA replicates are averaged to form $\widehat{\mathrm{I_j^{\star}(T)}}$,

$$\widehat{\mathrm{I_j^{\star}(T)}} = \langle \mathrm{I_k^{\star}(T)} \rangle_{1+12j}. \tag{4.17}$$

The $\widehat{\mathrm{I_j^{\star}(T)}}$ is then normalized on $[0, 1]$. Figure 4.3d shows the data $\widehat{\mathrm{I_1^{\star}(T)}}$ (for the $\mathrm{A_5C_5T_5}$ sequence). Each 96 well plate was run twice through the QPCR equipment and similarly processed. Each plate was also prepared in duplicate and this data was similarly processed; the four $\widehat{\mathrm{I_j^{\star}(T)}}$ values thus amassed where finally averaged to create a single melting curve for each of the seven DNA hairpins under study.

4.3.3 DNA Hairpin Melting Experimental Optimization

Through our investigations of the vast literature of oligonucleotide melting behavior, we understand that the experimental melting temperatures are highly dependent on the experimental conditions. Before we could begin collecting the high density experimental data that we would need to validate the model, we first needed to find the optimal experimental conditions for melting single-strand DNA hairpins. We found that the concentrations of the DNA and the SYBR I dye solution can shift the measured curve by as much as 18 K over the range of SYBR I concentrations (0.02X - 10X) and the DNA concentrations (0.02 - 0.2 μM) examined, consistent with other

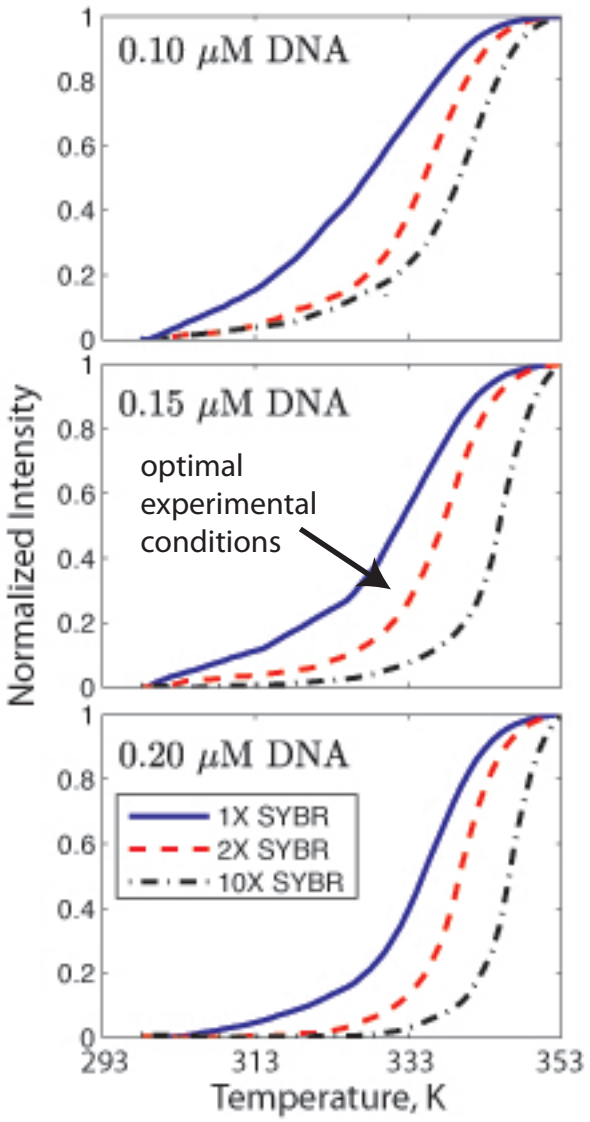

Figure 4.4: Plot of the normalized fluorescence intensity as a function of temperature for different DNA and SYBR Green I dye concentrations with the sequence $A_{10}C_5T_{10}$. The (red) dashed line in the center plot was chosen as the experimental condition for all subsequent studies: 0.15 μM DNA solution and 2X SYBR Green I dye. This sets the ratio of dye molecules to stem base pairs at 2X: 4.5×10^{13} stem base pairs.

experimental studies [12, 13, 236, 237]. To determine the experimental conditions that maximize the amount of useful data, we first sorted the *mfold* [68] predicted melting-point temperatures for a variety of small single-stranded DNA hairpins (ranging from 319.1 to 347.1 K), some of which are listed in Table 4.3. The parameters chosen for the initial *mfold* sorting were 1M monovalent sodium ions, which is similar to the 0.5M sodium and 0.5M potassium monovalent ions primarily comprising the biologically relevant buffer, buffer A [55]. The DNA hairpin sequence $A_{10}C_5T_{10}$ has the median

mfold predicted T_{MP} of 338.5 K [68]. We then examined this sequence using various combinations of the DNA and dye concentrations described above. Figure 4.4 shows nine of the 30 results thus obtained; the optimal set of concentrations for this sequence was found to be 2X SYBR I and 0.15 μM of $A_{10}C_5T_{10}$ DNA solution. We found that the melting-point temperature $T_{MP}^{exp} = 337$ K, under these conditions, is closest to the predicted *mfold* melting-point temperature $T_{MP}^{mfold} = 338.5$ K [68].

By generating these data with control over the specific experimental conditions and concentrations, we are better able to interpret the raw data and process it in a manner that provides a good correspondence with the computational simulations. In addition, after understanding that the specific experimental conditions (such as reagent concentrations) can shift the melting curve by as much as 18 K over a relatively small reagent parameter space, we found it vital to collect our own experimental data. Although a literature search can produce melting-point temperature data for a variety of DNA sequences, a highly reliable, self-consistent data set over the entire temperature range is necessary to arrive at meaningful conclusions concerning the quality of the simulation data.

4.3.4 DNA Hairpin Melting Simulation Method

At the start of the simulations, the molecule is initialed into one of two conformations illustrated in Figure 4.5: a linear chain (open configuration) or a square U chain (almost closed configuration). To ensure relaxation from this initial state, the simulations were carried out for 5 $\times 10^6$ BD time steps to erase memory of the initial configuration. We then obtained configuration data for approximately 5 $\times 10^8$ BD time steps with a time step of $\delta t = 0.01$, where a single time step corresponds to approximately a nanosecond. This total simulation times leads to a sufficiently large number of binding and unbinding events when $T \approx T_{MP}$, allowing reliable measurements of the phenomenon. Independent runs with different initial conditions were performed to ensure robustness with respect to the starting conformation of the molecules.

The simulations use a non dimensional temperature, but prior to this investigation we did not have an experimentally validated conversion to a real temperature. To limit the dimensionless temperature phase space that we needed to explore, we used a two step procedure. In the first step, we noted that the maximum characteristic hydrogen bonding potential in the model is approximately $0.123(\epsilon\ u_{HB})$ [172]. This

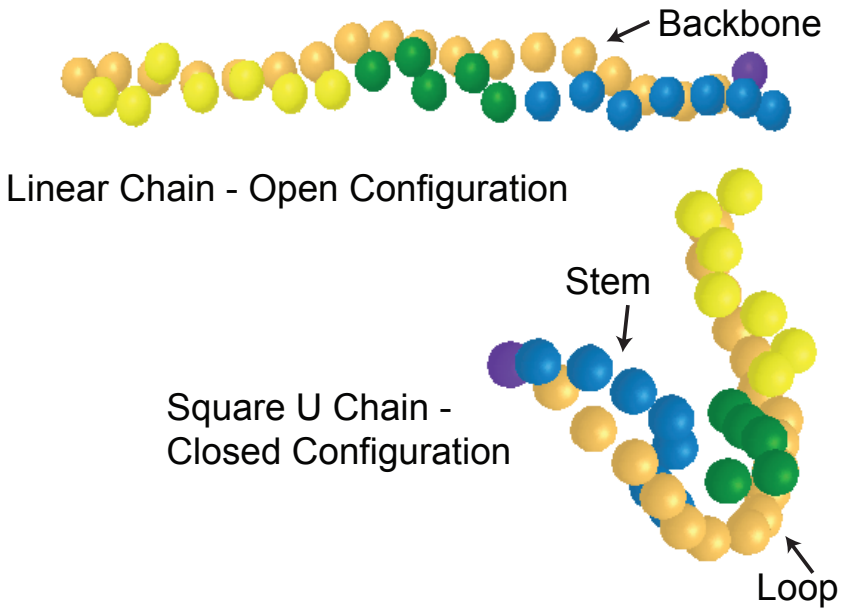

Figure 4.5: Example of a two bead model of the single-stranded sequence $5'-A_7C_5T_7-3'$. The chain is comprised of a series of contiguous backbone (orange) beads and a series of base beads: adenine (blue), cytosine (green) and thymine (yellow); the dark (purple) bead marks the $5'$ terminal of the chain. The chain is initialized in either a linear or square U configuration. The Watson–Crick base pairing stem and noninteracting loop sections of the hairpin are labeled.

leads to a characteristic energy of ϵ = -6.1 kcal/mol. With a choice of $u_{HB} = 2$ we find a temperature conversion of $T = 0.1 \rightarrow 310$ K. By first narrowing the temperature space through these estimates, we were able to make initial sweeps of simulations at non dimensional temperatures in the range of 0.1 - 0.6 in increments of 0.05. In the second step, we analyzed these data with the metric described in Section 4.3.5 and then fine tuned the temperature search near the melting point, and thus the transition region of the hairpin. Having narrowed the temperature range to within $\pm$ 0.1 of the melting point of the simulated sequence, the model was examined in more detail within this regime in temperature increments of 0.005. Additional simulations were added if the high temperature plateau was not sufficiently stable and clearly delineated. Three independent simulations were completed at every temperature point. The position of each of the simulated beads was saved every 1000 BD time steps (≈ 1 μs).

The complete parameter space that explicitly describes the model is comprised of 41 variables, some of which are discrete [172–174]. In this study, a small section of the parameter space was explored by varying the magnitude of the base stacking energy, ϵ_S, and the hydrogen bonding energy, ϵ_{HB}, while keeping the ratio of these quantities

fixed; data obtained for values $\epsilon_S/\epsilon_{HB} = 2.5/1$, $5/2$, and $10/4$ are presented here. We kept the ratio of the stacking and hydrogen bonding energies constant at the values approximated by experimental studies [12, 13], but allowed the magnitude of these quantities to change in relation to the other energies of the system. The system is written in reduced units with a non dimensional temperature of T.

4.3.5 DNA Hairpin Melting Simulation Data Analysis

We considered three different bonding metrics to determine the state of the single-stranded DNA hairpin in the simulation; that is, whether the hairpin was open or closed. All three metrics rely on the number of complementary interactions between the bases in the stem section of the nucleotide sequences. Complementary interactions were defined as follows: if two Watson–Crick complementary bases were found to be within a distance where the hydrogen bonding potential is effectively nonzero (since it decays quickly as a function of distance) then the two bases are considered bonded. This threshold distance is defined as $\sigma = 2^{1/6}$.

In metric 1 of Figure 4.6, all possible bonding pairs are calculated every 1000 BD time steps. There is a problem with this metric, especially for the block polymer-like sequences considered here. Since the model allows for multiple beads to bind to the same base bead via hydrogen bonding interactions, the number of bonding incidents described by this metric is often more then the number of bases in the stem section of the DNA hairpin. This artifact of two and three bead simulation models is prevalent throughout the literature [63, 160–162, 168–171, 238]. When we analyzed the instantaneous chain configurations, we found that the beads are often in a closed, zipper like configuration with the base beads slightly out of alignment [161, 162, 238]. This simulation artifact is due to steric and hydrogen bonding stabilization of this closed state form. Although the model does allow multiple binding incidents to occur, the change in free energy in the misaligned zipper configuration is not simply equal to twice the energy in the aligned system because the distances between the bonded beads are different.

Due to the shortcomings of metric 1, we designed additional metrics that could characterize the bound state. Metric 2, depicted in Figure 4.6, only considers bases to be bonded if they are paired "correctly" to lead to complete bonding in the stem. For example, for a sequence that is n bases long, the distance between the positions of the $i = 1$ bead and the $j = n$ bead is calculated (providing that the two bases in

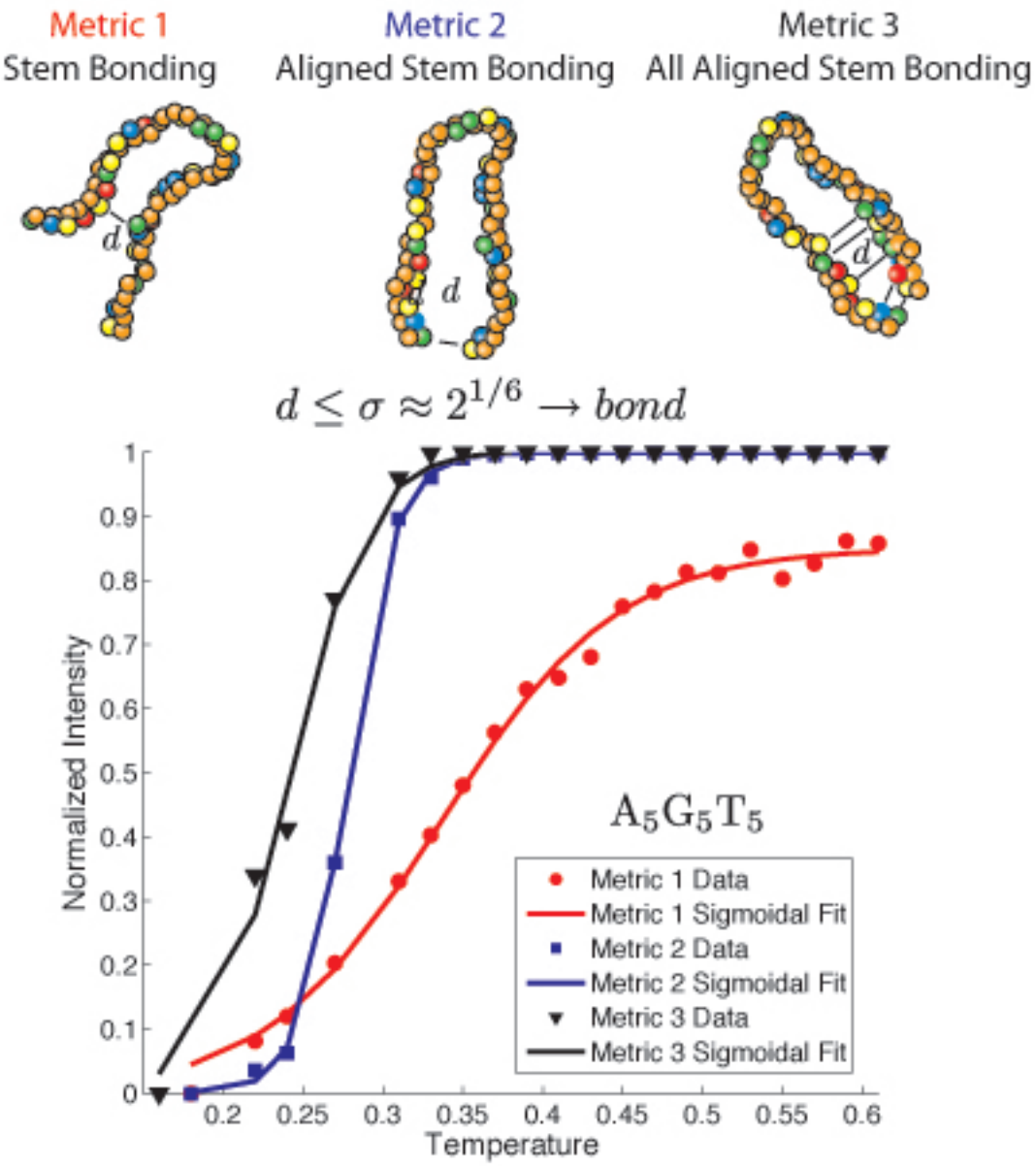

Figure 4.6: The three metrics used to determine the "closed" state in the simulation data. Metrics 1 and 2 enumerate all stem possible bonding pairs (1) and aligned stem possible bonding pairs (2). Metric 3 only considers the hairpin system to be closed when all of the possible pairs in the aligned stems are bonded. The $A_5G_5T_5$ simulation data are presented for each of the three metrics with non dimensional temperature and normalized intensity. A sigmoidal curve is fit to the data for each metric.

question are Watson–Crick complementary) and, if this distance is less than σ, they are counted as bonded. This calculation continues for the $i = 2$ and $j = n - 1$ beads and so on up the stem of the hairpin. This metric thus eliminates the problem of double counting in the first metric.

However, this second metric may still not fully capture the experimental system due to the nature of an intercalating dye. If double-stranded DNA is modeled as a ladder, then intercalating dyes like SYBR Green I fit between the rungs. Therefore, in order to bind to the DNA, at least two bases need to be bonded. Other simulation studies have used metrics that rely on contiguous bonded bases [168, 169]. As a result, we developed a third metric to capture this feature of the experiments. Metric 3, as depicted in Figure 4.6, requires that all of the complementary aligned bases be bonded in order for the hairpin to be deemed closed; this final metric is effectively bimodal.

We computed the average metric value for each simulation temperature run by computing the time average of the number of bonds found at every saved simulation frame (1000 BD time steps ≈ 1 μs). Each of the three simulation run replicates were then averaged together to find a single metric measurement for each non dimensional temperature. The MC algorithm transformed the data to produce a sigmoidal-shaped melting curve (with high temperature corresponding to high intensity) and normalized to a [0,1] scale with 0 corresponding to a fully closed state and 1 corresponding to a fully open hairpin. Normalization of each of the metrics, as in Figure 4.6, to the same [0,1] scale not only allowed for the three metrics to be compared among one another, but also with the experimental data.

Figure 4.6 also shows that each bonding metric defines a slightly different sigmoidal-shaped melting curve. The effective melting point temperature, T_{MP}, is defined where the chain has an equal probability of being open or closed. We obtain the melting temperature curve, T_M, by fitting the melting curve data with a sigmoidal distribution and finding the inflection point.

When we consider the sharpness of their transition regimes, we find that, by construction, metric 2 will form the sharpest melting curve, even though metric 3 is an "on-off" metric. (Metric 1, due to its multiple base bonding allowances, will have the broadest transition regime.) Imagine the base case sequence of $A_5C_5T_5$, which is comprised of stem length of five base pairs. In the completely closed state, all five pairs are bonded (to the aligned corresponding base, i.e., A_1 to T_{15}) and the instan-

taneous value of both metrics 2 and 3 is 1. If one bond is lost along the chain, the instantaneous value of metric 2 will be equal to 0.8 while the instantaneous value of metric 3 will be 0. Assuming, for the sake of argument, that the pair is bonded for half of the simulation time, the average value of metric 2 will be 0.9 and the average value of metric 3 will be 0.5. Therefore, with the bimodal metric 3, the transition regime will similarly become broader near the melting point temperatures.

4.3.6 DNA Hairpin Melting Simulation Optimization

In order to compare the experimental and simulation data, we first needed to determine both the conversion between simulation and experimental temperatures, T_{scale}, and the best ratio of stacking to bonding energy, ϵ_S/ϵ_{HB}. These two mutually dependent simulation parameters determine how well the simulation data matches the experimental data. We used the $A_5C_5T_5$ sequence to determine the best choices for these parameters because it is computationally efficient and the synthesis yield is high.

First, the three different metrics, as defined in Section 4.3.5, provide a spectrum of descriptions for the closed hairpin state. The simulation data for the test sequence $A_5C_5T_5$ were processed by metrics 1, 2, and 3 for each of the ϵ_S/ϵ_{HB} values. This produces a series of curves for the $A_5C_5T_5$ sequence in non dimensional temperature space, similar to that of Figure 4.6. Recall that we kept the ratio of the stacking and hydrogen bonding energies constant at the thermodynamic value [12, 13] of 2.5, but allowed the magnitude of these quantities to change. Due to the large number of simulation runs at every temperature that are needed to create a hairpin melting curve, only three values were investigated: ϵ_S/ϵ_{HB} = 2.5/1, 5/2, and 10/4. This allowed the base specific potentials to vary with respect to the other bead potentials (such as spring stiffness and bending potentials) in the system.

Next, the simulated melting curves for the different metrics were shifted by some T_{scale} = dT_{exp}/dT_{sim} which converts between the non dimensional simulation temperature and the dimensional experimental temperatures schemes,

$$\frac{dI_{sim}}{dT_{sim}} = \left(\frac{dI_{exp}}{dT_{exp}}\right)\left(\frac{dT_{exp}}{dT_{sim}}\right) = \left(\frac{dI_{exp}}{dT_{exp}}\right) T_{scale}. \tag{4.18}$$

The T_{scale} value for each metric melting curve was chosen so that the simulated melting-point temperature for this metric, T_{MP}, matches to the experimental melting-

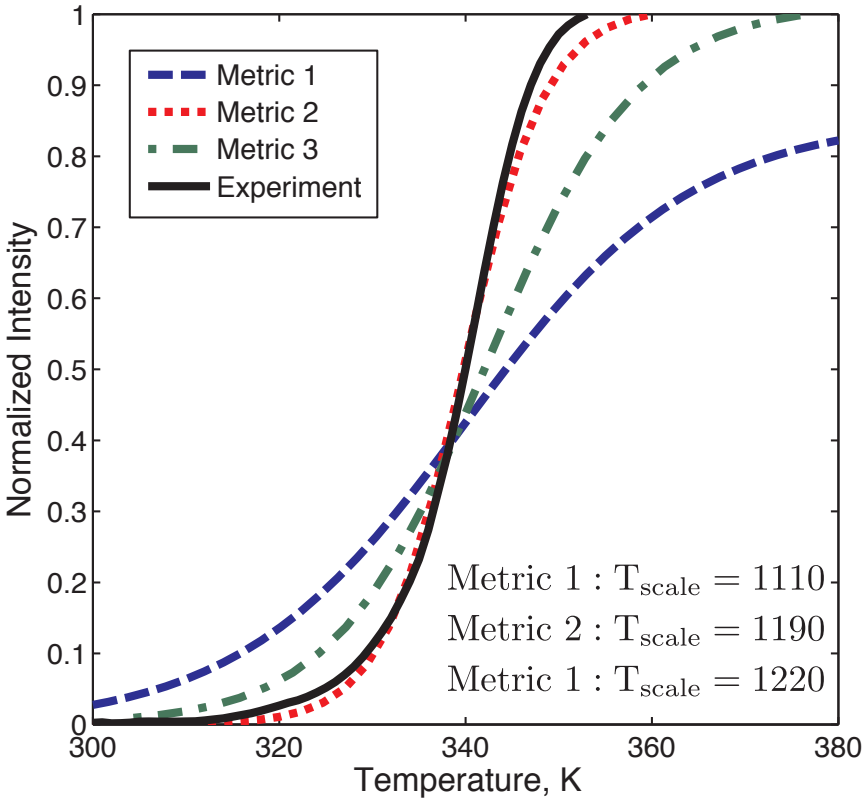

Figure 4.7: Plot of the three metrics with the experimental data overlaid for the $A_5C_5T_5$ base case sequence. Metric 2 best captures the slope in the transition regime of the experimental data.

point temperature $T_{MP}^{exp} = T_{MP}^{sim}(\text{Metric 1}) = T_{MP}^{sim}(\text{Metric 2}) = T_{MP}^{sim}(\text{Metric 3})$. As an example, we present in Figure 4.7 the $A_5C_5T_5$ sequence using each of the three metrics for the $\epsilon_S/\epsilon_{HB} = 5/2$.

To quantify the fit of the three melting metrics depicted in Figure 4.7 (with the bonding potential of $\epsilon_S/\epsilon_{HB} = 5/2$), we calculated the coefficient of multiple determination adjusted for the number of parameters in the sigmoidal model, R_a^2, values. We repeated the above process with the $\epsilon_S/\epsilon_{HB} = 2.5/1$ and $10/4$ bonding potentials. It should be noted that for $\epsilon_S/\epsilon_{HB} = 2.5/1$, some error in very low temperature data is expected. In this regime, extremely long simulation equilibration times are needed due to the low thermal energy of the system. The resulting T_{scale} and R_a^2 values for each bonding potential and metric are summarized in Table 4.5. From these data, we concluded that the $\epsilon_S/\epsilon_{HB} = 5/2$ and Metric 2 are the optimal bonding potential and

	Metric 1	Metric 2	Metric 3
$\frac{\epsilon_S}{\epsilon_{HB}} = \frac{2.5}{1}$	$T_{scale,K} = 1390$ $R_a^2 = 0.401$	$T_{scale,K} = 1460$ $R_a^2 = 0.678$	$T_{scale,K} = 1490$ $R_a^2 = 0.636$
$\frac{\epsilon_S}{\epsilon_{HB}} = \frac{5}{2}$	$T_{scale,K} = 1110$ $R_a^2 = 0.412$	$T_{scale,K} = \mathbf{1190}$ $R_a^2 = \mathbf{0.995}$	$T_{scale,K} = 1220$ $R_a^2 = 0.767$
$\frac{\epsilon_S}{\epsilon_{HB}} = \frac{10}{4}$	$T_{scale,K} = 1020$ $R_a^2 = 0.448$	$T_{scale,K} = 1030$ $R_a^2 = 0.716$	$T_{scale,K} = 1040$ $R_a^2 = 0.597$

Table 4.5: Summary of the T_{scale} and R_a^2 values for each of the ϵ_S/ϵ_{HB} values and metrics examined in the study. Row 2 is depicted graphically in Figure 4.7 and the center, bold cell contains the optimal values utilized for the remainder of the investigation.

metric, respectively. This choice produces a $T_{scale} = 1190$ K and $R_a^2 = 0.995$ for the base case sequence. These parameters and metric definition will be used to evaluate the model.

4.3.7 Simulation and Experimental Comparisons

With $T_{scale} = 1190$ K a simulation temperature of 0.3 corresponds to an experimental temperature of approximately 340 K. Although the thermodynamic estimate for the conversion factor in Section 4.3.4 is different, the value of T_{scale} found with the present method corresponds to our particular experimental system. Indeed, this experimental approach allows the model to be adjusted for any biological condition.

Let us now see if the value $T_{scale} = 1190$ K leads to similar agreement (using metric 2) for the other sequences. As seen in Figure 4.8, there is qualitative agreement. Unfortunately, the R_a^2 values reported in Table 4.6 indicate a problem with the model. With sequences similar to the base case of $A_5C_5T_5$, such as the $A_5C_{10}T_5$ sequence, we see high correspondence between the simulated curves and the experimental data. We find somewhat reduced matching with sequences containing poly-guanine sequences such as $A_5G_5T_5$, through 94% of the simulation data for these sequences lie within one temperature standard deviation of the median experimental data value. It is important to note that the "shoulder" regions of the melting curve, which is the initial transition from fully closed to 10% open and from fully open to 10% closed, show the greatest degree of mismatch to the experimental data. Finally, as the sequence stem gets longer, for example in the $A_7C_5T_7$ sequence, the agreement between the simulation and experimental data is again reduced. If we take the temperature range of the experimental data, depicted in Figure 4.3(d) by the horizontal bars, we find that 91% of the averaged metric 2 simulation data fall within the measured spread.

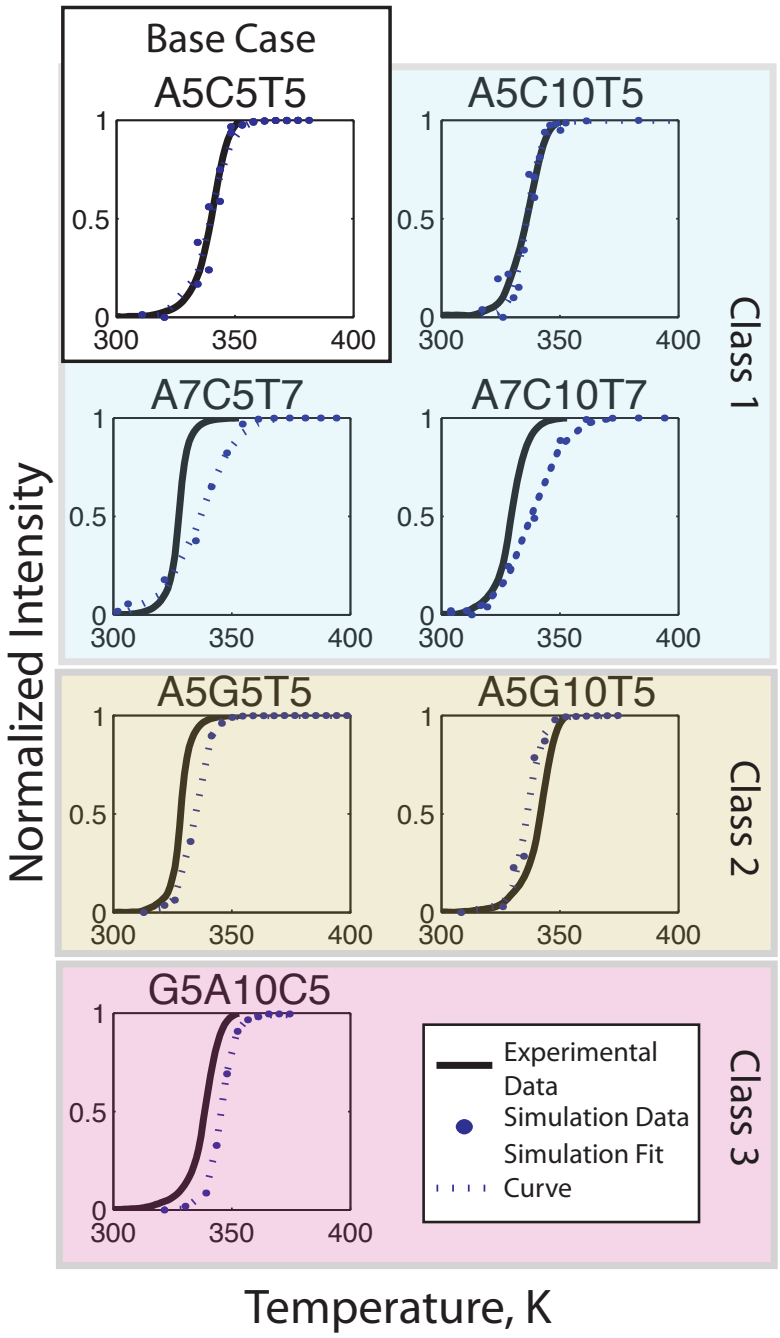

Figure 4.8: Comparison of the metric 2 simulation data and experimental sigmoidal fit curve with a T_{scale} = 1190 K factor. Each of the seven investigated sequences is depicted and sorted into their original classes. The R_a^2 values of the fits are reported in Table 4.6.

Perhaps the low R_a^2 values in Table 4.6 indicate that we chose the wrong base case. To test this possibility, we determined the optimal T_{scale} values for each sequence

Sequence	R_a^2
$A_5C_5T_5$	0.995
$A_5C_{10}T_5$	0.942
$A_7C_5T_7$	0.450
$A_7C_{10}T_7$	0.540
$A_5G_5T_5$	0.622
$A_5G_{10}T_5$	0.713
$G_5A_{10}C_5$	0.514

Table 4.6: Summary of the R_a^2 values for T_{scale}=1190 K for each of the sequences examined.

independently. However, this does not lead to all sequences having a value $R_a^2 > 0.9$, which we would consider to be a good fit. Indeed, simply changing the choice of T_{scale} to get the right melting point temperature, which is the usual approach in matching simulation to experiment, does not impy that the simulation will then mimic the full behavior of T_M. This is exemplified by the data for the $A_7C_5T_7$ sequence in Figure 4.9. Longer stem lengths are characteristic of this (and the other poor performing class I molecule $A_7C_{10}T_7$) and may explain the broadening of the transition regime.

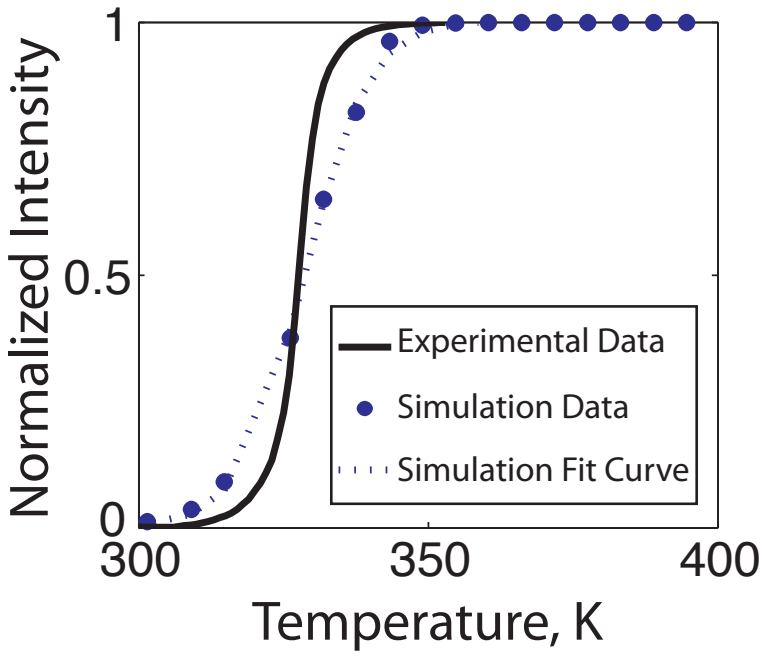

Figure 4.9: Comparison of the metric 2 simulation data and experimental sigmoidal fit curve for the $A_7C_5T_7$ sequence with T_{scale}=1150 K which is the conversion required to match the melting point temperature. The R_a^2 value is 0.707.

Qualitatively poor fits point toward parameters and features that may need to be adjusted or added to the two bead DNA model. Although the DNA model includes potentials describing both hydrogen bonding and nucleotide base stacking, it does not include the effects of cross stacking interactions between bases [12, 13, 15, 172–174]. Due to the fact that we see diminished R_a^2 values for sequences with longer stems, and thus more cross stacking interactions, we may need to add this feature to the model. From additional examination of Figure 4.8, we see that sequences with polyguanines also have reduced fits. While the two bead model includes pairwise interactions, it does not contain more complex (i.e., four base coordination) interactions that would be needed to describe the features of G-quartets [12, 13, 42]. The mismatch in the simulation and experimental systems with the $A_5G_5T_5$, $G_5A_{10}C_5$, and $A_5G_{10}T_5$ may be due to this lacking G-quartet coordination. Further study with incorporation of both cross stacking and multi base bead interactions could be conducted to determine if the experimental data can be better captured.

4.4 Conclusions

Due to the constant trade-off between simulation time and length scales, only the most basic of DNA behaviors (and thus interactions) are included in this two bead DNA model. After testing this model against experimental data, we found that it was unable to adequately capture even the most elementary of experimental systems. Therefore, before more complex structures, such as those described in Chapter 2, are attempted, the model must be revised. In the following chapter, we build a more realistic, yet still simplistic, model of DNA with three beads per nucleotide. This new architecture allows for additional sub features to be realized such as chirality, major and minor grooves, and backbone angles. Additional features to give a more refined base identity behavior including more realistically parameterized hydrogen bonding and stacking, the inclusion of non-Watson–Crick hydrogen bonding, and anisotropic potentials to eliminate unrealistic bonding. With these additional characteristics we seek to be able to capture not only the simplistic hairpin melting curves but also a wide array of other small DNA conformations.

5

Development of a New, Non-Canonical DNA Model

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of single datum of experience.

Albert Einstein, *On the Method of Theoretical Physics*, June 10, 1933 [239]

As we move from the previous two bead DNA model and look to incorporate a new host of features to better capture the available experimental data, it is important to remember Einstein's basic idea that everything should be as simple as it can be, but not simpler. As was described in Chapter 3, this is the ultimate goal in developing and utilizing models and their corresponding theories; to find the "irreducible basic elements" and have them be "as simple and as few as possible" to adequately represent the system at the length and time scales of interest [239]. In order to be able to describe both the canonical and non-canonical nucleic acid structures described for single-, double-, triple-, and quadruple-stranded structures in Sections 2.3, 2.2, 2.4, and 2.5, respectively, we must incorporate additional features into our previous DNA model described in Chapter 4.

The nucleic acid structures previously described in Chapters 1 and 2, possess several levels of complexity, ranging from the sequence of the bases (primary structure), to the base-pairing (secondary structure), and to its ultimate three-dimensional shape (tertiary structure). In designing a new DNA model we must be able to capture key features at each of these gradations. We will return to the idea of these three

complexity classifications (primary, secondary and tertiary structure) when we design metrics for evaluating our improved nucleic acid representation. Since the time-scales characterizing the structural dynamics (assembly and action) of the nucleic acid structures of Chapters 1 and 2 exceed the current capabilities of atomistic simulations, their study requires the usage of a coarse-grained model.

In this chapter we will amend, correct, and improve the previous two bead model of DNA with the basic features allowing us to capture key structural features described heretofore. Some of the necessary improvements include: chemical asymmetry of the backbone (i.e., a 5'-3' directionality); proper sizing of the sugar, phosphate, and nitrogenous base groupings; directional canonical and non-canonical base pairing, directional stacking interactions, and more realistic and experimentally derived parameters. These modifications are necessary not only to better capture relatively simple behavior such as that of single-stranded DNA hairpins but also move to more complicated and diverse formations of triple- and quadruple-stranded structures.

5.1 Three Bead Nucleic Acid Model

For the basic structural unit of the new DNA model, we adopted the model depicted schematically in Figure 5.1. Instead of the previous depiction of a nucleotide with two beads, this model utilizes a three bead approach. Whereas the sugar and phosphate groups were clustered into a single bead before, the backbone bead of Section 4.2, in our improved model the backbone groups are each allowed their own bead representations. In this structure we represent each nucleotide with a bead for the sugar (S), phosphate (P), and base (B) group, as illustrated in Figure 3.2. The addition of a third bead per nucleotide allows the model to have handedness when the DNA is in helical form since the repeating nucleotide unit is now asymmetrical in nature [160]. The termini of any nucleic acid can now be distinguished into their 3' and 5' ends; the phosphate group (which is only bound to a single sugar group) denotes the 5' end of each nucleic acid strand.

From a structural standpoint, our model resembles the 3-sites-per-nucleotide (3SPN) model in only the most literal sense; both the 3SPN model from de Pablo and coworkers [160, 161, 167] and the model we propose here represent each nucleotide as 3 beads. As will be apparent shortly, the force fields for the two models are quite different. The most notable difference is our use of non-spherical bonding potentials, which allow

us to avoid the need for dihedral potentials [160, 161, 164, 167–169] that introduce an unphysical bias towards the B-form of dsDNA when the DNA is single-stranded [160]. Rather, we can smoothly move between dsDNA and ssDNA. As a result, it would be inappropriate to view the new model as an extension of the existing 3SPN models [160, 161, 167].

The model is written in terms of a dimensionless length σ and energy ϵ. The value of σ is fixed by the sugar-phosphate distance in single-stranded DNA. In Figure 5.1, the backbone is depicted in a plane and the sugar-phosphate distance, $\sigma = 0.3$ nm, is shown. The degree of freedom embodied in ϵ is used to map the simulation temperature to the experimental one, as was conducted in Section 4.3.7 and will be described further in Section 5.2.2. Each of the potentials depicted in Figure 5.1 and governing the interactions of the model can be sorted into three categories: nonspecific interactions occurring between all beads, backbone-backbone interactions, and base-base interactions.

5.1.1 Nonspecific Interactions

In general the spacing between the sugar, phosphate group, and bases, along with their relative sizes, are enforced by a combination of excluded volume interactions and modified harmonic springs, defined by Equations 5.1 and 5.2, respectively; the sum of these potentials creates a relatively deep well that minimizes the fluctuations in these distances [172]. In the model every bead interacts with the other beads by excluded volume interactions in order to provide each bead with a physical size. For each bead i, the interaction with bead j is given by the truncated pairwise Lennard-Jones potential,

$$U_{\mathrm{EV}}(r_{ij}) = 4\epsilon \left[\left(\frac{\gamma}{r_{ij}} \right)^{12} - \left(\frac{\gamma}{r_{ij}} \right)^{6} \right] + \varepsilon, \tag{5.1}$$

for $r_{ij} \leq 2.5\gamma$. The energy $U_{\mathrm{EV}} = 0$, otherwise. To improve on the two bead DNA model where all of the beads were the same size, the backbone sugar and phosphate groups are given one size while the base groups are effectively made larger by using different parameters in Equation 5.1. The backbone beads have size σ and the base beads have size 1.5σ. The parameter, γ, in Equation 5.1 is calculated with the arithmetic average of the size of the i and j beads. However, it is important to note that we do not distinguish between the different sizes of the purine and pyrimidine bases, shown in Figure 1.1, instead all bases are considered to have the same overall

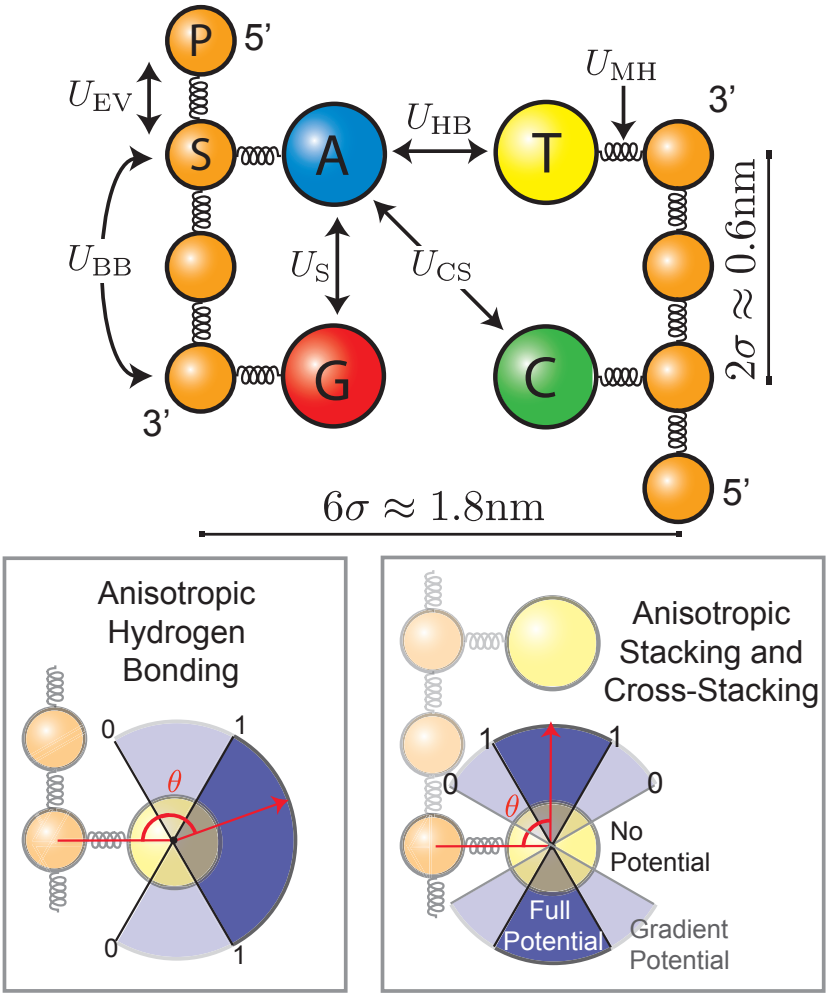

Figure 5.1: A schematic representation of a three bead coarse-grained model of DNA. Each nucleotide is represented by three beads, one for the phosphate group (P), one for the sugar group (S), and one for the base (A, C, G, T). The various interactions are labeled between the beads, where the hydrogen bonding interactions include Watson–Crick, Hoogsteen, and wobble type bonds. Directionality of the DNA chain is labeled, with the phosphate group marked as the 5'-end of the chain. The size spacing between each nucleotide group is shown to measure $2\sigma = 0.6$ nm and is one of the two free parameters used to dimensionalize the code. The base beads are 1.5 times the size of the backbone beads.

spherical size [160].

All bonded beads also interact through the modified harmonic (FENE) potential

$$U_{\mathrm{MH}}(r_{ij}) = -15\epsilon \left(\frac{R_0}{\sigma}\right)^2 \ln\left[1 - \left(\frac{r_{ij}}{R_0}\right)^2\right]. \tag{5.2}$$

For the backbone-backbone springs, the finite extensible length is 1.5σ, while for the backbone-base springs it is 2.25σ and the parameter R_0, in Equation 5.2, is calculated with the arithmetic average of the finite extensible length for the i and j beads. The differently sized springs connecting the (i) sugar-phosphate groups and the (ii) sugar-base groups, allow for the size of the nucleotide to be tuned to realistic dimensions. Although there is still no base identity derived size differentials (such as would distinguish purines from pyrimidines), these parameters give a generic nucleotide the approximate dimensions of 2σ by 3σ [13, 15]. The combination of excluded volume and spring forces maintains a relatively constant extension between bonded bead pairs.

5.1.2 Backbone-Backbone Interactions

The backbone stiffness is enforced with a bending potential,

$$U_{\mathrm{BB}}(\phi) = 12\epsilon(1 + \cos\phi)^2, \tag{5.3}$$

between all sugar trios along the same backbone. The stiff backbone bending potential has an equilibrium angle of π, [155, 168, 169] leading to a ssDNA persistence length (calculated from the decay of the autocorrelation function along the vector between consecutive sugar beads) of 1.7 nm. We make this calculation using the sugar beads, rather than including the phosphate beads as well, since the bending energy is defined between sugar trios. Our persistence length thus corresponds to nearly five nucleotides when we account for the natural relaxation of single-stranded DNA. This quantity is in line with experimental values of the flexibility of ssDNA and RNA measured with a variety of experimental approaches. As was detailed in Section 2.3, the values of the persistence length of ssDNA range from 0.75 nm to 5.2 nm [73–79], and seems to vary widely due to a variety of factors including the length of the sequences examined, the model used to examine the data, and the concentration and type of buffer used in each experiment. Our measured value of 1.7 nm is well in line with the experimental studies available.

5.1.3 Base-Base Interactions

The sequence dependent structure is captured by base specific potentials. There are three different types of hydrogen bonding interactions, Watson–Crick, Hoogsteen, and wobble bonds. Our nomenclature does not distinguish between protonated (reversed-Hoogsteen or wobble pairs) and non-protonated (traditional) Hoogsteen or wobble bonds; we simply use the term Hoogsteen and/or wobble bonds in all cases.

Taken in isolation, Hoogsteen bonds are rather strong; the interaction energy of a Hoogsteen A · T bond (5.2 kcal/mol) compares favorably with the equivalent Watson–Crick bond (5.7 kcal/mol) [12, 13, 240]. The reason why base pairing in double-stranded DNA is dominated by Watson–Crick bonds is the resultant stabilization of excluded volume interactions. Thermal transitions in dsDNA structure, such as melting, hybridization, and bubble formation, will also be affected if the sequence permits Hoogsteen-bonded secondary structure as it transitions to the open state. To capture the complexity of DNA structural dynamics in a coarse-grained model, it is essential to move beyond Watson–Crick base pairs.

Stacking interactions in the model occur between bases on the same backbone and enforce a sequence dependence for the interaction in the 5'-3' direction. Cross-stacking interactions occur between bases on a nearby strand when there is Watson–Crick hydrogen bonding between one of the bases involved in the cross-stacking. To enforce the directionality of the hydrogen bonding and stacking/cross-stacking interactions, we have modified their spherical potential functions with a smoothly varying prefactor, $f_{\rm k}(\theta)$ [164, 168, 169, 177]. Since the bonds have directionality, all of the base-base interactions (including excluded volume) are enforced at all times in the simulation, rather than alternating between the most relevant ones [160].

The sequence dependent interactions have the generic form

$$U_{\rm k}(r_{ij},\theta) = -\delta_{\rm k}^{ij} f_{\rm k}(\theta)\epsilon \left[\exp\left(20\frac{r_{ij}}{\sigma} - 30\right) + 1\right]^{-1}, \tag{5.4}$$

for $r_{ij} \leq 10\sigma$. The energy $U_{\rm k} = 0$, otherwise. This particular form of the potential has been used elsewhere [172, 184] to model hydrogen bonding and stacking in DNA. The parameter $\delta_{\rm k}^{ij}$ describes the strength of a bond of type k between base i and base j. The function $f_{\rm k}(\theta)$ appearing in Equation 5.4 accounts for the directionality of the hydrogen bonding interactions (Watson–Crick, Hoogsteen, or wobble) and the stacking interactions similar to the bead-pin model [170, 241]. Figure 5.2 illustrates

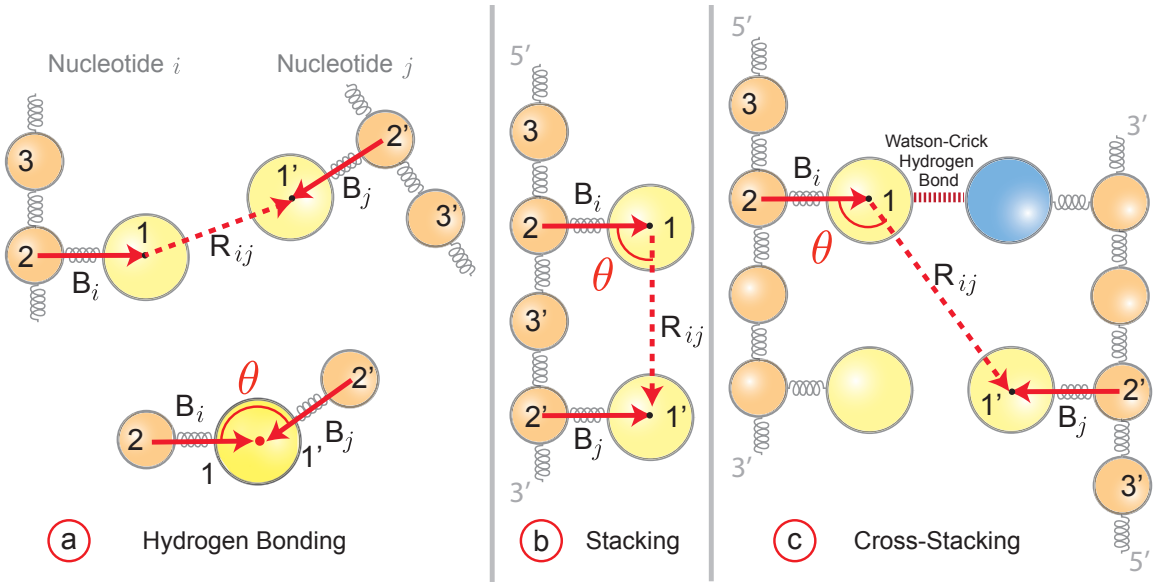

Figure 5.2: The definitions of the angle θ for (a) the hydrogen bonding, (b) stacking, and (c) cross-stacking interactions between nucleotide i and j. For each nucleotide, the base bead (1) the sugar bead (2), and the phosphate bead (3) are numbered. The vectors $\mathbf{B}_i$ and $\mathbf{B}_j$ (solid red lines) are drawn from the sugar bead to the base bead. The $\mathbf{R}_{ij}$ vector (dashed red line) is drawn between the interacting bases. The top part of (a) shows the nucleotide positions; the bottom part of (a) shows the definition of the angle θ. Both stacking and cross-stacking follow the 5'-3' direction along the backbone. For stacking interactions (b), the nucleotides i and j are contiguous along the backbone. In cross-stacking interactions (c), nucleotide i is involved in Watson–Crick hydrogen bonding and nucleotide j is displaced in the 5' direction from the other Watson–Crick bonded bead. The particular case illustrated in (c) corresponds to the $\uparrow^{5'\text{T}\cdot\text{A}\ 3'}_{3'\text{T}\cdot\text{T}\ 5'}\downarrow$ = 8.8 entry from Table **??**.

the definition of the angle θ for hydrogen bonding, stacking, and cross-stacking in terms of the vectors drawn from the backbone to the base, $\mathbf{B}_i$ and $\mathbf{B}_j$, and the vector drawn between the bases, $\mathbf{R}_{ij}$. In Equation 5.4, $r_{ij} = |\mathbf{R}_{ij}|$.

The hydrogen bonding of some base bead i is computed with all other base beads $j \neq i$, which allows us to move smoothly between secondary structure in ssDNA and dsDNA. The value of $\theta \in [0, \pi]$ depicted in Figure 5.2 is computed from the normalized dot product,

$$\cos(\theta) = \frac{\mathbf{B}_i \cdot \mathbf{B}_j}{|\mathbf{B}_i||\mathbf{B}_j|} \tag{5.5}$$

and the corresponding modulating function illustrated in Figure 5.1 is

$$f_{\text{HB}}(\theta) = \begin{cases} 0 & \text{for} \quad \theta \in [0, \pi/3] \\ |\cos(3\theta/2)| & \text{for} \quad \theta \in [\pi/3, 2\pi/3] \\ 1 & \text{for} \quad \theta \in [2\pi/3, \pi] \end{cases} \tag{5.6}$$

The full bonding strength is present over an angle of 120°, in light of the mirror symmetry in Figure 5.1, which is an approximation of the spread of the hydrogen donor and acceptor sites centered on the coarse-grained bead, as seen in Figure 1.5. The function evolves smoothly between the on/off states, which is important for some tertiary structure formations, and the particular functional form is convenient for computation [168, 169]. The off state prevents unphysical bonding through the backbone without the need for a cutoff length or additional beads [176, 177].

The key differences between stacking/cross-stacking and hydrogen bonding are the limitations on the value of j and the preferred alignment of base i and base j. For stacking, the bead j is displaced from bead i in the 3' direction on the chain. For cross-stacking, the i bead is involved in the Watson–Crick bond and the j bead is displaced in the 5' direction from the other Watson–Crick bonded bead. The definitions of θ, illustrated in the Figure 5.2, are computed by

$$\cos(\theta) = \frac{\mathbf{B}_i \cdot \mathbf{R}_{ij}}{|\mathbf{B}_i||\mathbf{R}_{ij}|} \tag{5.7}$$

and the corresponding modulating function, sketched in Figure 5.1, is

$$f_{\mathrm{S}}(\theta) = f_{\mathrm{CS}}(\theta) = \begin{cases} 0 & \text{for} \quad \theta \in [0, \pi/6] \\ |\cos(3\theta)| & \text{for} \quad \theta \in [\pi/6, \pi/3] \\ 1 & \text{for} \quad \theta \in [\pi/6, \pi/2] \end{cases} \tag{5.8}$$

with a mirror symmetry for $\theta \in [\pi/2, \pi]$. The full stacking and cross-stacking interactions are present over an angle of 60° centered on a perpendicular vector to the sugar-base vector, as depicted in Figure 5.1. There is no stacking or cross-stacking interactions over the angle of 60° parallel to the sugar-base vector. The function evolves smoothly between these on/off states in a computationally convenient manner. The directionality of the stacking and cross-stacking interactions mimic the parallel nature of the planar base ring formations that were described in Section 1.3.2.

Parameterization of Base-Base Interactions

The former two bead model utilized a general approximation of hydrogen bonding strength for the included Watson–Crick type bonding. As was explained in Section 1.3.1 and shown in Figure 1.6, a A · T Watson–Crick base pair has two hydrogen bonds while a G · C Watson–Crick base pair has three. In Section 4.2.3 the parameterization

for the two bead model uses the number of hydrogen bonds with the strength of a $A \cdot T$ base pair being two thirds that of a $G \cdot C$ base pair. However, this approach to hydrogen bonding was naive in two respects: (i) non-canonical hydrogen bonding are prevalent in nucleic acids, and (ii) hydrogen bonding strength is not additive.

First, the inclusion of Hoogsteen and wobble hydrogen bonds along with the canonical Watson–Crick hydrogen bonding patterns are necessary to fully capture many nucleic acid behaviors and structures. As was explained in Chapters 1 and C2, there is ample evidence that short ssDNA and RNA have the ability to fold into a multitude of structures that utilize these additional bonding patterns. For example, the $A \cdot G$ base pair is often found in tRNA, rRNA, ribozymes, and other oligonucleotides [12, 19, 242–251], the $A \cdot U$ bond in the Hoogsteen configuration is found in rRNA [19], the $G \cdot U$ and $G \cdot T$ wobbles are found in tRNA and other DNA structures [22, 252–254], and the $G \cdot G$ is found in telomeres, aptamers, DNAzymes, RNAzymes, and even in chromosomal DNA [13, 255–258]. As can be seen from the list above and from Chapter 2, many of the structures at the length scale of interest for this model include non-Watson–Crick hydrogen bonds. Their inclusion is paramount in order to be able to capture the behavior of nucleic acids at this scale. Secondly, hydrogen bonding strength depends on the location and the identity of the atoms involved in each bond. In order to improve the new three bead model, we need a more realistic set of parameters to describe different base-base hydrogen bonding interactions.

There are a set of general rules we can consider for the relative strength of hydrogen bonding interactions. From experimental nucleoside association studies [12, 259] it is roughly known that:

$$\begin{gathered} \text{G with } C > G \gg T > A, \\ \text{C with } G \gg C > T > A, \text{ and} \\ \text{A with } T > G \sim C. \end{gathered}$$

However, it should be noted that these rankings can only be considered rough estimates for nitrogenous bases in a nucleic acid polymer because nucleosides are free molecules and lack a phosphate group in their structure. These additions can alter their behavior somewhat, though the general trends should be much the same.

Similarly, the previous DNA model utilized a generalized parameter set for the stacking interactions between contiguous nucleotides. However, this technique to stacking interactions was too simple in two respects: (i) 3'-5' directionality of bases is important in stacking interactions, and (ii) stacking strengths do not vary in discrete

values. As was greatly detailed in Section 1.3.2, there are many aspects that affect the base stacking strength of two nucleotides; but a generalized ranking for the stacking interactions [15] can be considered to be

$$\text{purine - purine} > \text{pyrimidine - purine} > \text{pyrimidine - pyrimidine.}$$

Consequently, we can consider these qualitative guidelines for base pairing and stacking interactions along with more quantitative experimental values and quantum chemical calculations when designing the parameters to describe the base-base specific interactions.

The parameters δ_{k}^{ij} for hydrogen bonding and stacking are estimated from a range of experimental and computational data in the literature [12, 13, 28–30, 118, 240, 243, 247, 260–270]. Since we eventually chose ϵ to match the experimental data for hairpin melting curves discussed in Section 4.3.2 [174], we only need to determine the relative strengths of each interaction. We first grouped all of the references by the type of measurement, often with several publications per group.

All of the groups reporting hydrogen bonding energies included an estimate for the energy of a C-G Watson–Crick hydrogen bond, $\tilde{U}_{\mathrm{HB}}^{CG}$, along with values for other hydrogen bonds and/or stacking. We then rescaled all of the data within a given group relative to their reported value for $\tilde{U}_{\mathrm{HB}}^{CG}$. One group reported data for $\tilde{U}_{\mathrm{HB}}^{CG}$, $\tilde{U}_{\mathrm{S}}^{CG}$ and $\tilde{U}_{\mathrm{S}}^{GC}$ [265]. There is still a degree of freedom in this parameterization set which we used to conduct a series of simulations to adjust relative strengths so that the average U_{S} to U_{HB} ratio is approximately 2.5 as reported in literature [13]. The rescaled base specific parameters, δ_{k}^{ij}, appear in Equation 5.4. Note that, although the hydrogen bonding energies are symmetric with respect to ij, the stacking energies depend on the 5'-3' direction. From one set of experimental data [265], the relative values of δ_{k}^{ij} between hydrogen bonding and stacking is $\delta_{\mathrm{S}}^{CG} = 16.79\delta_{\mathrm{HB}}^{CG}$. We used the latter relationship to rescale the stacking data in the other references. In summary, so long as a group measured either a G-C Watson–Crick hydrogen bond, $\delta_{\mathrm{HB}}^{CG}$, C-G stacking, δ_{S}^{CG}, or G-C stacking, and δ_{S}^{GC} , we can use one of the latter trio to rescale the other hydrogen bonding or stacking data to a relative strength.

To merge these rescaled values into a single set of parameters for δ_{k}^{ij} for stacking and hydrogen bonding, we used quantum chemical calculations [12] as the guide. The latter calculations provide a rank-order for the strengths of different base-base interactions, and we ensured that our final set of parameters preserves this rank order. There are three possible cases we had to consider to determine the value for a given

Table 5.1: The base specific hydrogen bonding parameters, δ_{k}^{ij} for each base pair combination presented in Section 1.3.

	A	C	G	T
A	3.20	3.64	5.36	4.00
C	3.64	6.12	9.56	2.20
G	5.36	9.56	9.16	4.44
T	4.00	2.20	4.44	2.12

Table 5.2: The base specific stacking parameters, δ_{k}^{ij} for each stacking combination possible. The 5'-(top) base is listed in the left column and the 3'-(bottom) base pair is listed along the top of the table. The table is not diagonally symmetric, the 3'-5' direction of the bases affects the relative strength of stacking interaction.

	$\uparrow_{3'}$ A	$\uparrow_{3'}$ C	$\uparrow_{3'}$ G	$\uparrow_{3'}$ T
$^{5'}\uparrow$ A	59.07	72.27	107.91	42.02
$^{5'}\uparrow$ C	115.61	90.86	160.49	107.91
$^{5'}\uparrow$ G	74.58	106.59	90.86	72.27
$^{5'}\uparrow$ T	72.27	74.58	115.61	59.07

δ_{k}^{ij}: (i) If we had multiple values for a single δ_{k}^{ij} and they were close, we used the average. By close, we mean that using the average does not affect the rank order. (ii) If we had multiple values for a single δ_{k}^{ij} and they were not close, we picked the one that preserves rank order. (iii) If no value for δ_{k}^{ij} preserves the rank order, it was excluded from the data set. We did not encounter case (iii).

The values of the δ_{k}^{ij} parameters for Watson–Crick and Hoogsteen hydrogen bonds [12, 13, 118, 240, 243, 247, 260, 261, 264–266] are listed in Table 5.1. For hydrogen bonding, we do not allow any bonds between base i and $i+2$ on the same strand to avoid the formation of one member loops. Note that such loops also incur strong bending and excluded volume penalties, so this restriction may be superfluous. Stacking interactions only occur between contiguous bases on the sequence. For stacking [12, 13, 28–30, 118, 243, 247, 262–270], the strength, listed in Table 5.2, depends on the identity of i and j and their order in the 5'-3' sequence on that strand. Note that the overall strength of a hydrogen bond interaction is less than half of the overall strength of stacking interactions, as constructed in the δ_{k}^{ij} parameters [13, 174]. Therefore, even though the non-canonical base pairs have significant hydrogen bonding strengths, they can be considerably destabilized by stacking and cross-stacking interactions.

In addition to hydrogen bonding and intra-chain stacking interactions Equation 5.4 also includes inter-chain stacking. The inter-chain or cross-stacking interactions were

not included in the previous model. Cross-stacking, described in Section 1.3.2, occurs between strands or between noncontiguous bases on the same strand (for example, in the stem of a ssDNA hairpin). These are weak and poorly understood interactions. For cross-stacking between strands (or in a hairpin), we need to consider both bases i and j, as well as their complementary partners. We use a Gō-like potential that turns on the cross-stacking interaction if base i or j is Watson–Crick hydrogen bonded to the complementary strand (or hairpin stem). If one complimentary pair of bases forms a Watson–Crick hydrogen bond, as in Figure **??**, then the two cross-stacking interactions for the dimer are included. The value of the cross-stacking energy is very low when it turns on.

Estimating the value for the cross-stacking energy is not straightforward. We were able to identify one report on cross-stacking energies [13], which included a rubric stating that cross-stacking should be between 10-15% of the stacking energy. We set the $\delta^{ij}_{\mathrm{CS}}$ for $\uparrow^{5'\mathrm{C\cdot G}\ 3'}_{3'\mathrm{G\cdot C}\ 5'}\downarrow$ to be 15% of the $\uparrow^{5'\mathrm{C}}_{3'\mathrm{G}}$ dimer and then rescaled the remaining cross-stacking $\delta^{ij}_{\mathrm{CS}}$ relative to the corresponding value for $\uparrow^{5'\mathrm{G\cdot C}\ 3'}_{3'\mathrm{C\cdot G}\ 5'}\downarrow$. Since we require at least one Watson–Crick interaction in each dimer pair, the possible list of cross-stacking interactions [12, 13, 15, 118, 247, 262] in Table 5.3 is much larger than the 16 possible Watson–Crick dimer pairs. If we could not find experimental cross-stacking data, we set the value to zero. Since the cross-stacking interactions are weak, we do not expect the absence of data for some potential pairs to be a major concern.

5.2 Validation of Three Bead DNA Model

Returning to the idea that a useful coarse-grained model of DNA would be able to capture features at each of the levels of molecular complexity, we tested the newly designed three bead DNA model. It should be able to capture primary structure behavior such as the relaxed backbone configuration, secondary structure behavior such as the melting curves for DNA hairpins that can form non-canonical bonds in the open state, and also tertiary structure properties of double-stranded helices.

5.2.1 Relaxed Single-Stranded DNA Backbone Structure

As can be seen in Figure 5.3, relaxed single-stranded DNA exhibits significant stacking; the bases offset into a single helix conformation in order to maximize the stacking

Table 5.3: The base specific cross-stacking parameters, δ_{k}^{ij}. Cross-stacking interactions are only considered if one of the dimer base pairs is a Watson-Crick base pair. The 5'-(top) base pair is listed in the left column and the 3'-(bottom) base pair is listed along the top of the table. The table is read so that $\uparrow^{5'\,\mathrm{A\cdot T}\,3'}_{3'\,\mathrm{C\cdot G}\,5'}\downarrow$ =15.9.

	↑ AA ↓	↑ AC ↓	↑ AG ↓	↑ AT ↓	↑ CA ↓	↑ CC ↓	↑ CG ↓	↑ CT ↓	↑ GA ↓	↑ GC ↓	↑ GG ↓	↑ GT ↓	↑ TA ↓	↑ TC ↓	↑ TG ↓	↑ TT ↓
↑ AA ↓	-	-	-	8.8	-	-	16.5	-	-	12.1	-	-	11.0	-	-	-
↑ AC ↓	-	-	-	11.0	-	-	16.5	-	-	16.5	-	-	8.8	-	-	-
↑ AG ↓	-	-	-	8.8	-	-	15.4	-	-	14.3	-	-	12.1	-	-	-
↑ AT ↓	11.0	7.7	12.1	11.0	8.8	6.6	15.9	6.6	12.1	14.3	13.2	-	6.4	5.5	-	5.5
↑ CA ↓	-	-	-	6.6	-	-	11.0	-	-	12.1	-	-	7.7	-	-	-
↑ CC ↓	-	-	-	7.7	-	-	12.1	-	-	7.7	-	-	6.6	-	-	-
↑ CG ↓	12.1	12.1	17.6	15.8	16.5	7.7	20.2	11.0	14.3	24.6	15.4	-	14.3	5.5	-	7.7
↑ CT ↓	-	-	-	7.7	-	-	8.8	-	-	5.5	-	-	5.5	-	-	-
↑ GA ↓	-	-	-	8.8	-	-	15.4	-	-	17.6	-	-	12.1	-	-	-
↑ GC ↓	16.5	11.0	15.4	14.1	16.5	12.1	23.9	15.4	15.4	20.2	17.6	-	15.9	8.8	-	13.2
↑ GG ↓	-	-	-	8.8	-	-	17.6	-	-	15.4	-	-	13.2	-	-	-
↑ GT ↓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
↑ TA ↓	8.8	6.6	8.8	9.7	11.0	7.7	14.1	8.8	8.8	15.8	8.8	-	11.0	7.7	-	8.8
↑ TC ↓	-	-	-	8.8	-	-	15.4	-	-	11.0	-	-	6.6	-	-	-
↑ TG ↓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
↑ TT ↓	-	-	-	8.8	-	-	13.2	-	-	7.7	-	-	5.5	-	-	-

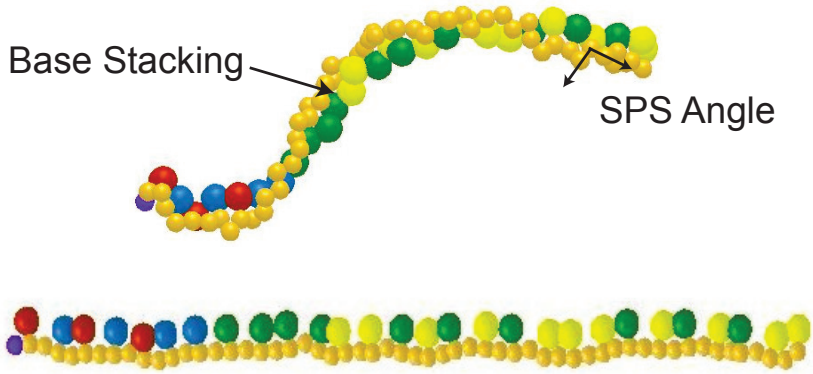

Figure 5.3: Snapshots from simulations of ssDNA shortly after its initialization as a comb (bottom) and in the relaxed state (top). An example of the sugar-phosphate-sugar (SPS) angle is depicted; this puckering shortens the apparently contour length of the relaxed molecule (top), measured from sugar to sugar, when compared to the completely extended state (bottom and Figure 5.1. Here the sugar and phosphate groups are shown as the small (orange) spheres with the bases A (blue), C (green), G (red), and T (yellow) as the larger beads. The dark (purple) bead on the end of each ssDNA backbone denotes the 5'-end of the molecule.

interactions and minimize the excluded volume and backbone bending interactions between consecutive beads. This helical formation due to stacking interactions was described for ssDNA in Section 1.3.2. Due to the level of coarse graining we cannot measure the glycosyl angle *per se*, however, we can measure the angle formed between contiguous sugar-phosphate-sugar (SPS) beads. The SPS angle is important not only in relaxed ssDNA structure, but also in dsDNA structures and will be discussed further in Section 5.2.3. In the absence of consistent base pairing (as in hairpins or dsDNA), the SPS angle changes dramatically in our model. For example, we obtained a SPS angle of $97^\circ \pm 52^\circ$ for the ssDNA sequence 5'-ATCATGCGATCATCCG-3' at a temperature of 340 K. The large deviation in the SPS angle, which results from temporal fluctuations, reflects the flexibility of ssDNA. Our calculated persistence length of ssDNA is 1.7 nm, as shown in Figure 5.4; as was detailed in Section 2.3, the values of the persistence length of ssDNA range from 0.75 nm to 5.2 nm [73–79], and seems to vary widely due to a variety of factors. The flexibility of a single nucleic acid chain allows our model to transition smoothly from ssDNA to dsDNA thereby permitting study of hybridization, melting, and other interchain interactions.

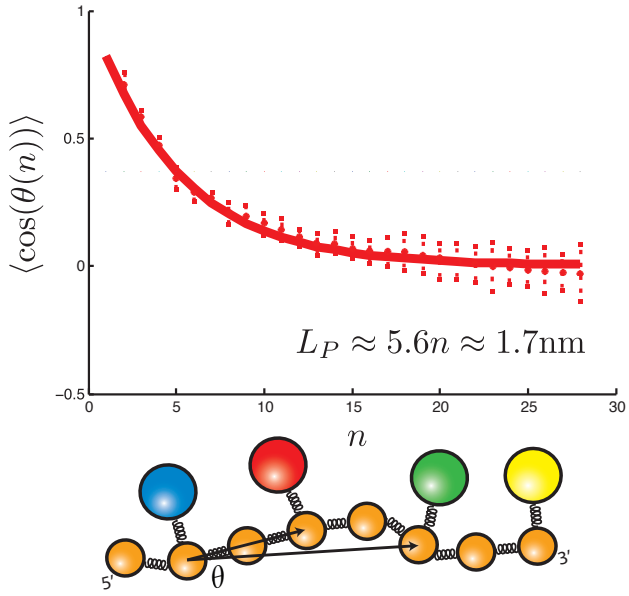

Figure 5.4: The persistence length is calculated for single-stranded DNA by finding the correlation of the angle, θ along the backbone of the nucleic acid, as depicted in the lower schematic. On average a persistence length of approximately 1.7 nm is found.

5.2.2 Melting of a DNA Hairpin

The second critical test for our model is its ability to capture thermally induced transitions. Some coarse-grained models are parameterized from the bottom up [157, 164, 182, 183, 202, 203], where the functional form and strength of the potentials are tuned to match trajectories from all atom simulations. We chose a top down approach [155, 160, 161, 167–170, 172, 175–177, 185, 271] for the reasons expounded by Ouldridge et al. [177]. After fixing the relative strengths of the base-base interactions with experimental thermodynamic data, as detailed in Section 5.1.3 [12, 13, 240], the model has a single free parameter relating the dimensionless temperature to the experimental temperature. We obtain this conversion factor by matching the model predictions for a test sequence to an experimentally obtained melting curve as illustrated in Section 4.3.2 [174]. The experimental data used to parameterize the model [12, 13, 174, 240] were obtained in an aqueous buffer solution. The model thus has implicit electrostatics, similar to others [160, 167–169, 175–177]. Our model is parameterized to match a single ionic strength [168, 169, 177], in this case 1X Buffer A [99], which is a model system for *in vivo* conditions and should be relevant for a number of *in vitro* biochemical experiments.

For this purpose, we considered the seven block polymer hairpins depicted in Figure

4.2. Our analysis followed along the lines of the method described in Section 4.3.5 [174]. To establish the mapping, we first simulated the 5'- $A_5C_5T_5$-3' hairpin between the dimensionless temperatures T = 0.25 and T = 0.50. The system was initialized as a comb and allowed to relax fully before collecting data. At low temperatures, it takes quite some time for the hairpin to close but the resulting closed state is stable. In prior work [174] and in Section 4.3.5, we showed that we obtained the same results for the fraction of bound bases independent of whether we start in the open state or the closed state, provided that we wait for the system to equilibrate.

We used the time for the largest hairpin to close at the lowest temperature as a very conservative estimate for the equilibration time and use this time for all of the simulations. For a given single-stranded DNA sequence, we first examined the simulation of the coldest dimensionless temperature, T = 0.25, and waited until the first closure event. The BD time step for which this happened was designated as the relaxation time for that sequence. For a simulation of this sequence at a given temperature, we waited until we reached this equilibration time and then sampled for 9 equilibration times. This allowed us to capture many opening and closing events at the melting temperature.

There are many possible ways for a hairpin to form Watson–Crick base pairs, but we have shown in Section 4.3.7 and prior work [174] that the best way to compare the open/closed state of the system to the experimental data is to time average the number of "correctly" bonded pairs at a given temperature. By correct, we mean that the pairing leads to a completely bonded stem. Since the bonds in the present model are directional, two bases were considered bonded if (i) they possess an allowed angle for hydrogen bonding (see Figure 5.1) and (ii) their center-to-center distance was less than 0.3 nm.

The simulations produced data at discrete temperatures, which we fit with a sigmoidal function. The experiments were conducted in a common and biologically relevant buffer, Buffer A [99]. We then obtained the conversion between the simulation and experimental temperature by shifting the simulated melting curve so that the melting temperature of the simulation, corresponding to the midpoint of the height of the sigmoidal function, corresponds to the midpoint in the fluorescence intensity of the experimental data [174]. This analysis led to the conversion factor T(K) = 1150 T. Our one degree of freedom was thus used to fit the melting temperature for the 5'- $A_5C_5T_5$-3' hairpin.

Table 5.4: Comparison of simulation and experimental data for DNA hairpin melting experiments.

Sequence	Chapter 4 Experiment [174]	Simulations [166, 174]			
		Chapter 4 (two bead model)		Chapter 5 (three bead model)	
	T_m(K)	T_m	R_a^2	T_m	R_a^2
5'- $A_5C_5T_5$-3'	341	341	0.995	341	0.996
5'- $A_5C_{10}T_5$-3'	337	338	0.942	338	0.979
5'- $A_7C_5T_7$-3'	328	343	0.450	327	0.940
5'- $A_7C_{10}T_7$-3'	330	343	0.540	329	0.942
5'- $A_5G_5T_5$-3'	329	340	0.622	331	0.928
5'- $A_5G_{10}T_5$-3'	341	336	0.713	338	0.902
5'- $G_5A_{10}C_5$-3'	338	350	0.514	339	0.915

For each hairpin in Table 5.4, we determined the simulated melting point and the coefficient of multiple determination adjusted for the number of parameters in the sigmoidal model, R_a^2, between the experimental and simulated melting curves. These R_a^2 values were obtained from plots similar to Figure 5.5. The plots for the other hairpins, which are essentially the same, are included as Figure 5.6. While it is difficult to propagate the error in the experimental data, we estimate that it is around $\pm$2K. The data for the two-bead model [172, 174] are included in Table 5.4 for comparison.

By definition, the simulated melting point for the 5'- $A_5C_5T_5$-3' hairpin is identical to the experimental value. All other simulation data in Table 5.4, Figure 5.5 and the Figure 5.6 can be considered predictive. Given this parameterization, the simulations certainly should capture the melting for the slightly larger loop in 5'- $A_5C_{10}T_5$-3'. Indeed, even the simple two-bead model [172] captures the latter experimental data. The challenge is to capture the data for all sequences, including the width of the transition [174, 177] and the shoulder [160]. The very high values of R_a^2 achieved by the current model indicate that we indeed accomplished this task. Moreover, the high R_a^2 values suggest that no bias was introduced by the arbitrary choice of 5'- $A_5C_5T_5$-3' for the parameterization. To confirm this conjecture, we repeated the parameterization procedure using each of the other sequences in Table 5.4. Out of the 42 predicted melting temperatures produced from all possible combinations, the largest difference between simulation and experiment was 3 K.

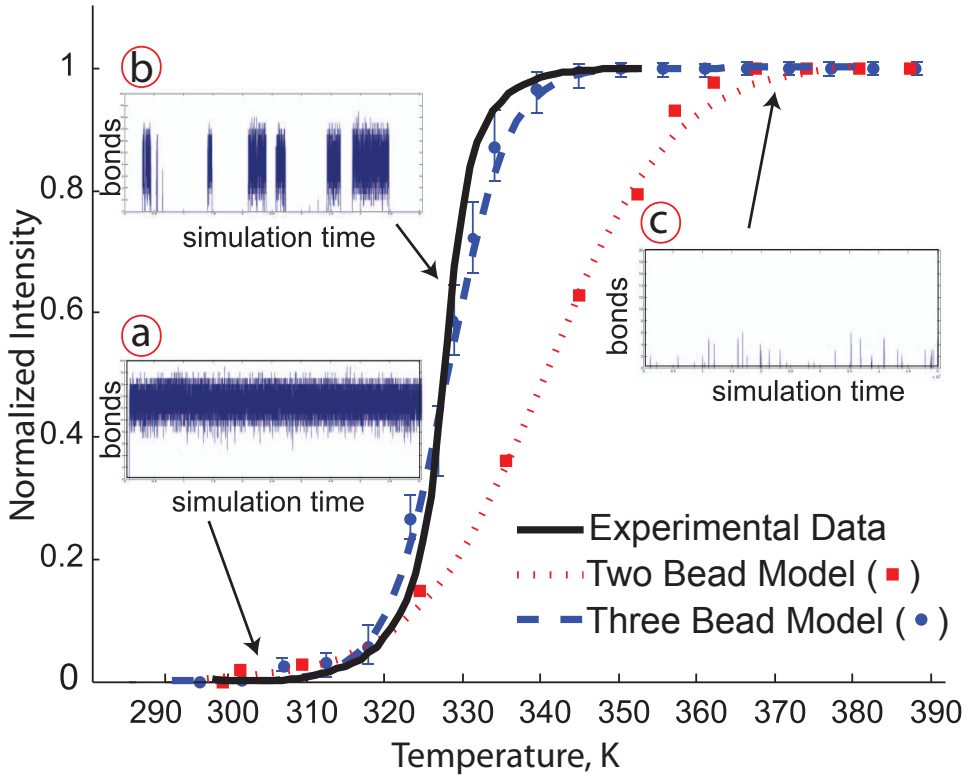

Figure 5.5: Comparison of the thermodynamics of experimental and simulated hairpin open-close transitions for the sequence 5'-$A_7C_5T_7$-3'. The solid (black) line is the experimental data and the dashed lines are the sigmoidal fits to the simulated data for (i) (blue) the new three bead model [166] and (ii) (red) the simulation data for the two bead model [174]. The insets show the number of correctly aligned bonds for (a) the closed state, (b) the transition, and (c) the open state as a function of the simulation time. The amount of data in the traces is 11% of the total sampling time.

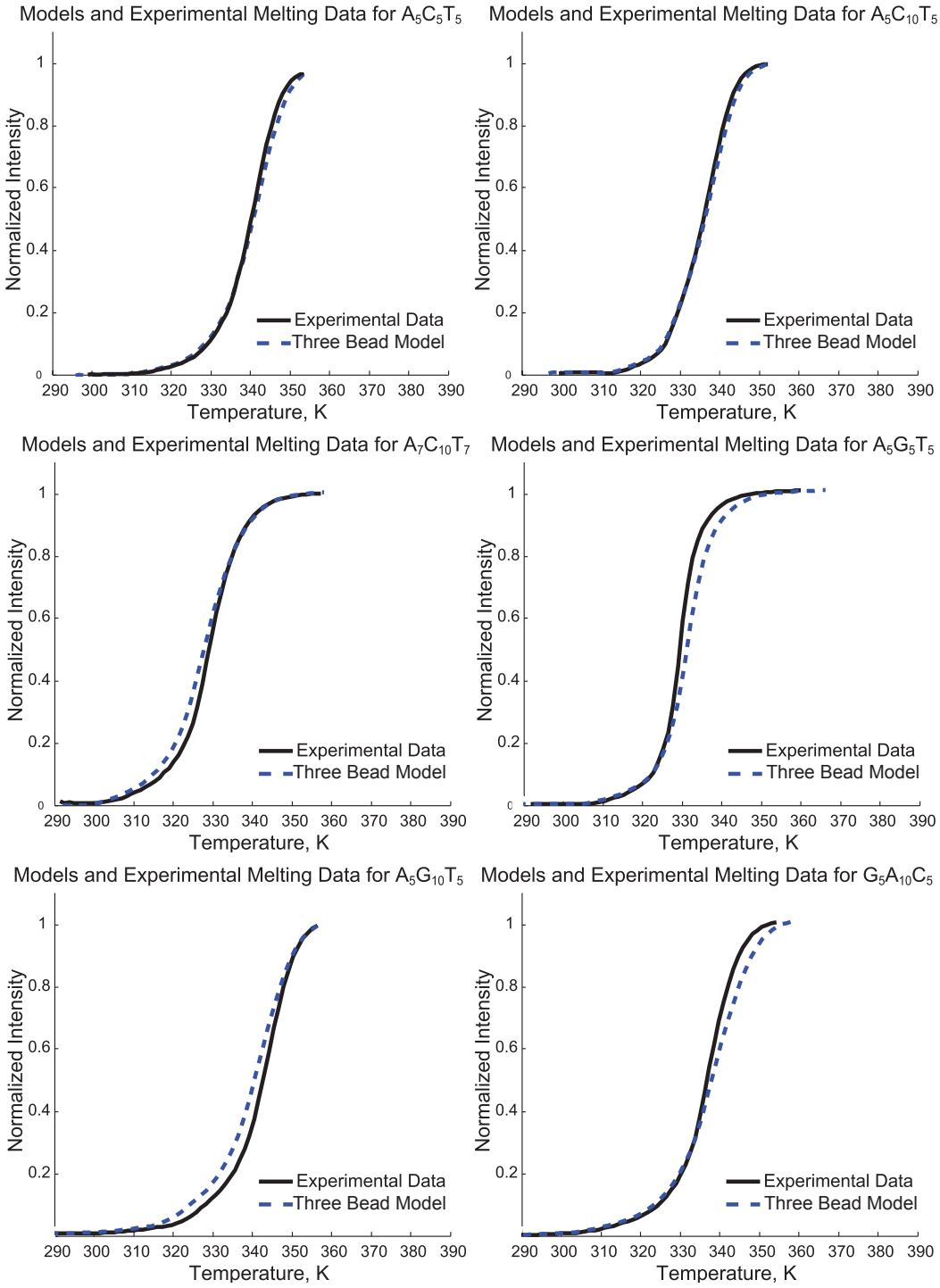

Figure 5.6: Comparison of the experimental data and three bead model for the other ssDNA sequences. The solid (black) line is the experimental data and the dashed (blue) line is the sigmoidal fit to the simulated data for the new three bead model [166]. Each inset depicts a row of Table 5.4.

Table 5.5: List of sequences used to evaluate the dsDNA structure.

Random 25-mer dsDNA Sequences Used
5'- CCGAGTACGTCGGGCGCTTATAGTG-3'
5'- CAGAACACTTTCTACACCCTGACGC-3'
5'- TGCCTGAACGATAAATCCGATGGCT-3'
5'- GGGTTCATCCGCTACCGTGCTCCCT-3'
5'- GTATGCCACGAATACTCTCTGCAGA-3'
5'- TTATCGCTCGAGGTGCTTGGCTGGC-3'
5'- TGTAAGCCACGAATACCGGCCCCGA-3'
5'- GATAAGCGTTTTAGAGTGTCATTTG-3'
5'- TAAGCTTGGGCTGTCTTTTAGGAGG-3'
5'- AAATGAATTCGCTCACGCCCGGTTA-3'

5.2.3 Structure of Double-Stranded DNA

Finally, as the last vital test for the new model's ability to match or predict experimental data, we look at its tertiary structure formations, particularly dsDNA. Our goal in simulating this system is not to study the mechanism of hybridization *per se*, but rather to establish that our model spontaneously forms a near B-form DNA structure over a wide range of sequences and temperatures. Watson–Crick bonds dominate in dsDNA [11–13, 15, 21, 59], so this system also demonstrates that the non-canonical bonds can be included in the model without disrupting the canonical conformation. We used ten random dsDNA sequences, listed in Table 5.5, containing 25 base pairs and performed simulations at five different dimensionless temperatures, corresponding to a temperature range of 290-315 K, based on the conversion factor we obtained in Section 5.2.2. We purposely chose temperatures in which the dsDNA does not melt in order to be better able to examine dsDNA structural characteristics in the canonical state. We presented melting data in Section 5.2.2 which highlighted the importance of non-Watson–Crick bonds. We initialized the two complementary ssDNA sequences as an anti-parallel ladder, with the backbones straight and the complementary bases separated by 1.5 nm.

At the start of the simulation, Watson–Crick bonds quickly formed between nearby, complementary bases on opposing strands. These bonds led to local twisting of the chain, with a mixture of right-handed and left-handed structures nucleating at different locations. The twists propagated along the sequence and the chain eventually achieved a homogenous chirality. Of 50 independent simulations conducted with the 10 random dsDNA sequences listed in Table 5.5, we found that right-handed helices are formed in 62% of the structures. This result is reasonable

since our model has no built-in handedness, torsional constraints that favor B-DNA [160, 161, 164, 167–169], bottom up parameterization from an all atom B-DNA model [157, 164, 182, 183, 202, 203], or a method to remove the stacking interactions in left-handed twist [176]. Indeed, other models of this type lead to equilibrated structures that are sometimes left-handed [155, 164, 172].

To estimate the equilibration time for the double-stranded DNA simulations, we initialized the sequence, 5'-CCGAGTACGTCGGGCGCTTATAGTG-3' and simulated the coldest dimensionless temperature, T = 0.25. We then computed the average number of bases per turn (calculated from one strand) as a function of time. We considered the duplex equilibrated when this value became constant and used the corresponding time as a conservative estimate for the equilibration of all sequences at all temperatures. We started sampling after two equilibration times and the sampling continued for 8 equilibration times.

In Figure 5.7, we provide a snapshot of the dsDNA configuration from our simulation and the structural data obtained for the right-handed helices, along with representative experimental data for A- and B-DNA [11, 13], described in Sections 2.2.2 and 2.2.1, respectively. The results are essentially unchanged if we include the left-handed helices, since the potentials are symmetric with respect to the handedness.The present model produces an overall double-stranded structure that is closest to B-DNA. The simulated structural data, averaged over all sequences and temperatures, agree well with experimental data. We only included data that can be reasonably resolved by the model at our degree of coarse-graining. For example, although it is possible to compute the roll using a multi-bead model for the bases [157], we cannot resolve it here because each base is only represented by a single sphere. The major and minor groove spacings were measured between the edges of the excluded volume cutoffs for the relevant beads, rather than from their centers, to correspond to the measurements obtained by NMR [13]. These distances are at the limit of what we can resolve at this level of coarse-graining, so the major and minor groove widths should be considered estimates.

We also estimated the persistence length of a 50 base pair dsDNA using an extrapolation method [272]. In contrast to our calculation of the persistence length of ssDNA for this model described in Section 5.2.1, we constructed the backbone vectors at a length scale corresponding to approximately one turn [160, 161, 177]. Since we have 10.8 bases per turn (see Figure 5.7), we constructed vectors between every 11 nucleotides and computed the initial decay of the autocorrelation function. Depending

	Simulation		Experiment		
Property	Value	Deviation	A-DNA	B-DNA	Reference
Major groove (nm)	1.0	0.22	0.27	1.17	[13]
Minor groove (nm)	0.6	0.12	1.10	0.57	[13]
Helix diameter (nm)	2.33	0.17	2.55	2.37	[11]
Rise (nm)	0.33	0.07	0.29	0.34	[13]
Base pairs per turn	10.8	0.61	11	10-10.6	[11]
SPS Angle (°)	60.7	10.6		≈ 62.5	[13]

Figure 5.7: Structural data for dsDNA. The standard deviation is over all sequences and temperatures. The SPS angle measures the angle formed between the phosphate and two sugar beads along the backbone of one of the ssDNA strands in the duplex. A depiction of double-helix DNA is included, the backbone is comprised of the smaller (orange) beads, with the light (orange) beads representing the phosphate beads and the dark (orange) beads representing the sugar beads. The 5'-end of the sequences is depicted by the dark (purple) beads. The four bases A, C, G, T are represented by the blue, green, red, and yellow beads, respectively. The major and minor grooves can be seen in the regular structure of the double helix.

on the reference bead, we obtained a persistence length of 47 ± 8 nm (sugar-to-sugar), 48 ± 7 nm (phosphate-to-phosphate), and 46 ± 10 nm (base-to-base). Although the measurement must be considered a rough estimate it shows that our model is at least approximately in line with the accepted standard of about 50 nm or 150 bp [11–13, 59].

Any model that includes stacking produces helicity. A particularly notable feature of our model is the sugar-phosphate-sugar (SPS) angle, which is close to but not the same as the sugar pucker (or glycosyl) angle. The 3SPN model [160, 161, 167] maintains an SPS angle close to the experimental value for B-DNA by imposing a dihedral angle potential on the backbone [160, 161, 164, 168, 169]. In our model, the SPS angle arises from the directionality of the stacking interactions without the need to also apply a dihedral potential. As would be the case with a spherical potential, the stacking interactions are increased as the bases along one strand move closer together. However, with our directional bonding, the stacking energy is most favorable when the vectors drawn from each sugar to its bonded base are parallel. To maximize the interaction, the backbone flexes and the phosphates are pushed towards the outside of the chain to form the SPS angle. Note that there is no bending penalty for forming a SPS angle because the bending energy is defined between the sugar trios. The result is the formation of a dihedral angle without the need for a dihedral potential. From a computational standpoint, our method and that employed in the 3SPN model [160, 161, 167] are roughly equivalent; the cost for computing the dihedral angles is somewhat less than that for computing the θ-dependent term appearing in the base-base interactions, but the θ-dependent term provides both the SPS angle and directional bonding.

From a versatility standpoint, our approach offers some advantages; for studying single-stranded DNA in the *in vivo* conditions mimicked by Buffer A [98], our method appears to have some conveniences compared to the 3SPN model [160, 161, 167]. As noted by Knotts et al., constraining a model *a priori* to favor B-DNA makes it difficult to study transitions to other forms. We suspect that our model does not suffer from the same limitation. It is true that we obtained the various energies for stacking, cross-stacking, and hydrogen bonding from experiments on B-DNA [12, 13, 21], and therefore this may explain some of the mimicry behaviors of dsDNA

to this form. In fact, in the double-stranded conformation our model shows high B-DNA structure, particularly with the SPS angle. In the absence of Watson–Crick base pairs, the SPS angle changes dramatically in our model; in dsDNA the average SPS angle is $60.7° \pm 10.6°$ which is much smaller and more constrained than the $97° \pm 52°$ that was measured in ssDNA. The tighter and less variant measure of the SPS angle denotes the additional stiffness of the dsDNA and is in line with experimental measurements for the angle and the persistence length, yet, we are able to capture both single- and double-stranded characteristics. However, in contrast to other models, our technique can transition smoothly from ssDNA to dsDNA thereby permitting study of hybridization, melting, and other interchain interactions.

5.3 New Features and Continued Limitations of the Model

The new three bead model has a number of improvements compared to the simple two bead model [172–174] defined in Section 4.2 and compared to the experimental hairpin melting data collected and analyzed by the methods described in Section 4.3. Explicitly, we now have directionality along the backbone, a major and minor groove, experimentally parameterized bonding energies, anisotropic bonding, and non-Watson–Crick bonds. We previously speculated that the main reason why the two bead model [172–174] fails to capture the melting transitions of these block-polymer sequences is the absence of Hoogsteen bonds [174]. As we can see in Figure 5.8, this certainly appears to be the case for the 5'-$A_5C_5T_5$-3' sequence. At low temperatures, the hairpin is stabilized by Watson–Crick bonds, as expected. When the hairpin opens in Figure 5.8, both the adenine and cytosine bases form Hoogsteen bonds, with the cytosines adopting an i-motif, as described in Section 2.5.2 [42, 273]. These Hoogsteen bonds are relatively strong and need to be undone in order to fold into the closed state. Thus they represent not only a change in the free energy landscape but nontrivial kinetic traps. While directional bonds are certainly important for modeling long $A \cdot T$ or $G \cdot C$ tracts [161], we suspect that Hoogsteen bonds will also be important when these simulations are intended to capture experimental data.

Our results thus far have highlighted the advantages of our new model. The structures can smoothly move between a flexible ssDNA and the more constrained dsDNA while still capturing most of the features of the double helix. In the single-stranded state,

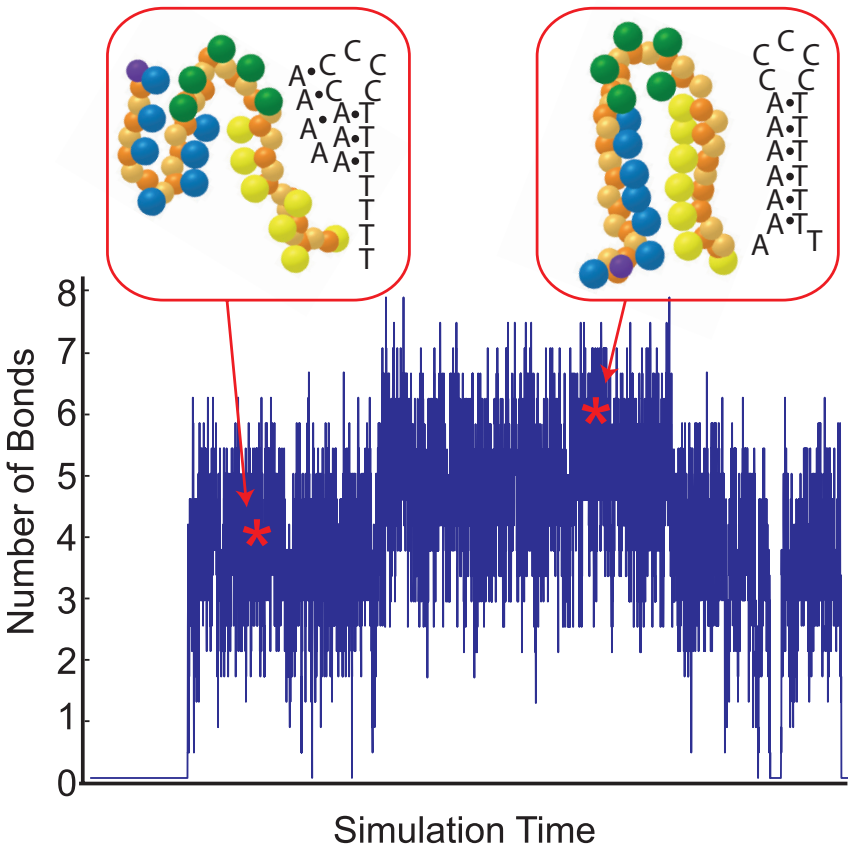

Figure 5.8: Detailed trajectory for the number of correct bonds (Metric 2 defined in Section 4.3.5) in the stem for the sequence 5'- $A_7C_5T_7$-3' at a temperature of 327 K, which is near the melting temperature. The snapshots show two examples of the hairpin configurations obtained at the times indicated by the (red) stars. The structure on the left, with three paired stem bases, is stabilized by Hoogsteen bonds. The structure on the right shows fraying of the hairpin ends. The light (orange) beads along the backbone are the phosphate groups while the dark (orange) beads are the sugar groups. The 5'-end of each ssDNA chain is marked by a dark (purple) bead. The base beads shows are A (blue), C (green) and T (yellow); G (red) is not included on this hairpin sequence.

we are also able to model important physical scenarios that require Hoogsteen bonds, such as G-quartets and I-motifs. However, there are some manifest shortcomings to the model that we discuss here.

The model does not possess any inherent chirality that enforces right-handed double helices. In other models, the chirality has been enforced using dihedral potentials [160], which prevent a smooth transition to single-stranded DNA, or by simply turning off the stacking interactions if the helix is left-handed [176]. Although we observed a number of left-handed helices, we do not view this as a critical shortcoming of the model. The initial conditions we used in our simulations for dsDNA are unbiased; two opposing combs have no initial handedness. The eventual handedness of the helix is strongly determined by the nucleation of a local region of twist, in particular if this occurs in a GC-rich region. Indeed, when we used the same random numbers but changed the sequence, all of the resulting helices had the same handedness (which happened, by chance, to be right-handed). If our goal was to investigate some property of double-stranded DNA, we simply need to initialize the chain as a right-handed helix. The energy barrier between handedness is enormous and well beyond the time scale for any reasonable isothermal simulation.

The parameterization we use here is only valid for a single ionic strength, since we determined the value of the energy scale ϵ using experimental data for Buffer A. The model can be modified to account for ionic strengths in the manner proposed by Knotts et al. [160]. First, a screened Debye-Huckel potential needs to be added between phosphate beads. Since electrostatic interactions on the backbone stiffen the DNA [274, 275], we then need to adjust the bending potential to recover a persistence length appropriate for single-stranded DNA. If the electrostatic potential is weak compared to the hydrogen bonding, then the value of ϵ is unaffected by the inclusion of explicit phosphate charges. If not, we can use a multiplicative factor for the ϵ in Equation 5.4 to set the relative strengths, analogous to the 3SPN model [160].

Perhaps the most critical issue is our use of the same excluded volume interaction, independent of the base identity. For the problems we studied here, this was not an issue but the base sizes will play an important role if the model is used to study mismatches. Our model thus incorrectly accounts for a non-Watson–Crick mismatch since the hydrogen bonding energies for the Hoogsteen bonds are similar to their Watson–Crick counterparts. In reality, the mismatch should lead to substantial excluded volume interactions, which in turn disrupt the stacking and thus the local stability of the duplex. Fortunately, the remedy to this problem is straightforward —

the homogeneous excluded volume interactions need to be replaced by a more realistic model. In the 3SPN model [160], for example, different bases are represented by different sized beads and bond angles.

Moving forward, we should also point out an additional issue with our model relative to the 3SPN model [160, 161, 167], namely, the use of anisotropic potentials. Most molecular dynamics solvers, such as GROMACS [276], only permit spherical potentials. We do not see an easy route towards using spherical potentials in a coarse-grained model and still moving smoothly between double-stranded DNA and single-stranded DNA. Removing the anisotropic potentials requires adding dihedral potentials, which then bias the shape of ssDNA towards the B-form of dsDNA.

5.4 Conclusions

We have shown here that restricting hydrogen bonding to Watson–Crick base pairs is insufficient to capture the higher order structure of many sequences of DNA, even for relatively pedestrian systems such as DNA hairpins. Rather, it is essential to include both canonical and non-canonical base pairs [11–13, 17, 18, 59]. Hoogsteen bonds stabilize multi-body secondary structures in single-stranded DNA, such as folded intra-strand G-quartets [11–13, 42, 59, 277] and i-motifs [12, 13, 42, 273]. They are also the glue that binds triple-stranded DNA [12, 13]. Although the applications of such a model to single- and triple-stranded DNA are apparent, the inclusion of Hoosteen bonds also does not impact simulations of double-stranded DNA. Indeed, while the conventional wisdom embodied in existing coarse-grained models [155, 157, 158, 160, 161, 163, 164, 167–170, 172, 173, 175–177, 182, 183, 185, 202, 203, 241, 271] postulates that dsDNA only utilizes Watson–Crick bonds, recent experimental data [21] showed transient formation of both A · T and G · C Hoogsteen pairs.

We will examine the advanced applications of the newly developed three bead model in the following chapter, Chapter 6. We focus our efforts on systems that cannot be simulated using other coarse grained models of DNA, but instead rely on the added components of the model detailed in this chapter. For example, we will look at modeling the behavior of a single-stranded structure that relies not only on Hoogsteen bonding, but also on multi-base interactions to form a set of G-quartets. In addition, as was described in Section 2.2, we will display several other non-natural forms of double-stranded DNA. Finally, we will explore the model's ability to investigate the

formation of the two triple-stranded DNA structures catalogued in Section 2.4.

6

Complex Nucleic Acid Secondary and Tertiary Modeled Structures

I cannot think of a single field in biology or medicine in which we can claim genuine understanding, and it seems to me the more we learn about living creatures, especially ourselves, the stranger life becomes. [...] We have not reached solutions; we have only begun to discover how to ask questions.

Lewis Thomas, *On Science and Certainty*, October 1980 [278]

6.1 Introduction to Complex Nucleic Acid Structures

New and complex features continue to be discovered for nucleic acids. Despite the passing of more than three decades since Thomas [278] described the strangeness of life, we continue to be puzzled by the processes fundamental to every living cell. As our understanding of DNA and RNA has grown, we have found that it is active in many different, elaborate, and convoluted conformations not just the standard canonical forms.

The present three bead model has a number of improvements compared to the simple two-bead model [172] used in our previous comparison with the experimental hairpin melting data [174]. Explicitly, we now have directionality along the backbone, a major/minor groove, experimentally parameterized bonding energies, anisotropic bonding, and non-Watson–Crick bonds. As was outlined in Chapter 5, the three

bead model developed not only can capture the experimental hairpin melting curves but also characteristic features of near B-form double-stranded DNA. However, it is how well this new model can capture intricate and complex structures that further tests the model's utility. In this chapter we will provide such examples: a single-stranded apatmer, several double-stranded complexes, and two triple-stranded structures.

6.2 Complex Structures of Single-Stranded Nucleic Acids

Single-stranded nucleic acids can form a wide array of intricate conformations as they fold and bind. Beyond the ssDNA hairpins that were used to validate the model, aptamers, and in particular DNAzymes can be captured with this model. As was discussed in Section 2.3, many of these structures rely on non-Watson–Crick bonds. Two such ssDNA structures have been modeled: (i) the thrombin aptamer that relies on Hoogsteen bonding, and (ii) the 10-23 DNAzyme of which the tertiary structure cannot be found experimentally.

6.2.1 Thrombin Aptamer

As discussed in Section 2.3.1, aptamers are sequences of ssDNA or RNA that bind selectively to proteins. While the methods for isolating aptamers from a random library of nucleic acids [39, 279, 280] are relatively well developed, the selection method provides little insight into the reasons for their high affinity and specificity towards particular proteins. However, it is reasonable to assume that the secondary and tertiary structures of aptamers substantially contribute to their activity. As a result, coarse-grained simulations could play an important role in understanding aptamer activity. However, since aptamers are short and single-stranded, they are often governed by non-Watson–Crick interactions and typically do not form the standard double-helix. These features preclude aptamers from being modeled by other nucleic acid models and make them uniquely suited for the three bead model developed in Chapter 5.

To illustrate the power of such simulations, we looked at the folding pathway of the DNA aptamer 5'- GGTTGGTGTGGTTGG-3', which binds to thrombin, a blood

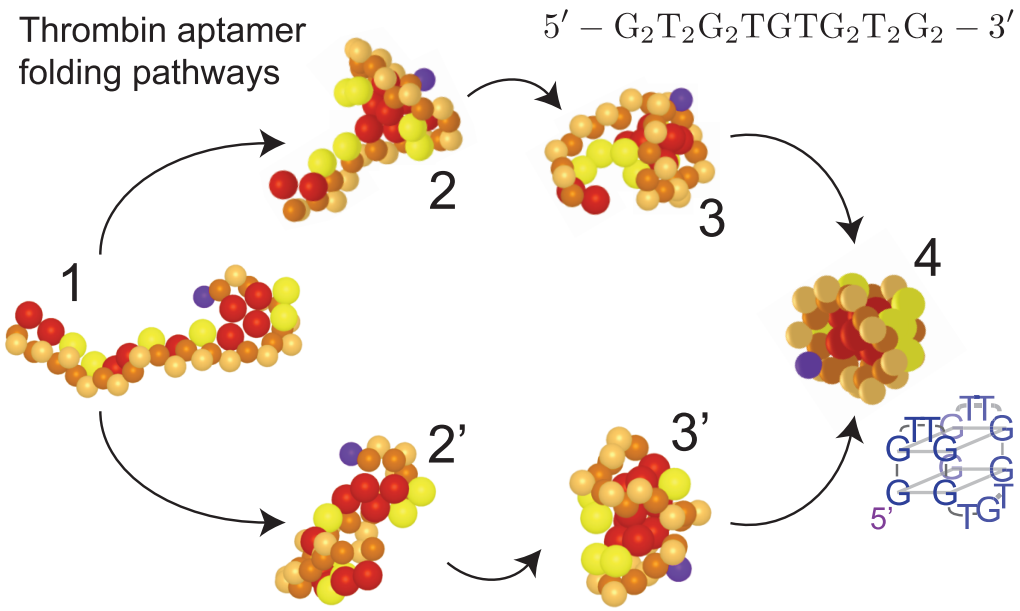

Figure 6.1: Folding pathways for the thrombin aptamer, 5'-$G_2T_2G_2TGTG_2T_2G_2$-3'. The numbers indicate steps in the most common pathway ($1 \rightarrow 2 \rightarrow 3 \rightarrow 4$) and a secondary pathway ($1 \rightarrow 2' \rightarrow 3' \rightarrow 4$). The wire-frame diagram shows a cartoon of the bead positions at the final snapshot time, the dark (purple) bead denotes the 5' end of the DNA molecule.

clotting protein [39, 148]. NMR studies [91] indicated that this aptamer forms two G-quartets. The thrombin aptamer contains only guanines and thymines, excluding Watson–Crick bonds as manner of base-base interactions and represents a sufficiently complex single-stranded nucleic acid structure. As this structure results from Hoogsteen bonds, the thrombin aptamer is an ideal candidate to study with the present non-canonical model. Cations such as K^+ or Na^+ may stabilize the G-quartet structure [43]. Although our model does not have explicit ions, the experimental data used to tune our model [12, 13, 174, 240] includes these ions and may implicitly account for such electrostatic effects [160, 167–169, 175–177].

To investigate the folding of this aptamer, we initialized the ssDNA as a comb and performed eight simulation runs. The simulation temperature was 298 K, which corresponds to experiments and should promote the folded state. Figure 6.1 shows the evolution of the structure as a function of time. We observed two distinct pathways. In both pathways, the distal guanines form a single G-quartet. In Figure 6.1, we show the case where this bonding occurred at the 5'-end (1), but this can also occur at the 3'-end. In the more common pathway (six of eight simulations), the next bonding step forms a triplex in the interior (2) while leaving the pair of guanines at the other end of the chain unbonded and able to fluctuate (3). To form the final structure (4),

the unpaired guanines on the free end need to disrupt the triplex and create a pair of G-quartets. While this folded state (4) is thermodynamically favorable, the kinetics for the final step ($3 \rightarrow 4$) are slow compared to the preceding steps. In the less common pathway (two of eight simulations), both distal ends fold in on themselves to create a pair of G-quartets (2'). The entire molecule then folds about the center axis (3') to stack the G-quartets (4). In both pathways, the final state, depicted in the wire diagram in Figure 6.1, is consistent with NMR data [91].

As shown by the thrombin example, the three bead, non-Watson–Crick model is able to capture the final, folded, NMR verified structure of a complex aptamer. In addition, it points towards possible pathways for the folding of the aptamer; this information may be fundamental to predicting the activity and binding of each molecule. Aptamer discovery, by definition of the SELEX process, involves the testing of libraries containing 10^{12} aptamer variants [279]. The combination of SELEX and high-throughput sequencing has demonstrated that SELEX can produce hundreds or even thousands of aptamers that satisfy some minimum characteristic, often binding affinity. However, continued rounds of SELEX often cannot further distinguish between the aptamers to find the strongest ones. This modeling approach offers a method to circumvent this problem; information garnered from modeled structural analysis may be able to further distinguish and sort these aptamers. Future work has been proposed to develop a tandem experimental and simulation aptamer sorting protocol.

6.2.2 10-23 DNAzyme

Both the two and three bead models can also be used to find the tertiary structure of the 10-23 DNAzyme and its corresponding substrate complex. Beginning with its basic secondary structure, these models can predict the 10-23 DNAzyme - RNA substrate global three-dimensional structure. This complex resists crystallization and thus cannot be found by experimental endeavors. Further details will be given in Chapter 7 on the development of this structure.

6.3 Complex Structures of Double-Stranded Nucleic Acids

Other non-traditional nucleic acid complexes can also be modeled; double-stranded structures beyond the near B-form dsDNA can be represented in the model. Since the model has no inherent chirality, and in fact forms left- and right-handed helices with theoretically equal probability, it can be used to model left-hand twist dsDNA molecules. In addition, both P- and S-forms of DNA can be captured. Finally, the model can be used as a rough approximation for dsRNA structures and further parameter refinement would allow it to better capture RNA characteristics. Here the three bead model's geometric characteristics are compared to Z-, P-, and S-DNA.

6.3.1 Left-Handed DNA

Although it does not capture the exact properties of Z-form DNA, having the ability to form left-handed structures gives the model increased utility. Two single strands of DNA are initialized as an anti-parallel ladder, with the backbones straight and the complementary bases separated by 1.5 nm. At the start of the simulation, Watson–Crick bonds quickly form between nearby, complementary bases on opposite strands. These bonds lead to local twisting of the chain with a mixture of right-handed and left-handed structures nucleating at different locations. Shown in Figure 6.2, twists propagate along the sequence of the chain and eventually achieve a homogeneous chirality. As was discussed in Section 5.2.3, of the limited simulations, 38% of the sequences twisted into a left-handed configuration. The data presented in Table 6.1 gives the basic geometric characteristics of the left-handed dsDNAs (which is the same as in Figure 5.7 except the rise per base pair is defined in the opposite direction). The geometric characteristics of Z-DNA is also included for easy comparison. Although the model does not closely capture the specific characteristics of Z-DNA (in fact it has the same A/B-DNA form but with left-handed twist), it does have the possibility to make stable left-handed helices. The propensity of the three bead model to form both right- and left-handed helices can be efficiently handled without the need to eliminate the possibility of such structures all together as is done with a dihedral potential in other models [160, 161, 164, 168, 169]. This can be done in our model by several approaches: (i) initializing a configuration with a right-handed twist to nucleate helix

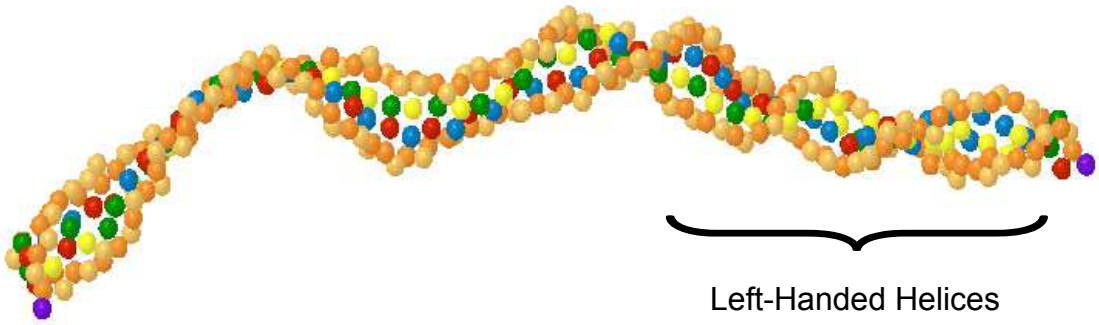

Figure 6.2: After two, straight, ssDNA strands are initialized, they quickly diffuse together and pair via hydrogen bonding. The complex spontaneously nucleates both left- and right-handed helices. In this simulation snapshot, the left-handed helices are propagating through the chain. The dsDNA is comprised of 58 nucleotides on each strand, with the sequence 5'- GGACAGGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT CTCTTTCTCATTATTTGC-3' and the other chain is the Watson–Crick complement. The base beads are shown as the same size as the backbone beads to facilitate viewing.

Table 6.1: A summary of geometric descriptors of the three bead simulation model and Z-form DNA.

Property	Simulation Model	Z-DNA
Handedness	left-handed	left-handed
Major groove (nm)	1.0	convex
Minor groove (nm)	0.6	0.40
Helix diameter (nm)	2.33	1.8
Rise (nm)	-0.33	-0.35 to -0.39
Base pairs per turn	10.8	12
SPS Angle (°)	60.7	-

formation; (ii) using the coordinates of a right-handed B-DNA sequence as the initial structure, or (iii) eliminating any left-handed helical conformations from a data set after simulation is complete. Despite some detractors, such as wasted simulation time, by not including a dihedral angle, the three bead model is able to smoothly move between single-stranded and double-stranded DNA conformations, as discussed in Section 5.3. In Figure 6.2, a 58 base pair sequence (for one strand, the other the Watson–Crick complement) was simulated.

6.3.2 P-DNA

Pauling, or P-DNA, discussed in Section 2.2.4, is a two oligonucleotide structure with its bases extended radially on the outside of the double-strand and the backbone

Figure 6.3: A simulation of P-DNA exhibits the characteristic phosphate and sugar groups (backbone beads in orange) in the interior and the nitrogenous bases extending radially outward on the surface of the duplex. The ends are fixed in place and a weak pulling force was applied to a near B-form DNA duplex. The dsDNA is comprised of 58 nucleotides on each strand, with the sequence 5'- GGACAGGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT CTCTTTCTCATTATTTGC-3' and the other chain is the Watson–Crick complement. The base beads are shown as the same size as the backbone beads to facilitate viewing.

Table 6.2: A summary of geometric descriptors of the three bead simulation model and P-form DNA.

Property	Simulation Model	P-DNA
Handedness	right-handed	right-handed
Major groove (nm)	-	-
Minor groove (nm)	-	-
Helix diameter (nm)	0.6 to 1.8	0.69 to 1.6
Rise (nm)	0.55	0.585
Base pairs per turn	4-5	2.62
SPS Angle (°)	172	-

molecules on the inside. It is believed to be produced as a result of relatively weak supercoiling, such as is probably encountered during replication, transcription, and recombinational repair [11, 13, 14, 59, 60, 62]. This can be mimicked by weakly pulling (less than 3 pN) on B-form dsDNA in the simulation since writhe and twist are held constant. Our simulations exhibit similar behavior to P-form DNA and can be seen in Figure 6.3. One double-strand of sequence 5'- GGACAGGTCTCTCTCTCTCT CTCTCTCTCTCTCTCTCTCT CTCTTTCTCATTATTTGC-3' , was simulated, beginning in its near B-form with one end fixed in space and the other end weakly (non-dimensional force of 0.1) pulled along its helical axis. Although limited simulations were conducted, we can take rough estimates for the geometric characteristics which are given in Table 6.2.

6.3.3 S-DNA

When the pulling forces are much stronger than in P-DNA formation, stretching of the molecule forms S-DNA. Extensional experiments on double-stranded B-DNA have shown that the over stretching transition occurs when the molecule is subjected to forces more than 65 pN. S-DNA can take two conformations depending on whether the ends are allowed to freely rotate. If the ends of the molecule are unrestricted, it will be "ladder-like" and can be considered an unwound helix. As can be seen in Figure 6.4, the simulation under similar conditions (unrestricted ends and strong pulling force that is 100 times the force used to produce P-DNA) produces a S-DNA, ladder-like conformation. This simulation was conducted at T = 0.25 and the force was slowly increased (0.1 force units per 1000 simulation steps) to allow time for the forces on the end to fully propagate the entire duplex. If we continue to increase the force beyond what is shown in Figure 6.4, the forces cause denaturation of the two strands. If the duplex is not allowed to freely rotate, forces at this scale cause breakages along the backbone.

Figure 6.4: A simulation of S-DNA exhibits the characteristic ladder-like configuration when strong forces are applied to a B-DNA and the ends are allowed to freely rotate. The dsDNA is comprised of 58 nucleotides on each strand, with the sequence 5'- GGACAGGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTTTCTCA TTATTTGC-3' and the other chain is the Watson–Crick complement. The base beads are shown as the same size as the backbone beads to facilitate viewing.

Table 6.3: A summary of geometric descriptors of the three bead simulation model and S-form DNA.

Property	Simulation Model	S-DNA
Major groove (nm)	-	-
Minor groove (nm)	-	-
Helix diameter (nm)	1.6	1.4
Rise (nm)	0.5	0.40
Base pairs per turn	-	ladder-like
SPS Angle (°)	180	nearly linear

Although the three bead per nucleotide model does not precisely capture all of the features of the Z-form, P-form, and S-form double-stranded DNA, it is able to capture many of the general features of these complex and intricate structures. This model can also be applied to other forms of DNA and even double-stranded RNA, as described in Section 2.2. With minor adjustments to the parameters, it may be possible to better simulate these and other non-canonical forms. However, even without these refinements the great strength of the three bead model is that it is able to smoothly move between different single-, double-, and even triple-stranded structures.

6.4 Complex Structures of Triple-Stranded Nucleic Acids

The generalized double helix can, under certain conditions, accommodate a third strand in its major groove as described in Section 2.4 [12, 13, 69]. Two of the segments are parallel with each other while the third is antiparallel to the first two. Triplex structures have been found to play important roles in gene regulation, DNA repair, and site specific modification or cleavage [13]. Two types of triplexes can form: (i) intramolecular structures and (ii) intermolecular structures.

6.4.1 H-DNA

Although the canonical Watson–Crick double helix is the most stable DNA conformation for an arbitrary sequence under usual conditions, some sequences within duplex DNA are capable of adopting structures quite different than the canonical B-form. H-DNA, described in Section 2.4.1 and named for the hydrogen ions that stabilize it, is formed when negative super-helical stress is applied on specific sequences contained in larger dsDNA [41]. The molecule kinks and folds in the middle while twisting to unpair a section of double helix. A bubble forms and one of the single-stranded sections wind into the nearby double helix where it binds via Hoogsteen bonds to form a triple stranded section. Here, we simulated a double strand with a palindromic sequence, 5'- GGACAGGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTTTCTC ATTATTTGC-3', the palindrome comprising the $(CT)_{18}$ repeats. Since the simula-

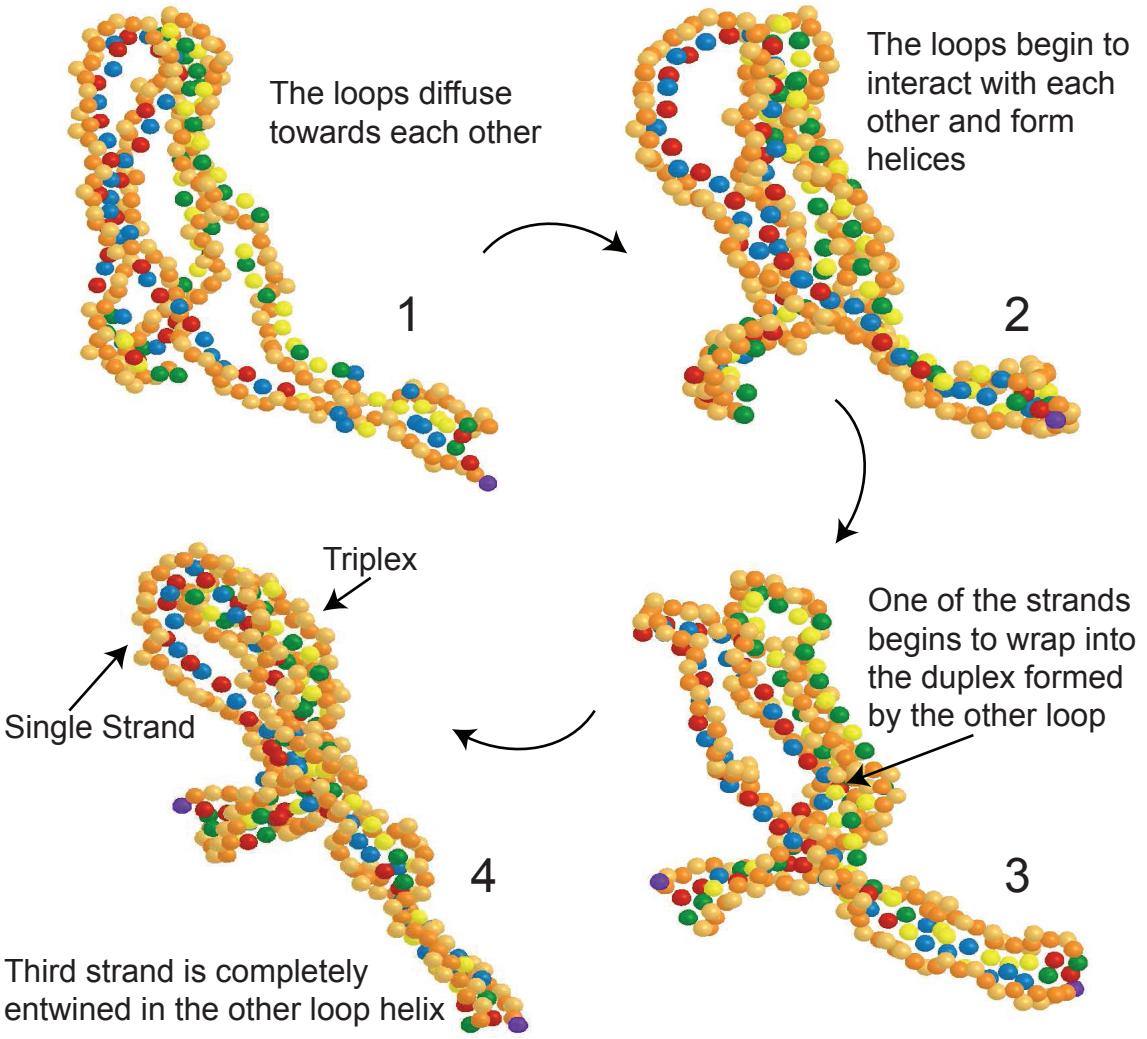

Figure 6.5: A simulation of H-DNA starting with a double-stranded, internally palindromic sequence. The simulation is initialized as two single-stranded DNAs. The simulation is biased by starting each DNA strand in an omega shape facing one another, ↄ ↄ, with the loops at a 90° angle to each other. The ends of the strands are initialized so that they are hydrogen bonded to each other. The original dsDNA is comprised of 58 nucleotides on each strand, with the sequence 5'- GGACAGGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTTTCTCA TTATTTGC-3' and the other chain is the Watson–Crick complement. The base beads are shown as the same size as the backbone beads to facilitate viewing. The formation pathway, $1 \rightarrow 4$ depicts how the two loops diffuse towards each other (1), begin to interact (2 and 3) and then finally one of the strands fully entwines in the other loop helix (4) forming both a triplex and a single-stranded structure within the regular double-stranded conformation.

tion is intended to be a proof of principle, we began the strand in a kinked conformation in the shape of two facing Ωs (ↄↃ) separated by 90° in order to bias the diffusion. As can be see in Figure 6.5, the H-DNA conformation is both single- and triple-stranded. The pathway found in this proof-of-principle example is as follows: Step 1, the loops begin to diffuse towards each other and stacking occurs; Step 2, the loops begin to interact with each other and form helices; Step 3, one of the strands begin to wrap into the duplex formed by the other loop; Step 4, the third strand is completely entwined in the other helix loop and a single-strand DNA strand remains.

6.4.2 Triplex Formation in DNA

Triplex formation plays an important role in the repair of a stalled replication fork or a break in dsDNA. In the simplest model, the strand invasion problem consists of a dsDNA and a ssDNA, where the ssDNA possesses the same sequence as one of the strands in the dsDNA. At the end of the process, the ssDNA is wrapped inside the major groove of the dsDNA and the complex is stabilized by a combination of Watson–Crick and Hoogsteen bonds. The *in vivo* process is more complicated, since the strand invasion is aided by proteins, such as RecA or Rad51 [13].

Simulating protein free strand invasion provides a particularly stringent test of the capabilities of our model. First, the major groove needs to be wide enough to accommodate the excluded volume of the invading strand. Second, the directionality of the bonding interactions needs to be strong enough to prevent unphysical bonding to multiple sites. Indeed, we frequently found that the spherical potentials appearing in the two bead model described in Chapter 4 [172, 174] led to a collapsed, globular state. Finally, stabilizing the triple-stranded structure requires Hoogsteen bonds.

To demonstrate that our model has the requisite fidelity to capture strand invasion, we used the single-stranded sequence 5'-ACTCAACCAAGTCATTCTGCGAAT AGTGTATGCGGCGACC-3' and a complementary double-strand. The dsDNA was relaxed into the B-form in the absence of the single-strand. We then initialized the ssDNA as a comb. If we define a polar coordinate system at the complementary 5'-end of the dsDNA with $\theta = 90°$ pointing along the backbone of the dsDNA, then the 3'-end of the ssDNA is located initially at a distance of 1.5 nm and an angle of $150°$ relative to the 5'-end of the dsDNA. All beads on the linear ssDNA strand are initially in the plane defined by (i) the line connecting the 5' and 3' beads of the

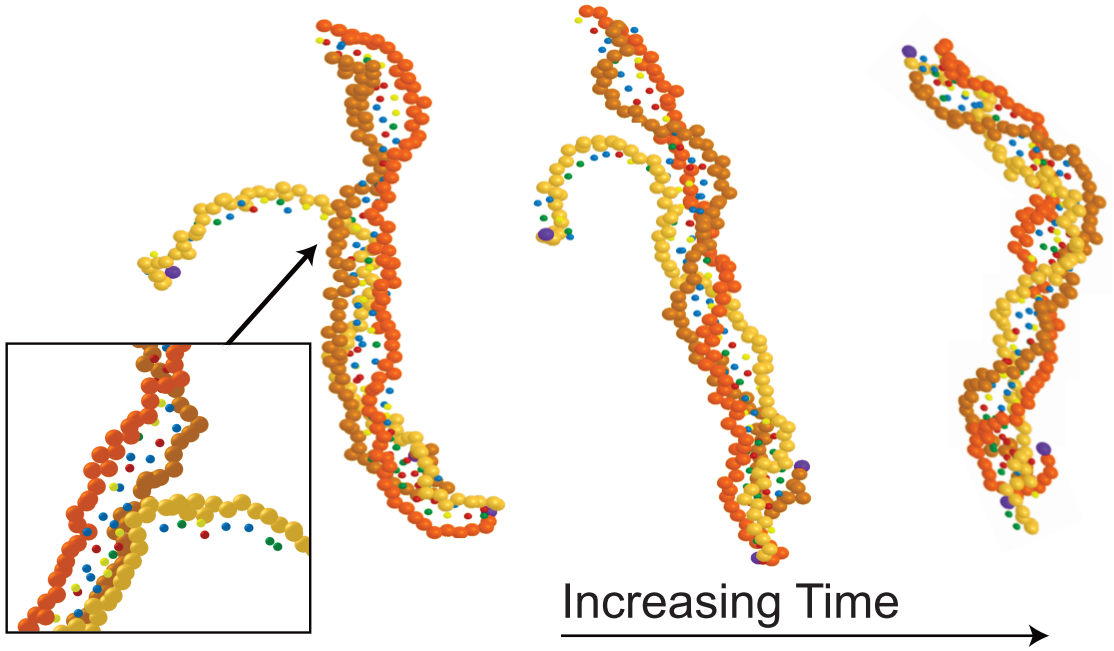

Figure 6.6: Simulation snapshots during the formation of a triple-stranded DNA. The backbone of the invading strand has the lightest color. For clarity, the base beads are represented as small spheres. The magnified inset depicts the indicated region of the chain, viewed from behind, where a dsDNA bubble is forming and the third, ssDNA, is being incorporated into the complex. At the final stage, the triplex is completely formed.

complementary strand of the dsDNA and (ii) the aforementioned line in the polar coordinate system. This initial condition promotes strand invasion in a reasonable amount of simulation time while allowing thermal motion to stack the ssDNA prior to invasion. The simulation was conducted at a temperature of 285 K.

Figure 6.6 shows several snapshots during the course of the simulation after the ssDNA has begun to disrupt the helical structure of the dsDNA. The ssDNA initially diffused towards the dsDNA. When the two DNAs came into close contact, the presence of the ssDNA opened a bubble in the dsDNA. This bubble allowed for rearrangement of the hydrogen bonding to minimize the energy of the combined Watson–Crick and Hoogsteen interactions between the bases on the three strands. The inset in Figure 6.6 highlights the bubble and the invading strand. As this region of the triplex stabilized, the backbone of the invading strand inserted into the major groove and the bubble propagated along the dsDNA. When the bubble reached the opposite side of the dsDNA, it closed to produce the final, triple-stranded state.

Both H-DNA and triplex formation are thought to be protein mediated structures *in vivo*. Although our model does not include proteins or their actions, by biasing the initialization of the simulation we are able to begin to examine the pathways and that may be important in the formation of these structures. Both of these triplex structures are little studied by either a simulation or an experimental approach, we hope that our model with its necessary Hoogsteen base-base interactions, and these simulations can provide a starting point for the further investigation of these phenomena.

6.5 Conclusions

Nucleic acids in *in vivo* systems continue to be studied, and as soon as we begin to think we understand how some genetic process works, we find that there is another layer beneath, that is more intricate and complex. As an example, transcription has proven to be a much more complicated process than was once thought. When Crick [80] first decoded the process by which DNA was translated into amino acids in 1968 it was assumed that we would soon completely understand the ways in which proteins are made in a cell. However, since then we have learned of the many and redundant regulatory pathways governed by signaling molecules and containing switches in our genomes that control how DNA can be translated into proteins. The full understanding in how these processes work in harmony is still a mystery at some levels. As Lewis Thomas [278] declared in the quotation at the beginning of this chapter, we cannot yet "claim genuine understanding" for "the more we learn about living creatures ... the stranger life becomes". Therefore it is the "strange" parts that we must continue to explore and study; we present this chapter on unusual and complex structures as a first step towards examining these features of nucleic acids whose utility and presence in living environments is still unknown.

The three bead per nucleotide simulation model is one such way of beginning to investigate these structures due to: (i) the presence of non-canonical hydrogen bonding possibilities, (ii) the ability to capture and smoothly transition between multiple global forms of DNA, and (iii) the ability to reach long simulation times. We have highlighted the importance of including Hoogsteen bonds in coarse-grained models of DNA as they are fundamental in capturing many complex structures. In this study we explored the role of Hoogsteen bonds in a relatively simple model of DNA. We expect that it will be straightforward to augment other coarse-grained models with non-Watson–Crick bonds if one wants to study more detailed interactions, such as the

role of solvation or ionic strength [160–163, 167, 238]. We expect that coarse-grained models incorporating Hoogsteen bonds will be useful in a number of scenarios beyond hybridization of dsDNA or ssDNA hairpins. As the first example in this chapter, we investigated the folding of an ssDNA aptamer that possesses a G-quartet. There are numerous other aptamers whose secondary and tertiary structure should be affected by Hoogsteen bonding and are thus amenable to simulation using our method. In another set of examples, we showed how the model could capture H-DNA and the dynamics of strand invasion leading to triplex formation. While the *in vivo* situation is much more complicated due to the presence of DNA binding proteins, the model presented here is the first step towards a sequence-specific, coarse-grained model of DNA regulation and repair.

In addition, the inclusion of anisotropic base-base potentials allows not only for the smooth transition between single- and double-stranded structures to be examined but also other types of dsDNAs such as left-handed, P-form, and S-form dsDNAs. With minimum parameter refinement we also believe that the model could be tuned to capture double-stranded RNA characteristics. Finally, although the model presented here does not include the complex and delicately balanced free energy terms found in all atom systems, this limitation should be balanced against the model's ability to reach long times. The structures described in this chapter involve a relatively large amount of nucleotides (over a hundred for H-DNA), on multiple strands, undergoing global rearrangements; by using a coarse-grained Brownian dynamic model we are able to simulate the number of beads necessary and at sufficiently long time scales to allow such complex structures to form.

However, it is just this coarseness in the model that limits the in depth understanding that can be gained. In Chapter 7 we are going to take a coarse-grained model and use it to find the global structure of a DNA and RNA complex that cannot be experimentally crystallized, the 10-23 DNAzyme. With the global structure we will implement a process to map the coarse-grained model to an atomistic model and then use a molecular dynamics approach to further understand the exact mechanism involved in the RNA substrate cleavage reaction. With such techniques we can match the strengths of the three bead model presented with other approaches to gain insight into the intricate and hereto impenetrable world of nucleic acids.

7

Unraveling the Mechanism of the 10-23 DNAzyme

Few scientists acquainted with the chemistry of biological systems at the molecular level can avoid being inspired. Evolution has produced chemical compounds exquisitely organized to accomplish the most complicated and delicate of tasks. Many organic chemists viewing crystal structures of enzyme systems or nucleic acids and knowing the marvels of specificity of the immune system must dream of designing and synthesizing simpler organic compounds that imitate working features of these naturally occurring compounds.

Donald J. Cram, *The Design of Molecular Hosts, Guests, and their Complexes*,
Nobel Lecture December 8, 1987 [281]

DNA has moved well beyond being simply the reservoir of genomic information. As has been seen in the previous chapters, DNA's ability to form complex and multifaceted structures facilitates the molecule's ability to function in unique and non-canonical ways. The majority of the structures heretofore examined occur naturally and are "compounds [that are] exquisitely organized to accomplish the most complicate and delicate of tasks" [281]. However, DNA is now being designed to "imitate [the] working features of these naturally occurring compounds" [281] for the directed assembly of colloidal crystals [282] and other complex nanoscale objects [283], acting as a replacement for antibodies [284], and even working as an enzyme (DNAzymes) [52]. Connecting the structure of these emerging technologies to their function is rarely straightforward, especially if the DNA has some engineered biological activity. In particular, crystallization of the active form is not always possible [285, 286] be-

cause the system is out of equilibrium or in a metastable state. When experimental methods have failed to fully unravel these complex structures in action, computational approaches have been used to gain fundamental understanding. In particular, recent advances in coarse grained DNA models [160, 166, 176] have opened up new, computational avenues for studying DNA-based technologies.

7.1 Introduction to DNAzymes

DNAzymes, also known as deoxyribozymes or DNA enzymes, are single-stranded DNA molecules with catalytic capabilities. Since the first DNAzyme was discovered over 15 years ago [52], hundreds of DNA sequences have been isolated that facilitate chemical transformations of biological and non-biological importance [106, 287]. Although deoxyribozymes have been excluded from natural evolution, the ability of DNAzymes to be specifically designed and created in the laboratory has opened an entirely new field of biological molecular function.

DNAzymes are generated *de novo* by *in vitro* selection. The discovery of DNA's catalytic function was made possible due to the development of the *in vitro* selection technique [279]. *In vitro* selection is a simple yet powerful combinatorial approach that allows simultaneous screening of 10^{13} to 10^{16} different DNA, RNA, or modified nucleic acids with an ability to bind a target of interest [106]. The mechanisms ascribed to natural selection evolution are also applied to *in vitro* selection but on a much faster time scale: variation, selection, and replication. Hundreds of deoxyribozymes have been found that catalyze more than a dozen different types of chemical reactions, most of which are involved in nucleic acid modifications such as cleavage and ligation of DNA and RNA strands [96, 106].

One such deoxyribozyme, named the 10-23 DNAzyme, was found to cleave an all RNA substrate in a biological cofactor range (i.e., Mg^{2+}) [55, 56]; it was named from its origin as the 23^{rd} clone of the 10^{th} cycle of *in vitro* selection. The core of the 10-23 DNAzyme is composed of only 15 nucleotides, flanked on each side by a substrate binding arm of 7 to 10 nucleotides that bind to the RNA target via Watson–Crick base pairing, as seen in Figure 2.6. This simple structure permits easy alteration of the substrate specificity to generate precise cleavage agents for almost any RNA molecule at a purine–pyrimidine junction. The ability of the 10-23 DNAzyme to cleave at the AU and GU sites allows the AUG start codon of any gene to be used

as a target. Under optimized reaction conditions, the 10-23 deoxyribozyme with a catalytic rate constant (k_{cat}) is greater than 10 min^{-1}. The catalytic efficiency k_{cat}/K_M is 10^9 $M^{-1}min^{-1}$, a value that is limited by the rate of RNA–DNA duplex formation. Under simulated physiological conditions, the 10-23 DNAzyme exhibits k_{cat} and K_M values comparable to those of the natural ribozymes and even protein endoribonucleases [56]. The chemical stability, high catalytic proficiency, mismatch discrimination, and the ease of synthesis of DNA have made the 10-23 DNAzyme an attractive alternative to ribozymes for site-specific cleavage of biological RNA targets [50].

The primary impetus for basic research into deoxyribozymes is their downstream practical applications. In particular, uses of RNA-cleaving DNAzymes such as the 10-23 DNAzyme range from *in vitro* chemical and biochemical experiments to *in vivo* applications as therapeutics [288]. The 10-23 DNAzyme, the most commonly studied deoxyribozyme [287], has found *in vitro* use as a molecular probe to assess site-specific RNA modifications (such as pseudouridylation, 2'- O−methylation, and m^5C formation [289–292]), an integral part of the 'DzyNA-PCR' procedure for real time detection of specific DNA sequences from biologically derived samples [106, 287, 290, 293], a DNA nanomotor [294–297], and a diagnostic test for genetic diseases like cystic fibrosis [298, 299]. The 10-23 motif has also been used in sensors for pH [300], fluorescence signaling [300], metal cofactors [96, 301–304], chemical substances, such as cocaine [92], and in DNA switches [106, 300, 305–308].

The utility of RNA-cleaving deoxyribozymes extends beyond the *in vitro* applications described above. In particular, DNA enzymes that cleave RNA maintain their activity in cellular settings and constitute a viable therapeutic strategy [106, 107, 309] for viral diseases such as HIV [51, 310–317], hepatitis [318–321], influenza [322], SARS [323], RSV [324], HPV [325], human rhinovirus [326], and Epstein-Barr [327]. They have also been used in the treatment of bacterial infections, such as tuberculosis, and antibiotic resistant strains of common infectious agents like MRSA [328–331]. Nearly all of the *in vivo* efforts have utilized the 10-23 structure also known as the type II motif.

Therapeutic targets for type II motif deoxyribozymes are not limited to viral and bacterial RNAs. They have come to be used to cleave a variety of cancer related gene products, with success both *in vitro* and in cell culture [332]. The list of studied cancers include: breast [327, 333], colon [327], lung [327], prostate [327], pancreas [334], leukemia [335–339], lymphoma [327], bone [340], neoplastic diseases [341], and general

tumor function (angiogensis, apoptosis, Egr1 etc.) [109, 117, 333, 342–359]. Finally they have also been used in the treatment of genetic diseases and cellular repair and regeneration. Both Huntington's disease and Saethre-Chotzen syndrome [109, 360, 361], have shown early progress utilizing a 10-23 derivative deoxyribozyme. Other targets for cellular processes have included the aging process (telomerase reverse transcriptase) [362], allergic asthma [363], wound healing (c-Jun protein) [350, 351, 364], central nervous system regeneration (axon re-growth and lesion repair) [365, 366], pain management [367], cardiovascular disease [347], and kidney failure [368].

Despite the wide spread applications of the 10-23 DNAzyme very little is known about its tertiary architecture from the structural biology viewpoint [369]. There only has been one attempt to crystallize a deoxyribozyme, the 10-23 DNAzyme, however the structure isolated was not the catalytically active DNA–RNA complex (instead it was a 2:2 DNA–RNA complex resembling a Holliday junction) [285, 286]. Therefore, neither tertiary nor mechanistic information could be inferred from the crystal structure study. Although there are biochemical data exploring various facets of RNA cleavage by the 10-23 DNAzyme [55, 106, 107, 115, 287, 308, 346, 369–374], a coherent picture of the structural basis for this activity has been elusive by experimental means.

Another route to gain further information about the tertiary structures and mechanism of complex molecules is through simulation studies. However, because the 10-23 DNAzyme lacks the crystal structure that is commonly used for initial, atomistic scale studies [375–378], this manner of investigation has been limited. In this chapter we present a new method for the examination of the 10-23 DNAzyme tertiary structure and dynamics utilizing a multi-scale simulation approach. Starting with a relatively simple coarse grained picture [172–174], we develop an atomistically detailed model for the 10-23 DNAzyme (as an example of a DNA based nanotechnology) using a multi-scale simulation approach that circumvents the need for a crystal structure [285, 286] as the initial condition for the simulation. Our approach to studying the 10-23 DNAzyme is generic and it is easily adapted to other DNA nanotechnologies that lack known crystal structures.

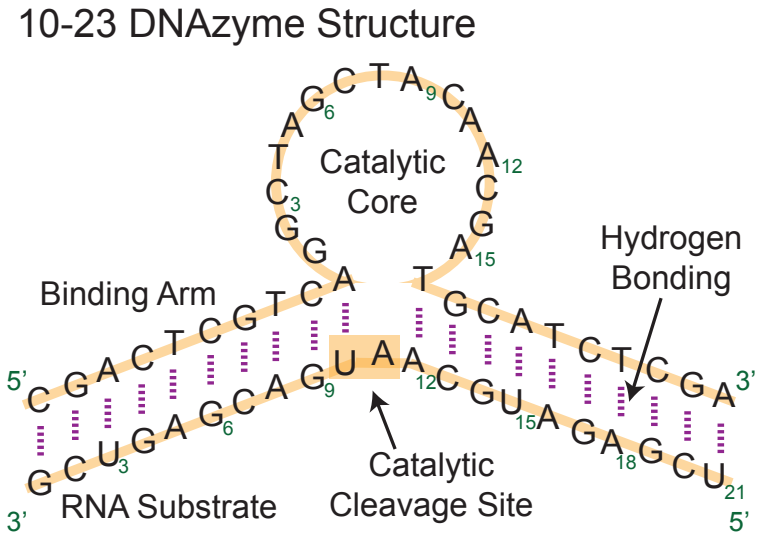

Figure 7.1: A schematic representation of the 10-23 DNAzyme (35 nucleotides) bound via Watson–Crick hydrogen bonds to its complementary RNA substrate (21 nucleotides) via two binding arms (10 nucleotides each). The catalytic cleavage site occurs on the 3' arm of the RNA substrate between the unpaired adenine (A_{11}) and the adjacent uracil (U_{10}). The catalytic core is highly conserved in mutation studies.

7.2 Multi-Scale Modeling Approach

To determine the structural basis for this DNAzyme's activity, we started with a two bead, coarse grained representation of each nucleotide [172–174]. The system consisted of a deoxyribozyme sequence 5'- CGACTCGTCA**GGCTAGCTACAACGA** TGCATCTCGA-3', where the catalytic core sequence is in bold text, and the binding substrate 3'- GCUGAGCAGU**A**ACGUAGAGCU-5', where the bold base is the unpaired adenine and shown in Figure 7.1. The cleavage reaction occurs between the U and the **A** on the 3' arm of the RNA substrate [55, 56]. After relaxing the structure by Brownian dynamics, we used a series of model refinements and relaxation steps to produce an atomistically detailed, fully solvated model in an electrolyte containing generic divalent counter-ions. We then simulated this system for 1.2 μs by molecular dynamics using the CHARMm27 force field [378]. This multi-scale approach reveals that the structure of the DNAzyme–RNA complex differs substantially from the simple schematic of Figure 2.6 that is often adopted in the biochemical literature.

7.2.1 Coarse-Grained Representation

Our starting point is a two bead representation of each nucleotide with one bead embodying the sugar and phosphate groups on the backbone and one bead representing the base identity, as described and in Chapter 4 and seen in Step 1 of Figure 7.4. In this model we do not distinguish between DNA and RNA in the model; we treat uracil and thymine as identical. We performed six independent, 4 second long BD simulations at 300 K, where one simulation time step is approximately 10 picoseconds and frame positions were recorded every 1000 BD time steps. The temperature and time step are approximate values [174].

The initial conditions for the simulations are rotations of the "cartoon" secondary structure used previously [173] and proposed by Santoro and Joyce [56]. The structure of the 10-23 DNAzyme and the RNA substrate are aligned so that the recognition binding arms are aligned to allow immediate Watson–Crick hydrogen bonding between the bases with the unpaired adenine base at the center of the substrate aligned with the catalytic loop of the deoxyribozyme. Our nomenclature calls this center substrate base the unpaired A or unpaired base for the remainder of our discussion. The resulting simulation structure is qualitatively independent of the initialization chosen.

We present a summary of the results from Kenward and Dorfman [173]. Notably, the DNAzyme-RNA substrate complex was seen to have a stable bent state over long simulation time. The binding to the substrate brings the ends of the catalytic core into close proximity. This reduction in the end-to-end distance is the result of the recognition arms bending away from the catalytic core which may induce sufficient tension to unstack the unpaired base. However, due to the level of coarse graining the detailed nature of the unstacked A was beyond the resolution of the two bead model and such conclusions could not be drawn. In addition, although it is understood that salt ions and metallic co-factors are essential to the functioning of many deoxyribozymes, including the 10-23 DNAzyme, the inclusion of ions in the study was beyond the model and computational resources then available.

In our subsequent coarse grained study of the 10-23 DNAzyme-RNA substrate, we examined the initial configuration and the relaxed bent state that it forms. We systematically investigated the multiple three-dimensional projections of the initial, two-dimensional, cartoon proposed by Santoro and Joyce [56] to understand any bias that the initialization configuration may cause. We next implemented a long time

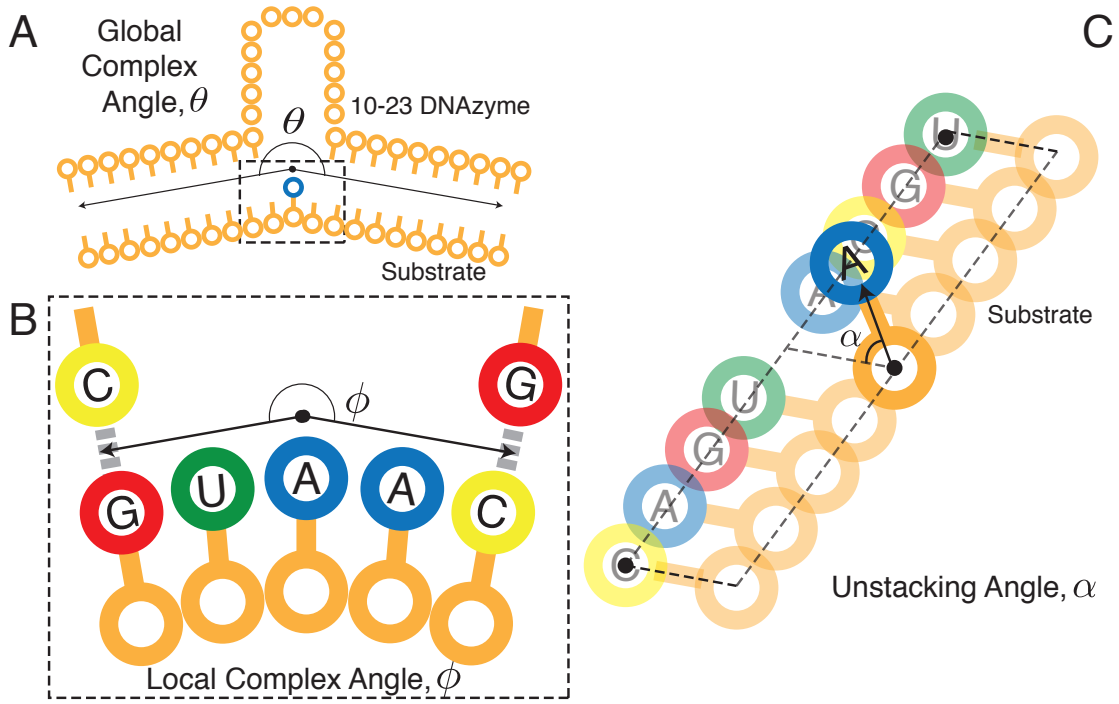

Figure 7.2: A schematic representation of the 10-23 DNAzyme-RNA substrate complex. The entire complex is shown in (A) with the global complex angle θ measured from the unpaired adenine to the midpoint of the hydrogen bond at each terminus. (B) The inset from the global structure is shown to describe the local complex angle ϕ. The local complex angle describes the angle formed from the unpaired A to the midpoint of the hydrogen bonds that are two bases away (G_9, C_{-1} and C_{13}, G_{17} on the RNA substrate and DNAzyme, respectively). For clarity, nucleotides not involved in the local angle calculation are omitted. (C) The unstacking angle α measures the position of the unpaired A base relative to its stacked neighbor beads. This angle is formed by drawing a plane (dashed lines) that contains the backbone beads of the unpaired A on the RNA substrate and the base beads four nucleotides away on each side (C_7 and U_{15}). The quantity α is the absolute value of the angle between the plane and a vector, shown with a solid (black) arrow, drawn from the backbone to the base bead of the unpaired A.

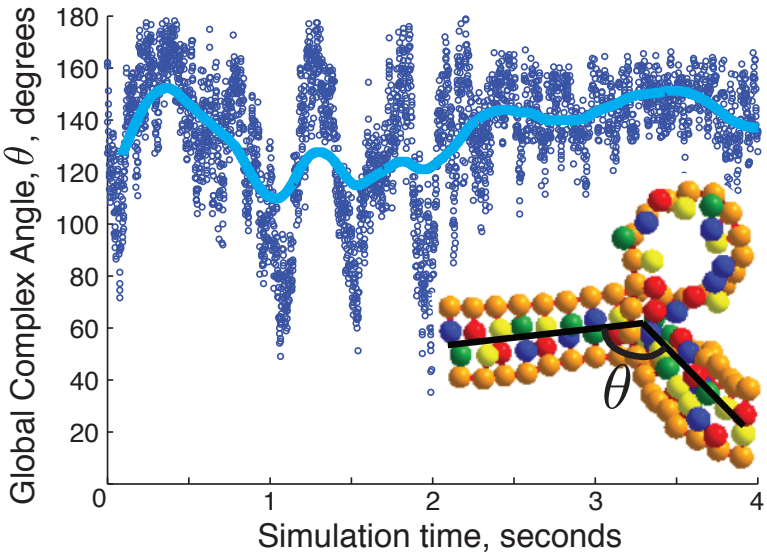

Figure 7.3: Plot of the global complex angle of the initialization projection depicted in Step 1 of Figure 7.4. The global complex angle, θ, formed between the unpaired A on the RNA substrate and the hydrogen bonds between the terminal ends of the recognition arms and the substrate. The light (blue) line represents the smoothed average of the data.

Brownian dynamics simulation parameterized such that it mimics biologically relevant solvation and temperature [99, 174]. After sufficient time, we found a stable tertiary configuration of the complex, as seen in Figure 7.3. The bead positions at 3.5 seconds serve as the initialization coordinates for the subsequent atomic model development and is depicted as Step 2 in Figure 7.4.

The interplay of the tension on the unpaired A by the bent substrate configuration, the stacking interactions, and the overall steric interactions lead to fluctuations that ultimately produce the configuration snapshot in Figure 7.3. The complex is strongly bent at the cleavage site; starting with an angle of 180° as the initial condition for the BD simulations [173], we ultimately arrived at an average angle of 143.9° ± 9.7°. In comparison with the previous work of Kenward and Dorfman [173], we report an averaged end-to-end distance of 5.525 ± 0.506 nm after converting from the dimensionless simulation units. The compares favorably with their 5.4 ± 0.4 nm reported and is close to the vale of 6.00 nm obtained during bulk FRET measurements [346].

We also examined the unpaired A relative to its stacked neighbor beads. In order to define this angle, illustrated in Figure 7.2 C, we first draw a plane between the following three points (measured from the center of mass of the atoms in each bead

grouping): the backbone bead of the unpaired A and the base beads four nucleotides away on each side (C_7 and U_{15}). We chose these nucleotides so that any local buckling at the catalytic site does not affect the measurement of the angle. The quantity α is the absolute value of the angle between the plane previously formed and a vector drawn from the backbone to the base bead of the unpaired A. Similar to the conclusions of Kenward and Dorfman [173], we see a partial unstacking of the unpaired A on the RNA substrate. However, at this degree of coarse graining and model choice it is not possible to further resolve details concerning the cleavage site. Therefore, while the coarse grained model is able to provide insight on the tertiary structure, it alone cannot provide enough detail to understand the complete system structure and function; we now begin the transition towards a full atomistic model and molecular dynamics study of the system.

7.2.2 Mapping Between Models

To map the BD coarse grained model, we picked a random configuration near the end of one simulation. The characteristic length scale of the coarse grained model is the base-backbone bond length, which we defined as the distance between the center of mass of the atoms in the DNA backbone (the deoxyribose and phosphate group for DNA) and the center of mass of the atoms in a guanine base in an atomistically detailed model. To create the atomistic model, we used the CHARMm27 force field [378] for the bases. In the first mapping step, we used united bases to represent each nucleotide. We placed the center of mass of the sugar/phosphate atoms of the united base at the location of the backbone bead from the coarse grained simulation. We then rotated the vector between the center of mass of the sugar/phosphate atoms and the center of mass of the base atoms in the united base model so that this vector was parallel to the vector between the coarse grained backbone bead and the coarse grained base bead, Steps 3 to 6 in Figure 7.4. Initially, each of these united bases was bonded together along the backbone by a harmonic constraining force of 100 kcal/mol between the centers of mass of neighboring sugar/phosphate units. Note that the united bases are electrically neutral; there are no charges in the model until after the solvation step.

The united bases already contain a number of structural features that are not captured by the coarse grained model. For example, although the coarse grained model includes measures of rise and mean rotation per base pair, it does not include propeller twist,

roll, or tilt of the base groups. We thus needed to make three sequential steps of relaxation: (i) the distance between each base and its corresponding location on the backbone, (ii) the distance between nucleotides along the backbone, and (iii) the relative orientation of adjacent bases.

To perform these relaxations, we first added a harmonic constraining force of 100 kcal/mol between the center of mass of each base and the corresponding center of mass of its sugar/phosphate unit. (Note that we already have a harmonic containing force of 100 kcal/mol between the centers of mass between neighboring sugar/phosphate units on the backbone to mimic the phosphodiester bond.) We then minimized the energy with a series of small minimizations (1000 steps) with both the steepest descent (SD) algorithm and the adapted basis Newton–Raphson optimizer (ABNR) algorithm. We performed standard checks and monitored the RMS gradient and energy change at each step to ensure that minimization has been reached.

A total of twelve serial minimizations were performed, with every other minimization conducted by the SD or ABNR algorithm. For each subsequent minimization, the harmonic containing force between each base and its sugar/phosphate unit was halved. Once we had relaxed the base-backbone distances, we removed the constraining force between the sugar/phosphate units and the bases and then performed the same minimization scheme while reducing the harmonic constraining force between the centers of mass of neighboring sugar/phosphate units. At the end of this second relaxation series, the harmonic constraints along the backbone were replaced with phosphodiester bonds. In the final set of relaxations, we used an initial harmonic constraining force of 100 kcal/mol on the relative orientation of neighboring bases in the 5' to 3' direction and followed the same relaxation procedure. At the conclusion of the relaxation steps these constraints were removed.

The system was solvated, in Steps 7 and 8 of Figure 7.4, using TIP3 water with explicit counter ions, both of which were modeled using the CHARMm27 force field. In both the salt and metal ion insertions, if the added ions overlapped any other molecule (water, DNA, or RNA), the ion was removed and the water molecule was replaced. Before solvation, we converted from united bases to an all-atom model, which has the hydrogens. We then solvated this nucleic acid complex by first creating and relaxing a 6 Å per side cubic box of TIP3 water and then filling the simulation space (90 Å per side cube) by periodic tiling the smaller relaxed water cube. The complex was located at center of the box and the water molecules which overlap the nucleic acids were removed, leaving a system with 22,764 water molecules. The

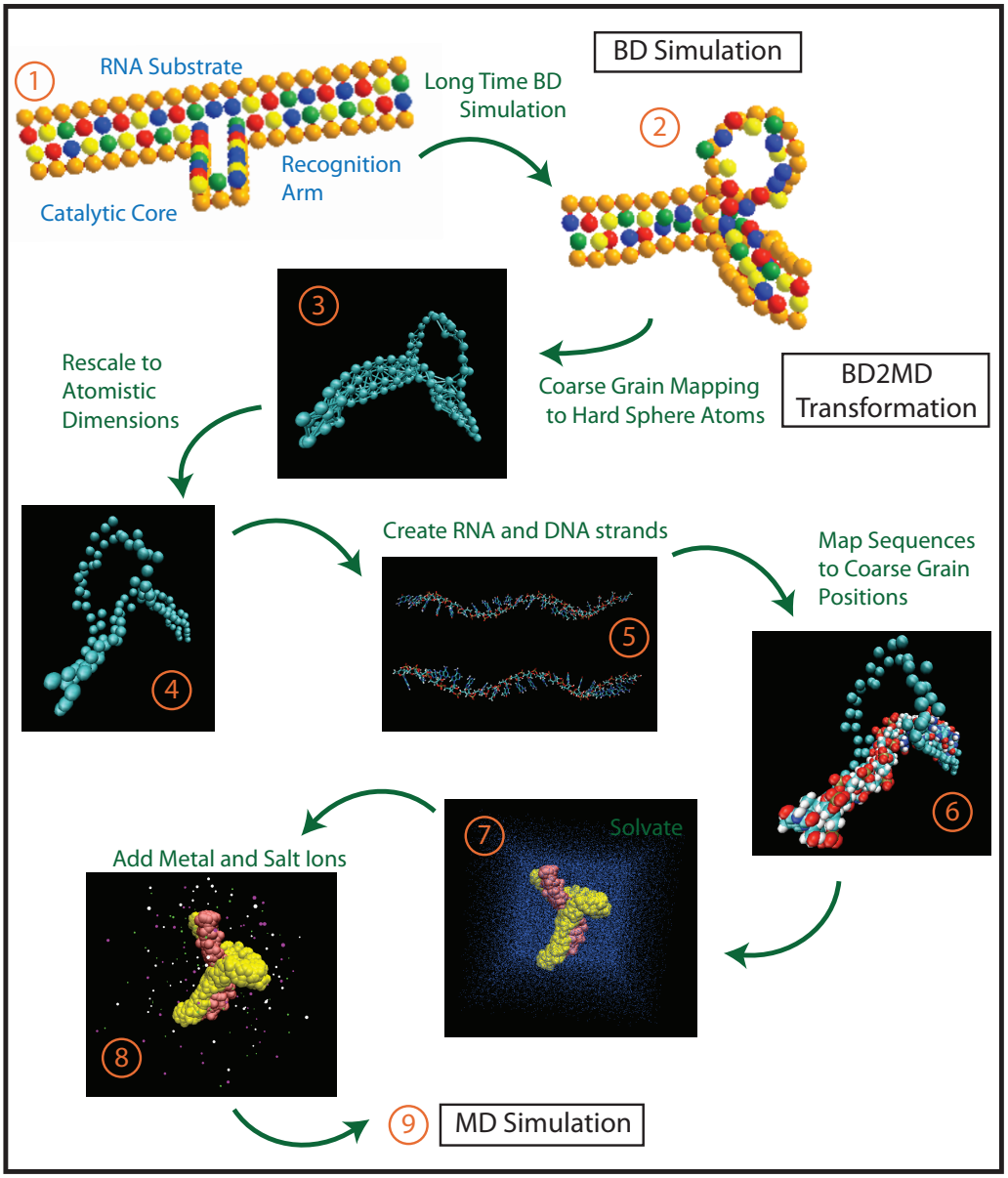

Figure 7.4: The tertiary structure of the 10-23 DNAzyme was explored serially with a multi-scale approach. Step 1: The projections of the proposed secondary structure were initialized into the coarse-grained Brownian dynamics simulation and a long time (order of seconds) simulation was conducted. Step 2: The relaxed and globally stable coarse grained complex positions were recorded. Step 3: The coarse grained bead positions were mapped into an atomistic nucleic acid model and represented by hard sphere atoms. Step 4: The coarse grained positions were rescaled to realistic atomistic lengths. Step 5: Two nucleic acid strands were created. Step 6: Each nucleic acid was mapped onto its corresponding coarse grain position. Step 7: Water atoms were added. Step 8: Metal and salt ions were added. Step 9: Full molecular dynamics simulations was conducted (on the order of microseconds). Between every transformation step the system was relaxed.

solvated system was relaxed through a series of twelve small energy minimizations (1000 steps), alternating between the SD algorithm and the ABNR algorithm. We performed standard checks and monitored the energy change at each step to ensure that the minimization has been reached.

After solvation and relaxation, the hydrogens on the backbones of the DNA and RNA were stripped (except for the terminal ends). To maintain electroneutrality, we removed 54 random water molecules and replaced them with sodium ions. To create a 150 nM buffer, which is the typical experimental system [55, 56], we then removed an additional 124 random water molecules and replaced them with 62 sodium ions and 62 chlorine ions. We relaxed the buffered system through a series of twelve small energy minimizations (1000 steps), alternating between the SD and ABNR algorithms. We performed standard checks and monitored the energy change at each step to ensure that minimization has been reached.

Although the divalent metal ions in experiments are typically found in the millimolar concentrations (physiological concentration is less than 5 mM) [55, 56], the corresponding number of M^{2+} ions in the simulation box (on the order of less than 5) is so low that the diffusion time for a given metal ion to interact with the complex is prohibitively long for the MD study. To make the computation feasible, we added 43 generic divalent metal ions, M^{2+}, by randomly removing water molecules and replacing them with ions. To keep the system electroneutral, we randomly removed 86 additional water molecules and replaced them with chlorine atoms. Since the CHARMm27 force fields are not fully established for specific metals such as copper(II) or magnesium, we used a generic divalent metal that has the correct divalent charge and a radius of approximately the size of divalent gold. After competing the ion insertion, we used the ABNR algorithm to relax the system until it reached the default settings.

7.2.3 Atomistic Model

After the final energy minimization, we simulated the dynamics of the system using the NAMD software package. We used a step size of two femtoseconds, and a tool simulation time of 1.227 microseconds. In order to determine whether the transformation process has preserved the tertiary structure in the coarser BD simulation, we examined the local complex angle, ϕ, defined in Figure 7.2 B. We choose to compare the local complex angle between the two systems since both systems can fluctuate at

the time scale simulated, there are many stresses in this area, and the structure at the cleavage site is important to the overall function of the deoxyribozyme complex. The local complex angle remains relatively stable at an average value of $140.3° \pm 0.7°$ as compared to the $141.6° \pm 2.7°$ for the Brownian and molecular dynamics studies, respectively. The overall similarity of the local complex angle measure from the BD and MD studies show that the transformation process has preserved the long time bent tertiary structure found in the BD simulation.

7.3 10-23 DNAzyme - RNA Substrate System Dynamics

While the BD simulations of the coarse grained model suggest that the unpaired adenine on the RNA substrate may be unstacked [173], Figure 7.5 shows the unstacking angle of $\alpha = 43.2° \pm 3.7°$. Figure 7.6 reveals that there are in fact two stable states for the complex: one state where the unpaired A is stacked and another state where the unstacking is substantial. The remainder of the bases in the RNA substrate remain stacked with the exception of the ends of the complex, which exhibit the typical fluctuations caused by fraying.

Unstacking of a base is energetically unfavorable since the π orbitals no longer overlap. The concomitant increase in free energy of approximately 6.1 kcal/mol at the top of the barrier in Figure 7.6 is eventually offset by new interactions between the hydrogens on the unpaired adenine base and reveal sites on the nearby backbone and base atoms. These hydrogen bonds are facilitated by twisting the planar ring of the adenine, as illustrated in Figure 7.5 C. Due to steric interactions, the twisting only begins after the base comes unstacked. The angle, while intuitively straightforward, is quite difficult to illustrate. For a given base, we first define a plane using the 1-carbon, 3-carbon, and 5-carbon of the base. (This removes any difference between the definition for a pyrimidine and a purine base.) We then define a normal vector for this plane pointing from the 5' to the 3' direction along the sequence. The twisting angle, β, is defined as the angle formed between the normal vectors for adjacent bases in the 5' to 3' direction. The value $\beta = 0$ corresponds to parallel normal vectors. This unstacked and twisted state is quite stable, with a free energy approximately 1.8 kcal/mol lower than the original stacked state in Figure 7.6. The relatively low barrier and small energy difference between the stacked and unstacked state are in line with a previous

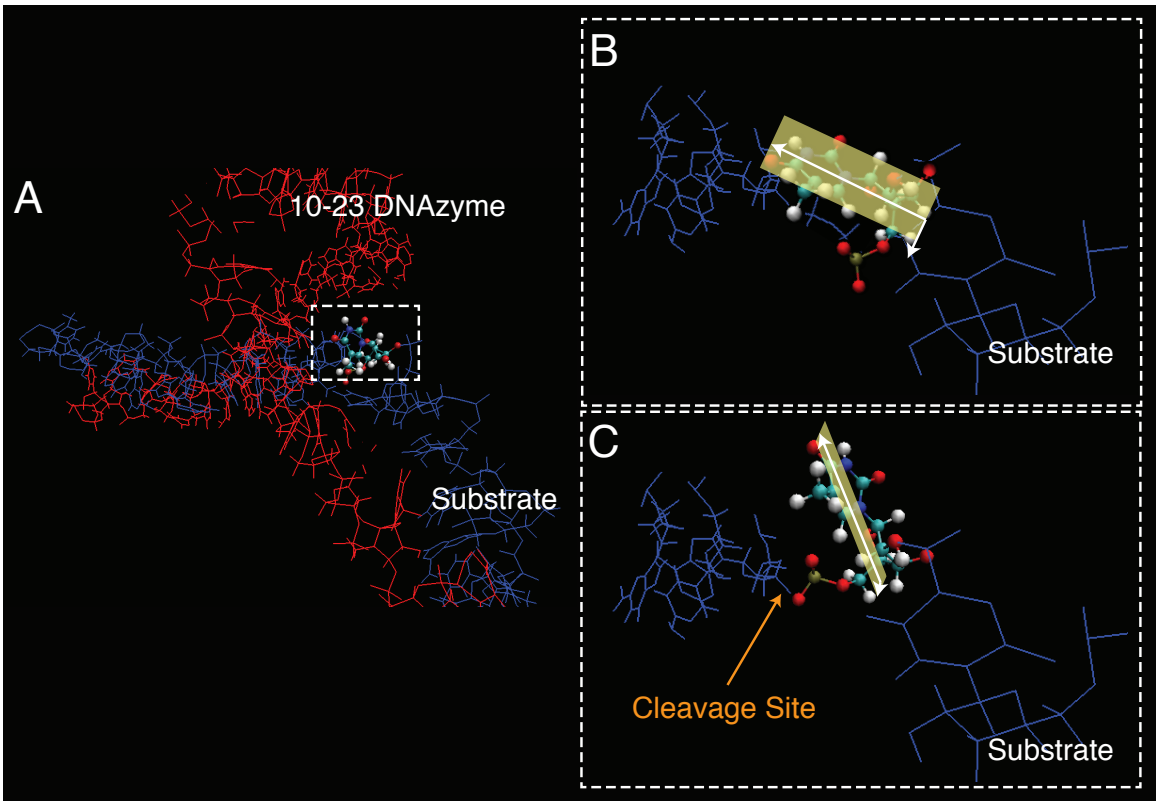

Figure 7.5: The initial configuration (A) of the 10-23 DNAzyme (red) and RNA substrate (blue) complex obtained after the mapping procedure. The unpaired A base is shown with a more detailed ("ball-rod" type) representation while the other RNA residues have only a simpler ("line type") depiction; (B) is the zoom of the boxed area of (A) showing the initial stacking of the unpaired A. The complex is oriented such that the catalytic loop of the 10-23 DNAzyme (not shown) points out of the page. (C) Snapshot of the configuration after 1.2 μs simulation. On average, the unpaired A has (i) rotated out of the plane of the other stacked nucleotides by $\alpha = 43.2° \pm 3.7°$ and (ii) twisted an average of $\beta = 80.8° \pm 4.1°$. These angular averages correspond to the last 50 ns of the simulation where the DNAzyme is in the unstacked and twisted state (see Figure 7.6).

results for the energetics of base flipping in a double helix [379].

With one exception [380], divalent metals are essential for the activity of RNA cleaving DNAzymes. It is thus reasonable that the DNAzyme, when bound to its RNA substrate, creates regions of electronegativity that attract these divalent metals near the cleavage site. Figure 7.7 shows the areas of highest negative charge density for the structure in Figure 7.5 C. We can readily identify three areas that constitute binding pockets for the divalent ions: (A) is near the A_{11} and A_{12} bases of the catalytic core, (B) is between the G_6 on the catalytic core and the cleavage site of the RNA substrate, and (C) is on the underside of the RNA substrate. The catalytic loop also kinks inward and towards the cleavage site near the G_6 in the catalytic core; this kink

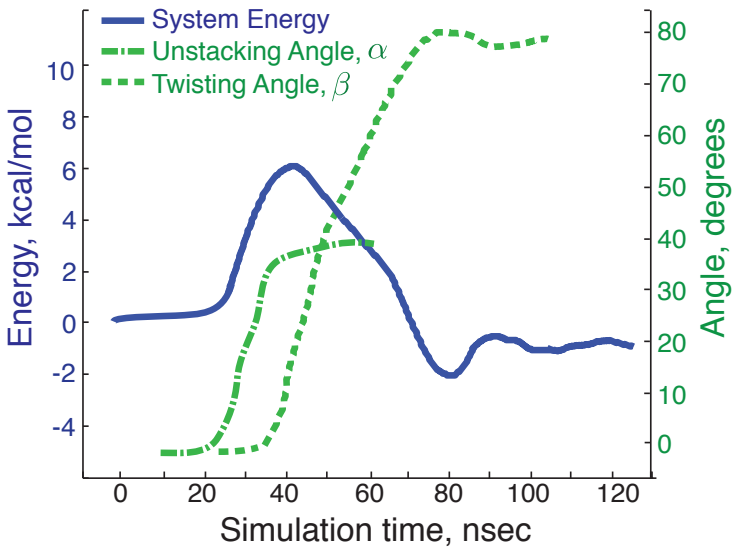

Figure 7.6: The free energy pathway for the unstacking and twisting of the unpaired adenine on the RNA substrate. The free energy of the system is the left (blue) axis. The energy barrier to unstacking is approximately 6.1 kcal/mol and occurs as the unpaired adenine begins to rotate out of a stacked configuration with the other RNA substrate bases. As the nitrogenous base begins to twist the fee energy falls to an eventual plateau corresponding to a 1.8 kcal/mol decrease compared to the stacked state.

may aid in forming the ion trap found between the catalytic loop and the cleavage site.

7.4 Activity Mechanism

Taken together, our results suggest a plausible model for the activity of the 10-23 DNAzyme. The binding of the DNAzyme to the RNA substrate bends the complex and begins to unstack the unpaired adenine [173]. In the bent configuration, the unpaired A unstacks and twists, exposing the cleavage site. The energetics of the unstacking and twisting are commensurate with thermal energy, so the DNAzyme complex should alternate between the stacked and unstacked state. These structural transitions do not require the presence of the divalent metal ions. However, based on Figure 7.7, it appears that the trapping of the divalent metal in binding pocket C further stabilizes the bent configuration. The divalent metal in binding pocket B is drawn close to the cleavage site, along with the G_6 nucleotide. Based on the structure of the complex, we speculate that these entities are responsible for the cleavage of the bond.

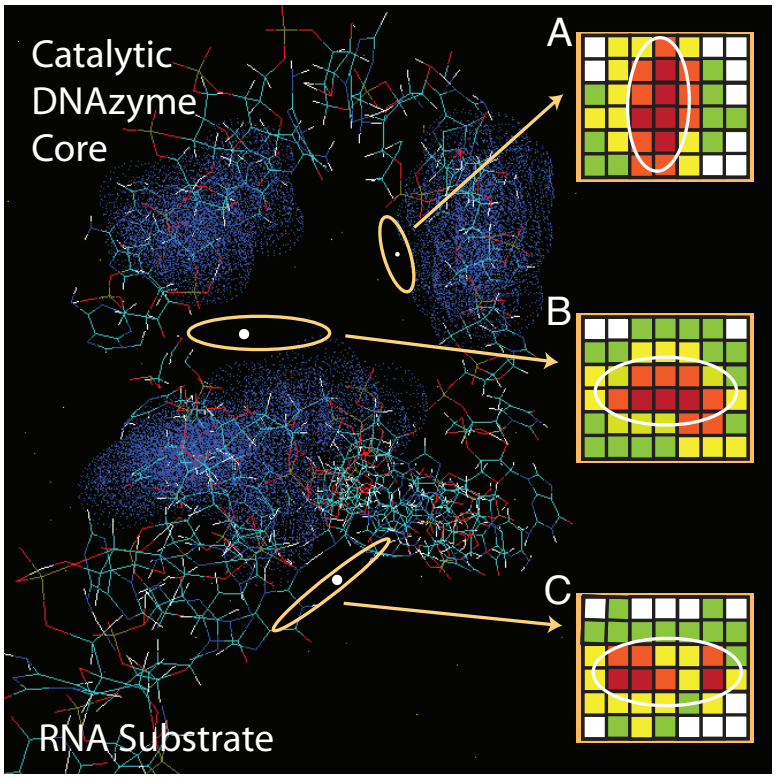

Figure 7.7: A map of the electrostatic distribution near the 10-23 DNAzyme - RNA substrate complex. The (blue) cloud areas depict the highest negative charge density. There are three areas of metal ion trapping. The time averaged diffusion disks, depicting the areas of highest probability for locating a divalent metal ion during the last 50 ns of simulation, are sketched on the main figure and their heat map diagrams are in the insets (A, B, and C). The dark (red) squares are the areas of highest probability and the white squares have the lowest probability. The planes for the histograms were defined to run parallel to the average backbone vector of the molecule nearest the highest metal concentration in that section. The histograms are rescaled to make equal divisions of the sampled space for each ion and are not in absolute distances: in A, B, and C each pixel represents approximately 0.1 nm, 0.2 nm, and 0.2 nm, respectively.

Our model for the active complex explains a number of experimental observations that are otherwise puzzling in the absence of any structural data. First, using an intercalator [371] or adding wobble bonds near the cleavage site [370] increases DNAzyme activity. Both intercalators and wobble bonds destabilize the double-helical structure of the recognition arms and make it easier to form a kink. As the bent structure is critical to providing the space for unstacking and twisting the unpaired adenine (Figure 7.5) and creating the electrostatic distribution or attracting the divalent counter ions (Figure 7.7), it is reasonable that anything that disrupts the double helix without interfering with the base pairing in the recognition arms would increase the DNAzyme activity. Second, as the concentration of Mg^{2+} ions increases, fluorescence resonant energy transfer (FRET) experiments [346] show that the end-to-end distance of the

complex decreases but there is no reaction. As the Mg^{2+} concentration increases further, the end-to-end distance stabilizes and the reaction rate increases. This two-fold action of the divalent metal ion is consistent with the need to (i) trap an ion in binding pocket C to stabilize the bent structure and then (ii) trap an ion in binding pocket B to pull the G_6 towards the cleavage site. Third, mutation studies indicate that each of the four guanines in the DNAzyme catalytic core are highly conserved [115, 308, 372], with G_1 and G_6 being vital to the metal ion binding and is essential for DNAzyme activity [374]. Our model explains the critical importance of G_6, since it is involved in binding pocket B and the nucleotide is pulled towards the catalytic site by the kink.

7.5 Conclusions

In summary, we used a straightforward simulation approach to obtain a detailed picture of a DNA nanotechnology, the 10-23 DNAzyme, without the need for a crystal structure. The 10-23 DNAzyme has been used for hundreds of applications due to its biological compatibility, catalytic efficiency, physiological reagent conditions, and simple cleavage site requirements (needing only adjacent purine and pyrimidine nucleotides) means that not only can it be utilized *in vivo*, but that there are also numerous and potential target sites in any given RNA [55, 56, 114, 298]. Our results provide a clear picture of the activated state and we used the model to rationalize a number of experimental observations [115, 308, 346, 370–372, 374]. While our starting point was a rather elementary nucleic acid mode, the difference between the coarse grained potential and the CHARMm potentials did not lead to any instabilities during the mapping procedure. This gives us confidence that simple coarse grained nucleic acid models such as the one we used here [172–174] could provide the initial configurations needed in molecular dynamic simulations for a wide range of nucleic acid technologies.

8

Conclusions and Future Considerations of Research

> *Science has radically changed the condition of human life on earth. It has expanded our knowledge and our power but not our capacity to use them with wisdom.*
>
> Senator J. William Fulbright, *Old Myths and New Realities, and Other Commentaries*, 1964 [381]

Since the discovery of DNA, researchers have been attempting to decode the detailed sequences, structures, properties, and abilities of this molecule. Long ago, nucleic acids expanded their role from just the transmission, expression, and conservation of genetic information. They are now instrumental throughout many fields in both *in vitro* and *in vivo* applications, as detailed in Chapters 1 to 7. The models developed here will be useful in the continued understanding and engineering of DNA and RNA molecules for nanotechnology, genetic engineering, and therapeutic applications. In this way, as Senator Fulbright described, "science has radically changed the condition of human life". However, it is yet to be seen whether we, as a society, can "use [this knowledge] with wisdom" as we move into the future [381]. As scientific researchers, it is important to understand that technical details are not the only aspects worth considering; ethical, legal, and other policy issues are also fundamental.

In this chapter we seek to unravel some of the societal implications of both the current nucleic acid research and the future implications of continued research. This treatise begins the investigation of complex nucleic acid structure from sequence data. Future research will continue to unfold these multifaceted structures and connect them

with their fundamental functions both in our genomes and as engineered tools in a variety of applications. The combination of more extensive understanding of nucleic acid structures and the ready availability of whole genome sequence data will further transform "human life on earth" [381] as it ushers in a new age of personalized medicine. Begun with the completion of the Human Genome Project and advanced by progress on the $1000 Genome Project, routine whole genome sequencing will revolutionize the entire health and biotechnology fields. However, with such a paradigm changing technology, an entirely new set of issues that require societal deliberation, research scrutiny, and regulatory consideration will arise. A brief overview of some of the considerations concerning these issues are discussed in this chapter.

8.1 Future Research Advancements for Complex Nucleic Acid Systems

As we begin to unravel some of the structural characteristics of complex nucleic acid systems, we find that the *in vitro* and *in vivo* systems are incredibly elaborate. The intricate balance of interactions between the elements of each nucleotide comprising DNA and RNA molecules, the components of the solvent, and any external forces is delicate. Although the model developed in this treatise is coarse, it can provide a first approximation of the features of complex nucleic acid structures. These elaborate arrangements are the fundamental keys to deciphering the most essential cellular processes such as regulation, transcription, and cellular repair. For example, it is believed that hairpin structures, bulges, P-DNA, Z-DNA, H-DNA, and G-quartets all function as genetic flags and switches to control how DNA is translated into proteins. In addition, triple-stranded structures are vital to cellular repair. These structures, examined in the preceding chapters are by no means an extensive collection of functional nucleic acid configurations, additional structures may be vital to future molecule design and ultimate applications. Such nucleic acid configurations that have only recently been tentatively identified which exhibit genomic importance included telomeres, TALENs (TAL effectors), and zinc fingers. The characteristics of these features can similarly be scrutinized and their attributes added to the genetic tools available for specific applications.

As researchers continue to utilize nucleic acids for novel engineered applications, the in depth understanding of how sequence, structure, and function interrelate will be

essential. By understanding the secrets locked in our genomes, future researchers hopefully can harness this power to treat a variety of diseases, cancers, and other conditions. As more genomic data becomes available (through research or routine genome sequencing), additional complex nucleic acid structures will be discovered, examined, and utilized.

8.2 Whole Genome Sequencing

Complete, or whole, genome sequencing has been perceived, almost from the beginning of our understanding of DNA, as one of the ultimate goals in biological research. In fact, a 1987 report to the U.S. Department of Energy stated boldly that "the ultimate goal of this initiative is to understand the human genome" and "knowledge of the human genome is as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine" [382]. Understanding of the great depth and power of the information in the human genome has been attempted through several different and highly successful "projects" over the past three decades. These range from the panoramic view of our genomes provided by the Human Genome Project, to the portrait of hereditary subpopulations elucidated by the HapMap Project, to finally the highly focused and individualized perspective of our DNA from the $1000 Genome Project.

8.2.1 Human Genome Project

The Human Genome Project was an international research effort (sponsored by the Department of Energy and lead by the National Human Genome Research Institute (NHGRI)) to determine the entire sequence of nucleotides comprising human DNA and to identify the genes that encode the genetic information. Specifically, the listed project goals set forth formally in 1990 included (i) to determine the sequence of the three billion nucleotides that comprise the genome, (ii) to identify all of the approximate 20,000 to 25,000 genes in human DNA, to map their location in the genome and discover their function, (iii) to store this information in databases openly available, (iv) to improve the tools for data analysis, (v) to transfer related technologies to the private sector, and (vi) to address the ethical, legal, and social issues that arise from such a project [382].

In a 13 year long, international effort researchers announced an essentially complete reference genome in 2003, however analysis of the data (goals ii - iv) will continue for many years to come. The reference genome consists of the combined samples of a small number of (somewhat) anonymous donors; this spliced sequence serves as a scaffold for future work to identify differences among individuals [383]. Investigating the differences between individual DNA samples fell to a new project, the global HapMap Project, to help find, understand, and correlate what makes each human unique [384].

8.2.2 HapMap Project

Although no two humans are genetically identical (even monozygotic (identical) twins have infrequent genetic differences due to mutations occurring during development), we do share 99.9% of our genomes with each other. The HapMap Project was designed to study not the entire genome, as in the Human Genome Project, but instead to focus on the 0.1% differences between any two people. Slight variations in human DNA sequences can have a major impact on the characteristics of the individual (in fact, these are what define each of us to be unique), their propensity for disease and their response to environmental factors such as infectious microbes, toxins, and drugs; these differences denote the haplotype of the individual [383]. One of the most common types of sequence variations is the single nucleotide polymorphism (SNP). SNPs are sites in a genome where individuals differ in their DNA sequence, often by a single base. It is currently estimated that the human genome has at least 10 million SNPs, though due to systematic errors in sequencing techniques, this number may be an overestimate [71]. These markers are often associated together in blocks of varying length and complexity that are inherited in tandem and can diverge greatly between different populations and with different conditions. The HapMap Project aims to develop a haplotype map of the human genome, which will describe the common patterns of human genetic variation.

Beginning in 2002, the project selected a sample of 269 individuals and then searched their genomes for several million well-defined SNPs [383]. By determining the genotype of each individual for the particular SNPs examined, a haplotype map was constructed. When the data was collected and correlated with respect to ethnic, geographic, and medical information, patterns began to emerge that pointed to particular variants between subpopulations describing disease and population characteristics

[385]. The HapMap Project fully published its highly correlated data in 2009 [385]. Although the HapMap Project made great progress in our understanding of the origins of the differences between humans, it sampled a limited (and necessarily foreknown) collection of SNPs. The great promise will be when we can sequence whole genomes (not just focus on part of the 0.1% that differentiate each of us) and correlate the genetic, ethnic, environmental, and medical information of vast populations. With this goal in mind, the $1000 Genome Project was initiated.

8.2.3 $1000 Genome Project

Completing the draft sequence of the human genome has been a massive achievement that marks the beginning of the genetics age for society. However, capitalizing on that investment and realizing the potential of the Human Genome Project requires better understanding of what genetic variation means for biology and the ability to sequence a multitude of whole genomes. The HapMap Project began to examine the genetic differences between subpopulations with the sequencing of a small subset of known polymorphisms. To continue towards a truly genetics age, we must have the ability to sequence millions of genomes quickly and cheaply. With this goal in mind (goal v in the Human Genome Project), the NHGRI sponsored the $1000 Genome Project to develop technologies able to rapidly sequence a person's entire genome for less than $1000. One thousand dollars is the conceptual cost point needed for routine analysis of individual genomes [383]. Inexpensive, whole genome sequencing will enable the detection and typing of known, as well as unknown, polymorphisms, and will also offer feasible, large scale population pattern discovery.

Technological progress has been swift and has dramatically driven down the costs and increased the throughput to unprecedented levels. The cost of a human genome sequencing in 2009 was $100,000 per genome, $20,000 in 2011, and $5000 in 2012. Within the next 12 months it is anticipated that whole genome sequencing will cost less than $1000 and take a mere 15 minutes [383, 386–388]. This is a considerable advancement from the first 13 year, $2.7 billion sequencing effort of the Human Genome Project.

Once this goal has been reached, a new sequencing paradigm based on single molecules will be faster, cheaper, and more sensitive, and will permit routine analysis at the whole-genome level. Such information can be used in a wide array of applications including functional/comparative genomics, exploration of microbial diversity for the

agricultural biology field, pathogen identification, transcriptome characterization (and in particular characterization of alternative splice variants), genotype-phenotype correlations, human and animal disease association, pharmacogenomics, the development of new molecular diagnostics and drugs, and personalized medicine [383, 386–388].

8.2.4 The Promise of Personalized Medicine

There is little doubt that the ability to quickly and inexpensively sequence whole genomes will revolutionize the entire health and biotechnology fields. This is a paradigm changing technology. Genomic medicine is a powerful way to personalize health care at the individual level by using patients' genetic information. By identifying the genetic factors associated with disease it is possible to design more effective drugs, to prescribe the best treatments for each patient, to identify and monitor individuals at high risk for disease, to provide prognoses for common diseases, and to avoid adverse drug reactions. Ultimately this will lead to lower costs and more effective treatments. Whole genome sequencing has already begun to transform clinical care. It is, albeit at a limited scale, currently being used to identify new and/or rare diseases, diagnose certain cancers, personalize more successful cancer treatments, and develop new diagnostic tests, medical devices, and patents.

Led by individual tissue allele tests (which look for particular SNPs and other biomarkers) to aid drug therapy decisions, the personalized medicine testing market exceeded $28 billion in 2011 [389]. These tests are used to determine the appropriate therapeutics for an individual patient. The most dynamic part of the market, and the tests that have turned personalized medicine from concept to reality, are these tissue allele tests that determine therapy for cancer. To improve patient survival rates, therapies in the oncology marketplace are being combined with predictive biomarkers to help select patients who will respond to specific drugs. Although whole genome sequencing is not yet routinely used in these applications (except for extremely rare cancers), it could soon supplant tissue allele tests since one whole genome sequence run would provide the information contained in any and all tissue allele tests.

The promise of personalized medicine, in particular for understanding mechanisms of disease, identifying markers or risk of disease, improving diagnosis and definition of disease, validating human targets, and studying response, including adverse effects, to drug treatment is profound. The use of whole genome sequencing as a means of

tailoring healthcare management to an individual's specific needs could soon be commonplace. This technology, combined with detailed understanding of the structures and functions of nucleic acids, will soon "radically [change] the condition of human life," as Fulbright described in the quotation beginning this chapter [381]. It remains to be seen if we can "use them with wisdom"; to investigate not only the technical but societal impacts of whole genome sequencing (goal vi in the Human Genome Project aims) and personalized medicine, then we must examine the policy implications, the stakeholders, and the risks of such technological advancement.

8.3 Policy Implications of Whole Genome Sequencing

From time to time a technology comes along that revolutionizes an entire sector. There is little doubt that the ability to quickly and cheaply sequence whole genomes is one such technology. However, being able to read entire genomes and unlock their mysteries will result in an entirely new set of issues that require societal deliberation, research scrutiny, and regulatory consideration. As we stand at the brink of the age of genetics, we must start looking now at the ethical and legal implications and regulatory challenges of whole genome sequencing. Specifically, recommendations are needed to government, health providers, biotechnology companies, and researchers. Without widespread deliberation, application of this technology may not only be slowed, but may also cause dramatic issues and harsh complications for government, industry, researchers, and patients and their families. In the context of rapidly advancing science and an unprepared healthcare environment, many vital questions arise particularly concerning protection, privacy, perception, practicality, and patentability of genomic information.

8.3.1 Protection for Genetic Information

The process of implementing whole genome sequencing has begun largely in the major government-funded and non-profit research institutes. This technology has been supported by leveraging the strong political will that exists to see real human health benefits from the large investment already made in genetics, and in particular in the Human Genome Project and its various ramifications. The nearly unanimously

federal congressional passage (414-1 House, 95-0 Senate) of the Genetic Information Nondiscrimination Act (GINA) of 2008 (122 Stat. 881) is just the beginning of the governmental protections needed. GINA ensures that genetic information for particular diseases (found from biomarker analysis, known SNPs, and particular gene tests) cannot be used by health insurers or employers to deny or differentiate coverage. It does not prevent discrimination against people when applying for life insurance or long term care and disability insurance. In addition, as research advances and more complex information can be gleaned from genetic codes, GINA fails to protect this information. Further, neither GINA nor the Health Insurance Portability and Accountability Act (HIPAA) of 1996 (110 Stat. 1936), which was originally designed to address the security and privacy of health data, explicitly protect complex genetic information. In particular, due to the limitations of our understanding of genetic information at the time of their creation, neither act explicitly protects a patient with an increased likelihood for a particular complex condition but is written to protect for particular diagnoses. Explicit extensions of GINA and HIPAA to cover whole genome information are needed. Several states currently have more comprehensive safeguards for their citizens, their legislation can serve as a starting point for further patient protections both at the State and the Federal levels [390–394].

8.3.2 Privacy of Genetic Information

For patients, genomic privacy, particularly when combined with electronic health records, poses several issues. For example, personal genomic information, once listed on electronic health records, could be widely available to many healthcare practitioners including many who may not need access to it such as dentists, optometrists, or emergency room doctors. It is foreseeable that many individuals with sensitive and stigmatizing information in their health records will be very concerned about creating a longitudinal, comprehensive record that can be accessed by any healthcare provider; access controls on who has permission to view certain parts of medical information will be needed. However, how to determine which specialties have access to particular medical datum will be difficult.

In addition, there are concerns on how privacy and confidentiality can be ensured for research purposes. Currently there are three options for handling this: using de-identified data that removes the genomic information from the patient, allowing individuals to opt out from having their data used, and inserting an opt-in provision

into consent forms. All three of these options have their own problems and difficulties, mostly stemming from discomfort on the part of patients and incomplete or limited data sets for the researchers. Genomic consent is a new concept for patients and therefore great care must be given when describing the risks and benefits. Whole genome information, particularly when done routinely for large populations, is incredibly powerful in that it does not suppose a particular condition and then look for it, but rather because it allows for spontaneous patterns to emerge. Limiting either the identifying data (which may have unknown research relevance) or allowing patients to choose which parts of their genomic sequence are included may bias the research samples dramatically.

Finally, there are other issues concerning non-medical uses of genomic data such as disclosures required by law, for certain judicial proceedings, for tissue and organ donation purposes, for research purposes, for national security and intelligence activities, and for workers' compensation activities. It is important to note that in the United States people are compelled to release their health records 25 million times per year for a range of reasons, specifically for life insurance applications and specialized employment applications [383, 386]. It is inevitable that genomic information will be used in ways beyond research and medical treatment applications.

The real remaining questions are only how broadly, with what purposes, and with what consequences will genomic information be used. Ultimately, as a research community we must decide how we can continue to support science and research and also protect individual privacy.

8.3.3 Perception of Genetic Information

As this technology becomes fully available, patients will need to be advised on the benefits and risks of whole genome sequencing. Many ethical difficulties arise in these situations. While the benefits have been discussed previously, risks concerning the revelation of information that the patient would prefer not to know or have known by others are possible. Late-onset diseases or behavioral tendencies that are difficult or impossible to prevent can lead to stigmatization or discrimination. Patients may not even want to know about certain susceptibility conditions for which preventative actions are possible because of their unwillingness to engage in the behavioral or lifestyle changes that would then be necessary [395, 396]. In addition, if a patient does decide to have their genome sequenced to assess their risk of one particular disease,

the results may also have information about risks for other diseases. It is unclear how much of this information should be communicated to patients. Should physicians only give information for which there is clinical utility, or is it unethical to withhold any information related to patient health? If the patient only wishes to know information directly related to the reason for his or her genome sequencing, it is unclear what responsibility the healthcare providers have in sharing gene variants associated with increased risk for disease. To further complicate matters, the thresholds for the clinical validity of a gene variant and disease risk is undefined and patients' risk of disease based on their genomic profiles will change as research advances.

Next, we inherit our genomes; they are closely tied to our immediate and extended family. There are many unanswered questions concerning the responsibility of a family member to inform or shield their family from their genetic information and particular diseases or risk factors [396]. Sequencing of embryos, fetuses, and infants further muddles these issues since the previous discussion has assumed that a competent adult is making their own decisions [395]. The rights, responsibilities, and consequences of whole genome data inside of the family unit are complicated and unclear at this time. Continued and focused efforts to protect and educate people about their genetic information and to develop guidelines to navigate these complex issues are greatly needed before routine whole genome sequencing and personalized medicine can be spread widely throughout society.

Translating whole genome information to doctors and patients will be difficult. The different responsibilities of the primary care physician, genetic counselor, and medical geneticist need to be examined and perhaps redefined. Currently there are only about 2,000 genetic counselors in the United States (as compared to over 560,000 physicians) [397]. If genetic counselors are to play a central role as genomic medicine is integrated into clinical practice, there is a need for more genetic counselors. At the same time, other health care providers will need to learn more genomics so they can better inform and treat their patients. The standard practice of care for doctors with regard to genomic medicine needs to be defined. Finally, if patients are expected to make important lifestyle decisions based on their complex and individual genomic data, they will need appropriate support to help them understand their genetic proclivities, the implications for their families, and to make responsible lifestyle changes.

8.3.4 Practicality of Genetic Information

Genomics is now beginning to trickle out of the academic labs and into the clinical medical field. With the rise of electronic medical records and the ability to sift enormous quantities of data, biobanks of genomic and traditional medical information are being created. Researchers have begun to take full advantage of these resources; however, the vast majority of hospital and other clinics do not have the infrastructure necessary to begin to implement this and other similar genome-based medical innovations.

Further, health insurance companies are unlikely to cover sequencing in the absence of proven clinical utility. Initial government payments, in the form of research grants, incentives for insurance companies, or funding through Medicaid or Medicare, for genome sequencing to provided coverage are possible remedies until clinical utility is fully developed. Once improved patient outcome is proven, and the consequential cost saving measures are fully understood, whole genome sequencing will move from being a theoretical promise to that of a practical application.

8.3.5 Patentability of Genetic Information

There are also other outstanding legal issues that are currently being debated, one of particular concern to the biotechnology industry is the ability to patent all or part of a gene or genome. The United States Patent and Trademark Office has granted thousands of patents on human genes. In fact, about 20% of human genes are patented. A gene patent holder has the right to prevent anyone from studying, testing, or even examining a particular gene. This is a hotly contended issue: on one side, claims have been made that as a result, scientific research and genetic testing has been delayed, limited, or even shut down due to concerns about gene patents; on the other side biotechnology companies make enormous investments into the discovery, examination of a gene, and development of any related diagnostic tests or drug treatment regimes.

An example of this raging debate is a 2009 lawsuit, *Association for Molecular Pathology, et al., vs. United States Patent and Trademark Office, et al.*, filed against Myriad Genetics and the United States Patent and Trademark Office by professional medical organizations, doctors, and patients. The complaint challenged specific claims on isolated genes and diagnostic methods in several of Myriad Genetics patents on the

BRCA1 and BRCA2 genes. BRCA1 and BRCA2 are two gene mutations which can dramatically increase a woman's likelihood of developing breast cancer. The plaintiffs wanted certain claims declared invalid on the grounds that they are not patentable subject matter, that is that the isolated genes are unpatentable products of nature and that diagnostic method claims are mere throughput processes that do not yield any real world transformation. Patentable subject matter, along with novelty, inventive step, transformation, utility, and industrial applicability are the fundamental requirements of patentability. The initial case was heard in United States District Court and was ruled, on March 29, 2010, that the patent claims were invalid. Upon appeal in the Federal Court, the decision was overturned in part and upheld in part. The July 29, 2011 decision overturned the District Court's finding that the claims covering isolated gene sequences are invalid and also overturned the invalidity of some of the diagnostic claims; the Federal Circuit upheld the finding that the claims for the diagnostic methods that only compare or analyze sequences (and thus have no transformative step) are invalid. Further appeals are likely; on October 12, 2011 the plaintiffs petitioned the Supreme Court to hear the case. In addition, Congress is also currently examining whether patents on genetic material should be treated differently from other intellectual property. Fundamentally, this issue is far from decided.

The patentability of genetic information is only going to get more complicated with the introduction of routine whole genome sequencing. Instead of involving a single gene at one location on the human genome, whole genome sequence data necessarily will involve thousands of genes and millions of SNPs across many chromosomes. In addition, a paradigm of free and open access to sequencing data has been set with the public release of the preliminary reference sequence data from the Human Genome Project in 2000. The decision, from the project's onset, to release sequence data daily has fueled genomics ever since and has led to other open data sets such as the Cancer Genome Atlas [398, 399]. This precedent for whole genome sequencing has served the research community well; how such a system can mesh with the patentability of genes is currently unknown.

8.4 Conclusions

Translating the technical details of nucleic acid sequences and their complex structures developed in Chapters 1 to 7 and combining these concepts with whole genome sequence information paves the way to truly understanding our DNA. Humans are now

at the brink of a new era, the genetic age, in which rapid, inexpensive, and detailed knowledge of our individual chromosomes will be commonplace. The promise of personalized medicine, in particular for understanding mechanisms of disease, identifying markers or risk of disease, improving diagnosis and definition of disease, validating targets for treatments and studying the response (including adverse effects) to drug treatments, is vast. Continued research in both the complex physical and biochemical interactions of nucleic acids and in the public policies concerning their usage in society are needed as we move towards the future and realize the possibilities of the genetic age.

Bibliography

[1] Crick, F. *What Mad Pursuit: A Personal View of Scientific Discovery*; Basic Books, 1988.

[2] Ridley, M. *Genome: The Autobiography of a Species in 23 Chapters*; Harper Perennial, 1999.

[3] Avery, O. T., MacLeod, C. M., and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *The Journal of Experimental Medicine 79*, 137.

[4] Hershey, A. D., and Chase, M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology 36*, 39.

[5] Lederberg, J. (1994) The transformation of genetics by DNA: an anniversary celebration of Avery, MacLeod and McCarty (1944). *Genetics 136*, 423.

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

[7] Watson, J. D., and Crick, F. H. C. (1953) The structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology 18*, 123.

[8] Franklin, R. E., and Gosling, R. G. (1953) Molecular configuration in sodium thymonucleate. *Nature 171*, 740–741.

[9] Wilkins, M. H., Stokes, A. R., and Wilson, H. R. (1953) Molecular structure of deoxypentose nucleic acids. *Nature 171*, 738.

[10] Chargaff, E. (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Cellular and Molecular Life Sciences 6*, 201–209.

[11] Garrett, R. H., and Grisham, C. M. *Principles of Biochemistry: with a Human Focus*; Brooks/Cole Publishing Company, 2001.

[12] Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag, New York, 1984.

[13] Bloomfield, V. A., Crothers, D. M., and Tinoco, I. *Nucleic Acids: Structures, Properties, and Functions*; University Science Books, 2000.

[14] Nelson, D. L., and Cox, M. M. *Lehninger Principles of Biochemistry*; 2000.

[15] Calladine, C. R., and Drew, H. R. *Understanding DNA: The Molecule and How It Works*; Academic Press, San Diego, 1998.

[16] Kyogoku, Y., Lord, R. C., and Rich, A. (1967) The effect of substituents on the hydrogen bonding of adenine and uracil derivatives. *Proceedings of the National Academy of Sciences of the United States of America 57*, 250.

[17] Hoogsteen, K. (1959) The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine. *Acta Crystallographica 12*, 822–823.

[18] Hoogsteen, K. (1963) The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallographica 16*, 907–916.

[19] Wimberly, B., Varani, G., and Tinoco, I. (1993) The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochemistry 32*, 1078–1087.

[20] Quigley, G. J., Ughetto, G., van der Marel, G. A., van Boom, J. H., Wang, A. H., and Rich, A. (1986) Non-Watson-Crick GC and AT base pairs in a DNA-antibiotic complex. *Science 232*, 1255.

[21] Nikolova, E. N., Kim, E., Wise, A. A., O'Brien, P. J., Andricioaei, I., and Al-Hashimi, H. M. (2011) Transient Hoogsteen base pairs in canonical duplex DNA. *Nature*

[22] Crick, F. H. C. (1966) Codon–anticodon pairing: the wobble hypothesis. *Journal of Molecular Biology 19*, 548–555.

[23] Strobel, M., Lyons, C. S., and Mittal, K. L. *Plasma surface modification of polymers: relevance to adhesion*; Vsp, 1994.

[24] Lezius, A. G., and Domin, E. (1973) A wobbly double helix. *Nature 244*, 169–170.

[25] Romaniuk, P. J. Stability studies of short, imperfect RNA double helices. Ph.D. thesis, McMaster University, 1979.

[26] Hermann, T., and Westhof, E. (1999) Non-Watson-Crick base pairs in RNA-protein recognition. *Chemistry and Biology -London 6*, 335–343.

[27] Goodman, L., and Tso, P. *Basic Principles in Nucleic Acid Chemistry*; Academic Press, New York, 1974; Vol. 1; p 93.

[28] Solie, T. N., and Schellman, J. A. (1968) The interaction of nucleosides in aqueous solution. *Journal of Molecular Biology 33*, 61–77.

[29] Gill, S. J., Downing, M., and Sheats, G. F. (1967) The enthalpy of self-association of purine derivatives in water. *Biochemistry 6*, 272–276.

[30] Tribolet, R., and Sigel, H. (1987) Self-association and protonation of adenosine 5′-monophosphate in comparison with its 2′-and 3′-analogues and tubercidin 5′-monophosphate (7-deaza-AMP). *European Journal of Biochemistry 163*, 353–363.

[31] Weaver, W. (1948) Science and Complexity. *American Scientist 36*, 536–544.

[32] Klug, A. (2004) The discovery of the DNA double helix. *Journal of Molecular Biology 335*, 3–26.

[33] Maddox, B. *Rosalind Franklin: The Dark Lady of DNA*; Harper Perennial, 2003.

[34] Sayre, A. *Rosalind Franklin and DNA*; Norton, New York, 1975.

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

[36] Pauling, L., and Corey, R. B. (1953) A proposed structure for the nucleic acids. *Proceedings of the National Academy of Sciences of the United States of America 39*, 84.

[37] Wells, R. D., and Harvey, S. C. *Unusual DNA structures*; St. Martins Press, New York, 1988.

[38] Watson, J., Baker, T., and S. Bell, A. G. *Molecular Biology of the Gene*; Addison-Wesley Publishing Company, 2003.

[39] Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H., and Toole, J. J. (1992) Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature 355*, 564–566.

[40] Maier, B., Bensimon, D., and Croquette, V. (2000) Replication by a single DNA polymerase of a stretched single-stranded DNA. *Proceedings of the National Academy of Sciences of the United States of America 97*, 12002.

[41] Frank-Kamenetskii, M. D., and Mirkin, S. M. (1995) Triplex DNA structures. *Annual Review of Biochemistry 64*, 65–95.

[42] Phan, A. T., and Mergny, J. L. (2002) Human telomeric DNA: G-quadruplex, i-motif and Watson–Crick double helix. *Nucleic Acids Research 30*, 4618.

[43] Davis, J. T. (2004) G-Quartets 40 years later: from 5 GMP to molecular biology and supramolecular chemistry. *Angewandte Chemie International Edition 43*, 668–698.

[44] Tinoco, I., and Bustamante, C. (1999) How RNA folds. *Journal of Molecular Biology 293*, 271–281.

[45] Chen, S. J. (2008) RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *The Annual Review of Biophysics*

[46] Sen, S., and Nilsson, L. (2001) MD simulations of homomorphous PNA, DNA, and RNA single strands: characterization and comparison of conformations and dynamics. *Journal of the American Chemical Society 123*, 7414–7422.

[47] Wereszczynski, J., and Andricioaei, I. (2006) On structural transitions, thermodynamic equilibrium, and the phase diagram of DNA and RNA duplexes under torque and tension. *Proceedings of the National Academy of Sciences 103*, 16200.

[48] Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell 31*, 147–157.

[49] Joyce, G. F. (1991) The rise and fall of the RNA world. *The New Biologist 3*, 399.

[50] Klussmann, S. *The aptamer handbook: functional oligonucleotides and their applications*; Wiley, 2006.

[51] Sun, L. Q., Cairns, M. J., Saravolac, E. G., Baker, A., and Gerlach, W. L. (2000) Catalytic nucleic acids: from lab to applications. *Pharmacological Reviews 52*, 325–348.

[52] Breaker, R. R., and Joyce, G. F. (1994) A DNA enzyme that cleaves RNA. *Chemistry & Biology 1*, 223.

[53] Breaker, R. R. (1997) DNA enzymes. *Nature Biotechnology 15*.

[54] Breaker, R. S. (1999) Catalytic DNA: in training and seeking employment. *Nature Biotechnology 17*, 422–423.

[55] Santoro, S. W., and Joyce, G. F. (1997) A general purpose RNA-cleaving DNA enzyme. *Proceedings of the National Academy of Sciences of the United States of America 94*, 4262.

[56] Santoro, S. W., and Joyce, G. F. (1998) Mechanism and utility of an RNA-cleaving DNA enzyme. *Biochemistry 37*, 13330–13342.

[57] Flory, P. J., and Volkenstein, M. (1969) Statistical mechanics of chain molecules. *Biopolymers 8*, 699–700.

[58] Hsieh, C. C., Balducci, A., and Doyle, P. S. (2008) Ionic effects on the equilibrium dynamics of DNA confined in nanoslits. *Nano Letters 8*, 1683–1688.

[59] Nelson, P., Radosavljevic, M., and Bromberg, S. *Biological Physics*; Freeman, New York, 2008.

[60] Ghosh, A., and Bansal, M. (2003) A glossary of DNA structures from A to Z. *Acta Crystallographica Section D: Biological Crystallography 59*, 620–626.

[61] Harvey, S. C. (1983) DNA structural dynamics: longitudinal breathing as a possible mechanism for the B–Z transition. *Nucleic Acids Research 11*, 4867.

[62] Allemand, J. F., Bensimon, D., Lavery, R., and Croquette, V. (1998) Stretched and overwound DNA forms a Pauling-like structure with exposed bases. *Proceedings of the National Academy of Sciences 95*, 14152.

[63] Mergell, B., Ejtehadi, M. R., and Everaers, R. (2003) Modeling DNA structure, elasticity, and deformations at the base-pair level. *Physical Review E 68*, 21911.

[64] Cluzel, P., Lebrun, A., Heller, C., Lavery, R., Viovy, J. L., Chatenay, D., and Caron, F. (1996) DNA: an extensible molecule. *Science 271*, 792.

[65] Leger, J. F., Romano, G., Sarkar, A., Robert, J., Bourdieu, L., Chatenay, D., and Marko, J. F. (1999) Structural transitions of a twisted and stretched DNA molecule. *Physical Review Letters 83*, 1066–1069.

[66] Koo, H. S., Wu, H. M., and Crothers, D. M. (1986) DNA bending at adenine· thymine tracts. *Nature 320*, 501–506.

[67] Nadeau, J. G., and Crothers, D. M. (1989) Structural basis for DNA bending. *Proceedings of the National Academy of Sciences 86*, 2622.

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

[69] Arnott, S., L., H. D. W., Dover, S. D., Fuller, W., and Hodgson, A. R. (1973) Structures of synthetic polynucleotides in the A-RNA and A'-RNA conformations: X-ray diffraction analyses of the molecular conformations of polyadenylic acid· polyuridylic acid and polyinosinic acid· polycytidylic acid. *Journal of Molecular Biology 81*, 107–108.

[70] Stein, A., Takasuka, T. E., and Collings, C. K. (2010) Are nucleosome positions in vivo primarily determined by histone–DNA sequence preferences? *Nucleic Acids Research 38*, 709–719.

[71] Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., and Takahashi, H. (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*

[72] Bennett, S. (2004) Solexa Ltd. *Pharmacogenomics 5*, 433–438.

[73] Smith, S. B., Cui, Y., and Bustamante, C. (1996) Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science 271*, 795.

[74] Rivetti, C., Walker, C., and Bustamante, C. (1998) Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. *Journal of Molecular Biology 280*, 41–59.

[75] Kuznetsov, S. V., Shen, Y., Benight, A. S., and Ansari, A. (2001) A semiflexible polymer model applied to loop formation in DNA hairpins. *Biophysical Journal 81*, 2864–2875.

[76] Achter, E. K., and Kelsenfeld, G. (1971) The conformation of single-strand polynucleotides in solution: sedimentation studies of apurinic acid. *Biopolymers 10*, 1625–1634.

[77] Murphy, M. C., Rasnik, I., Cheng, W., Lohman, T. M., and Ha, T. (2004) Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy. *Biophysical Journal 86*, 2530–2537.

[78] Mills, J. B., Vacano, E., and Hagerman, P. J. (1999) Flexibility of single-stranded DNA: use of gapped duplex helices to determine the persistence lengths of Poly (dT) and Poly (dA) 1. *Journal of Molecular Biology 285*, 245–257.

[79] Tinland, B., Pluen, A., Sturm, J., and Weill, G. (1997) Persistence length of single-stranded DNA. *Macromolecules 30*, 5763–5765.

[80] Crick, F. H. C. (1968) The origin of the genetic code. *Journal of Molecular Biology 38*, 367–379.

[81] Jayasena, S. D. (1999) Aptamers: an emerging class of molecules that rival antibodies in diagnostics. *Clinical Chemistry 45*, 1628.

[82] James, W. (2000) Aptamers. *Encyclopedia of Analytical Chemistry*

[83] Brody, E. N., and Gold, L. (2000) Aptamers as therapeutic and diagnostic agents. *Reviews in Molecular Biotechnology 74*, 5–13.

[84] Klug, S. J., and Famulok, M. (1994) All you wanted to know about SELEX. *Molecular Biology Reports 20*, 97–107.

[85] Fitzwater, T., and Polisky, B. (1996) A SELEX primer. *Methods in Enzymology 267*, 275–301.

[86] Gold, L., Brown, D., He, Y., Shtatland, T., Singer, B. S., and Wu, Y. (1997) From oligonucleotide shapes to genomic SELEX: novel biological regulatory loops. *Proceedings of the National Academy of Sciences 94*, 59.

[87] Gold, L., Brody, E., Heilig, J., and Singer, B. (2002) One, two, infinity: genomes filled with aptamers. *Chemistry & Biology 9*, 1259.

[88] Hamaguchi, N., Ellington, A., and Stanton, M. (2001) Aptamer beacons for the direct detection of proteins. *Analytical Biochemistry 294*, 126–131.

[89] Mobley, D. L., and Dill, K. A. (2009) Binding of small-molecule ligands to proteins. *Structure 17*, 489–498.

[90] Huang, Y. F., Chang, H. T., and Tan, W. (2008) Cancer cell targeting using multiple aptamers conjugated on nanorods. *Analytical Chemistry 80*, 567–72.

[91] Macaya, R. F., Schultze, P., Smith, F. W., Roe, J. A., and Feigon, J. (1993) Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution. *Proceedings of the National Academy of Sciences of the United States of America 90*, 3745.

[92] Stojanovic, M. N., De Prada, P., and Landry, D. W. (2001) Aptamer-based folding fluorescent sensor for cocaine. *Journal of the American Chemical Society 123*, 4928–4931.

[93] Robertson, D. L., and Joyce, G. F. (1990) Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature 344*, 467–468.

[94] Beaudry, A. A., and Joyce, G. F. (1992) Directed evolution of an RNA enzyme. *Science 257*, 635.

[95] Emilsson, G. M., and Breaker, R. R. (2002) Deoxyribozymes: new activities and new applications. *Cellular and Molecular Life Sciences 59*, 596–607.

[96] Li, Y., and Breaker, R. R. (1999) Deoxyribozymes: new players in the ancient game of biocatalysis. *Current Opinion in Structural Biology 9*, 315–323.

[97] Breaker, R. R., and Joyce, G. F. (1995) A DNA enzyme with Mg2+-dependent RNA phosphoesterase activity. *Chemistry & Biology 2*, 655–660.

[98] Carmi, N., Shultz, L. A., and Breaker, R. R. (1996) In vitro selection of self-cleaving DNAs. *Chemistry & Biology 3*, 1039–1046.

[99] Carmi, N., Balkhi, S. R., and Breaker, R. R. (1998) Cleaving DNA with DNA. *Proceedings of the National Academy of Sciences of the United States of America 95*, 2233.

[100] Carmi, N., and Breaker, R. R. (2001) Characterization of a DNA-cleaving deoxyribozyme. *Bioorganic & Medicinal Chemistry 9*, 2589–2600.

[101] Cuenoud, B., and Szostak, J. W. (1995) A DNA metalloenzyme with DNA ligase activity. *Nature 375*, 611–614.

[102] Cruz, R. P. G., Withers, J. B., and Li, Y. (2004) Dinucleotide junction cleavage versatility of 8-17 deoxyribozyme. *Chemistry & Biology 11*, 57–67.

[103] Li, Y., and Sen, D. (1996) A catalytic DNA for porphyrin metallation. *Nature Structural Biology 3*, 743.

[104] Gilbert, W. (1986) Origin of life: the RNA world. *Nature 319*.

[105] Perreault, J. P., Wu, T., Cousineau, B., Ogilvie, K. K., and Cedergren, R. (1990) Mixed deoxyribo-and ribo-oligonucleotides with catalytic activity. *Nature*

[106] Achenbach, J. C., Chiuman, W., Cruz, R. P. G., and Li, Y. (2004) DNAzymes: from creation in vitro to application in vivo. *Current Pharmaceutical Biotechnology 5*, 321–336.

[107] Benson, V. L., Khachigian, L. M., and Lowe, H. C. (2008) DNAzymes and cardiovascular disease. *British Journal of Pharmacology 154*, 741–748.

[108] Garibotti, A. V., Knudsen, S. M., Ellington, A. D., and Seeman, N. C. (2006) Functional DNAzymes organized into two-dimensional arrays. *Nano Letters 6*, 1505–1507.

[109] Khachigian, L. M. (2002) DNAzymes: cutting a path to a new class of therapeutics. *Current Opinions in Molecular Therapeutics 4*, 119–121.

[110] Sioud, M., and Iversen, P. O. (2005) Ribozymes, DNAzymes and small interfering RNAs as therapeutics. *Current Drug Targets 6*, 647–653.

[111] Wang, D. Y., Lai, B. H. Y., Feldman, A. R., and Sen, D. (2002) A general approach for the use of oligonucleotide effectors to regulate the catalysis of RNA-cleaving ribozymes and DNAzymes. *Nucleic Acids Research 30*, 1735.

[112] Smith, R. M., and Hansen, D. E. (1998) The pH-rate profile for the hydrolysis of a peptide bond. *Journal of the American Chemical Society 120*, 8910–8913.

[113] Cairns, M. J., King, A., and Sun, L. Q. (2003) Optimization of the 10–23 DNAzyme–substrate pairing interactions enhanced RNA cleavage activity at purine–cytosine target sites. *Nucleic Acids Research 31*, 2883.

[114] Cairns, M. J., and Sun, L. Q. (2004) Target-site selection for the 10-23 DNAzyme. *Methods in Molecular Biology 252*, 267–278.

[115] Schubert, S., Gül, D. C., Grunert, H. P., Zeichhardt, H., Erdmann, V. A., and Kurreck, J. (2003) RNA cleaving '10-23'DNAzymes with enhanced stability and activity. *Nucleic Acids Research 31*, 5982.

[116] Breaker, R. R. Molecular Biology: Making Catalytic DNAs. 2000.

[117] Pun, S. H., Tack, F., Bellocq, N. C., Cheng, J., Grubbs, B. H., Jensen, G. S., Davis, M. E., Brewster, M., Janicot, M., and Janssens, B. (2004) Targeted delivery of RNA-cleaving DNA enzyme (DNAzyme) to tumor tissue by transferrin-modified, cyclodextrin-based particles. *Cancer Biology & Therapy 3*, 641.

[118] Cheng, Y. K., and Pettitt, B. M. (1992) Stabilities of double-and triple-strand helical nucleic acids. *Progress in Biophysics and Molecular Biology 58*, 225.

[119] Mannuss, A., Dukowic-Schulze, S., Suer, S., Hartung, F., Pacher, M., and Puchta, H. (2010) RAD5A, RECQ4A, and MUS81 have specific functions in homologous recombination and define different pathways of DNA repair in Arabidopsis thaliana. *The Plant Cell Online 22*, 3318–3330.

[120] Seidman, M. M., and Glazer, P. M. (2003) The potential for gene repair via triple helix formation. *Journal of Clinical Investigation 112*, 487–494.

[121] Gehring, K., Leroy, J. L., and Guéron, M. (1993) A tetrameric DNA structure with protonated cytosine-cytosine base pairs.

[122] Box, G., and Draper, N. *Empirical Model-Building and Response Surfaces*; John Wiley & Sons, Oxford, 1987.

[123] Orozco, M., Noy, A., and Pérez, A. (2008) Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Current Opinion in Structural Biology 18*, 185–193.

[124] de Pablo, J. J. (2011) Coarse-grained simulations of macromolecules: from DNA to nanocomposites. *Annual Review of Physical Chemistry 62*, 555–574.

[125] Weiner, S. J., Kollman, P. A., Nguyen, D. T., and Case, D. A. (1986) An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry 7*, 230–252.

[126] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society 117*, 5179–5197.

[127] Sponer, J., Leszczynski, J., and Hobza, P. (1996) Hydrogen bonding and stacking of DNA bases: a review of quantum-chemical ab initio studies. *Journal of Biomolecular Structure & Dynamics 14*, 117.

[128] Cieplak, P. (1998) Nucleic acid force fields. *Encyclopedia of Computational Chemistry*

[129] Foloppe, N., and MacKerell Jr, A. D. (2000) All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry 21*, 86–104.

[130] MacKerell Jr, A. D., and Banavali, N. K. (2000) All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *Journal of Computational Chemistry 21*, 105–120.

[131] Brameld, K., Dasgupta, S., and Goddard III, W. A. (1997) Distance dependent hydrogen bond potentials for nucleic acid base pairs from ab initio quantum mechanical calculations (LMP2/cc-pVTZ). *The Journal of Physical Chemistry B 101*, 4851–4859.

[132] Elstner, M., Hobza, P., Frauenheim, T., Suhai, S., and Kaxiras, E. (2001) Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density-functional-theory based treatment. *The Journal of Chemical Physics 114*, 5149.

[133] MacKerell Jr, A., Banavali, N., and Foloppe, N. (2001) Development and current status of the CHARMM force field for nucleic acids. *Biopolymers 56*, 257–265.

[134] Cheatham III, T. E., and Young, M. A. (2001) Molecular dynamics simulation of nucleic acids: successes, limitations, and promise. *Biopolymers 56*, 232–256.

[135] Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham III, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of [alpha]/[gamma] conformers. *Biophysical Journal 92*, 3817–3829.

[136] Beveridge, D. L., and Ravishanker, G. (1994) Molecular dynamics studies of DNA. *Current Opinion in Structural Biology 4*, 246–255.

[137] Cheatham III, T. E., and Kollman, P. A. (2000) Molecular dynamics simulation of nucleic acids. *Annual Review of Physical Chemistry 51*, 435–471.

[138] Kannan, S., and Zacharias, M. (2007) Folding of a DNA hairpin loop structure in explicit solvent using replica-exchange molecular dynamics simulations. *Biophysical Journal 93*, 3218–3228.

[139] Lavery, R., Zakrzewska, K., Beveridge, D., Bishop, T. C., Case, D. A.,

Cheatham, T., Dixit, S., Jayaram, B., Lankas, F., and Laughton, C. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Research 38*, 299.

[140] Levitt, M. Computer Simulation of DNA Double-Helix Dynamics. 1983.

[141] McDowell, S. E., Nad'a Špačková, J., and Walter, N. G. (2007) Molecular dynamics simulations of RNA: an in silico single molecule approach. *Biopolymers 85*, 169.

[142] Norberg, J., and Nilsson, L. (2002) Molecular dynamics applied to nucleic acids. *Accounts of Chemical Research 35*, 465–472.

[143] Noy, A., Soteras, I., Luque, F. J., and Orozco, M. (2009) The impact of monovalent ion force field model in nucleic acids simulations. *Physical Chemistry Chemical Physics 11*, 10596–10607.

[144] Ponomarev, S. Y., Thayer, K. M., and Beveridge, D. L. (2004) Ion motions in molecular dynamics simulations on DNA. *Proceedings of the National Academy of Sciences of the United States of America 101*, 14771.

[145] Ponomarev, S. Y., Putkaradze, V., and Bishop, T. C. (2009) Relaxation dynamics of nucleosomal DNA. *Physical Chemistry Chemical Physics 11*, 10633–10643.

[146] Ponomarev, S. Y., Bishop, T. C., and Putkaradze, V. (2009) DNA Relaxation Dynamics in 1ID3 Yeast Nucleosome MD Simulation. *Biophysical Journal 96*, 577.

[147] Lankaš, F., Gonzalez, O., Heffler, L. M., Stoll, G., Moakher, M., and Maddocks, J. H. (2009) On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Physical Chemistry Chemical Physics 11*, 10565–10588.

[148] Reshetnikov, R. V., Golovin, A. V., and Kopylov, A. M. (2010) Comparison of models of thrombin-binding 15-mer DNA aptamer by molecular dynamics simulation. *Biochemistry (Moscow) 75*, 1017–1024.

[149] Wang, H., and Laughton, C. A. (2009) Evaluation of molecular modelling methods to predict the sequence-selectivity of DNA minor groove binding ligands. *Physical Chemistry Chemical Physics 11*, 10722–10728.

[150] Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G.,

Profeta, S., and Weiner, P. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society 106*, 765–784.

[151] Zheng, G., Czapla, L., Srinivasan, A. R., and Olson, W. K. (2010) How stiff is DNA? *Physical Chemistry Chemical Physics 12*, 1399–1406.

[152] Manning, R. S., Maddocks, J. H., and Kahn, J. D. (1996) A continuum rod model of sequence-dependent DNA structure. *The Journal of Chemical Physics 105*, 5626.

[153] Wocjan, T., Krieger, J., Krichevsky, O., and Langowski, J. (2009) Dynamics of a fluorophore attached to superhelical DNA: FCS experiments simulated by Brownian dynamics. *Physical Chemistry Chemical Physics 11*, 10671–10681.

[154] Swigon, D. (2009) The Mathematics of DNA Sturcture, Mechanics, and Dynamics. *Mathematics of DNA Structure, Function and Interactions* 293–320.

[155] Tepper, H. L., and Voth, G. A. (2005) A coarse-grained model for double-helix molecules in solution: spontaneous helix formation and equilibrium properties. *The Journal of Chemical Physics 122*, 124906.

[156] Pasquali, S., and Derreumaux, P. (2010) HiRE-RNA: a high resolution coarse-grained energy model for RNA. *The Journal of Physical Chemistry B*

[157] Dans, P. D., Zeida, A., Machado, M. R., and Pantano, S. (2010) A coarse grained model for atomic-detailed DNA simulations with explicit electrostatics. *Journal of Chemical Theory and Computation 6*, 1711–1725.

[158] Zhang, F., and Collins, M. A. (1995) Model simulations of DNA dynamics. *Physical Review E 52*, 4217–4224.

[159] Bruant, N., Flatters, D., Lavery, R., and Genest, D. (1999) From atomic to mesoscopic descriptions of the internal dynamics of DNA. *Biophysical Journal 77*, 2366–2376.

[160] Knotts IV, T. A., Rathore, N., Schwartz, D. C., and De Pablo, J. J. (2007) A coarse grain model for DNA. *The Journal of Chemical Physics 126*, 084901.

[161] Sambriski, E. J., Schwartz, D. C., and de Pablo, J. J. (2009) A mesoscale model of DNA and its renaturation. *Biophysical Journal 96*, 1675–1690.

[162] Sambriski, E. J., Schwartz, D. C., and de Pablo, J. J. (2009) Uncovering path-

ways in DNA oligonucleotide hybridization via transition state analysis. *Proceedings of the National Academy of Sciences 106*, 18125.

[163] DeMille, R. C., Cheatham III, T. E., and Molinero, V. (2010) A coarse-grained model of DNA with explicit solvation by water and ions. *The Journal of Physical Chemistry B* 889.

[164] Morriss-Andrews, A., Rottler, J., and Plotkin, S. S. (2010) A systematically coarse-grained model for DNA and its predictions for persistence length, stacking, twist, and chirality. *The Journal of Chemical Physics 132*, 035105.

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

[166] Linak, M. C., Tourdot, R., and Dorfman, K. D. (2011) Moving beyond Watson–Crick models of coarse grained DNA dynamics. *The Journal of Chemical Physics 135*, 205102.

[167] Ortiz, V., and de Pablo, J. J. (2011) Molecular origins of DNA flexibility: sequence effects on conformational and mechanical properties. *Physical Review Letters 106*, 238107.

[168] Drukker, K., and Schatz, G. C. (2000) A model for simulating dynamics of DNA denaturation. *Journal of Physical Chemistry B 104*, 6108–6111.

[169] Drukker, K., Wu, G., and Schatz, G. C. (2001) Model simulations of DNA denaturation dynamics. *The Journal of Chemical Physics 114*, 579.

[170] Sales-Pardo, M., Guimera, R., Moreira, A. A., Widom, J., and Amaral, L. A. N. (2005) Mesoscopic modeling for nucleic acid chain dynamics. *Physical Review E 71*, 51902.

[171] Buyukdagli, S., Sanrey, M., and Joyeux, M. (2006) Towards more realistic dynamical models for DNA secondary structure. *Chemical Physics Letters 419*, 434–438.

[172] Kenward, M., and Dorfman, K. D. (2009) Brownian dynamics simulations of single-stranded DNA hairpins. *The Journal of Chemical Physics 130*, 095101.

[173] Kenward, M., and Dorfman, K. D. (2009) Coarse-Grained Brownian Dynamics Simulations of the 10-23 DNAzyme. *Biophysical Journal 97*, 2785–2793.

[174] Linak, M. C., and Dorfman, K. D. (2010) Analysis of a DNA simulation model

through hairpin melting experiments. *The Journal of Chemical Physics 133*, 125101.

[175] Ouldridge, T. E., Johnston, I. G., Louis, A. A., and Doye, J. P. K. (2009) The self-assembly of DNA Holliday junctions studied with a minimal model. *The Journal of Chemical Physics 130*, 065101.

[176] Ouldridge, T. E., Louis, A. A., and Doye, J. P. K. (2009) DNA nanotweezers studied with a coarse-grained model of DNA. *Physical Review Letters 104*.

[177] Ouldridge, T. E., Louis, A. A., and Doye, J. P. K. (2011) Structural, mechanical and thermodynamic properties of a coarse-grained DNA model. *Journal of Chemical Physics 134*, 085101.

[178] de la Torre, G. (1994) Hydrodynamic properties of a double-helical model for DNA. *Biophysical Journal 66*, 1573–1579.

[179] Huertas, M. L., Navarro, S., López Martínez, M. C., and García de la Torre, J. (1997) Simulation of the conformation and dynamics of a double-helical model for DNA. *Biophysical Journal 73*, 3142–3153.

[180] Sheng, Y. J., Chen, J. Z. Y., and Tsao, H. K. (2002) Open-to-closed transition of a hard-sphere chain with attractive ends. *Macromolecules 35*, 9624–9627.

[181] Trovato, F., and Tozzini, V. (2008) Supercoiling and local denaturation of plasmids with a minimalist DNA model. *The Journal of Physical Chemistry B 112*, 13197–13200.

[182] Savelyev, A., and Papoian, G. A. (2009) Molecular renormalization group coarse-graining of polymer chains: Application to double-stranded DNA. *Biophysical Journal 96*, 4044–4052.

[183] Savelyev, A., and Papoian, G. A. (2010) Chemically accurate coarse graining of double-stranded DNA. *Proceedings of the National Academy of Sciences 107*, 20340.

[184] Mielke, S. P., Grønbech-Jensen, N., Krishnan, V. V., Fink, W. H., and Benham, C. J. (2005) Brownian dynamics simulations of sequence-dependent duplex denaturation in dynamically superhelical DNA. *The Journal of Chemical Physics 123*, 124911.

[185] Mielke, S. P., Grønbech-Jensen, N., and Benham, C. (2008) Brownian dynamics

of double-stranded DNA in periodic systems with discrete salt. *Physical Review E 77*, 031924.

[186] Fyta, M. G., Melchionna, S., Kaxiras, E., and Succi, S. (2007) Multiscale coupling of molecular dynamics and hydrodynamics: application to DNA translocation through a nanopore. *Arxiv preprint physics/0701029*

[187] Carmesin, I., and Kremer, K. (1988) The bond fluctuation method: a new effective algorithm for the dynamics of polymers in all spatial dimensions. *Macromolecules 21*, 2819–2823.

[188] Somasi, M., Khomami, B., Woo, N. J., Hur, J. S., and Shaqfeh, E. S. G. (2002) Brownian dynamics simulations of bead-rod and bead-spring chains: numerical algorithms and coarse-graining issues. *Journal of Non-Newtonian Fluid Mechanics 108*, 227–255.

[189] Hsieh, C. C., Jain, S., and Larson, R. G. (2006) Brownian dynamics simulations with stiff finitely extensible nonlinear elastic-Fraenkel springs as approximations to rods in bead-rod models. *The Journal of Chemical Physics 124*, 044911.

[190] Bubis, R., Kantor, Y., and Kardar, M. (2009) Configurations of polymers attached to probes. *Europhysics Letters 88*, 48001.

[191] Chen, Y. L., Graham, M. D., de Pablo, J. J., Randall, G. C., Gupta, M., and Doyle, P. S. (2004) Conformation and dynamics of single DNA molecules in parallel-plate slit microchannels. *Physical Review E 70*, 60901.

[192] Chen, Y. L., Ma, H., Graham, M. D., and de Pablo, J. J. (2007) Modeling DNA in confinement: a comparison between the Brownian dynamics and lattice Boltzmann method. *Macromolecules 40*, 5978–5984.

[193] Iniesta, A., and de la Torre, J. G. (1989) A second-order algorithm for the simulation of the Brownian dynamics of macromolecular models. *The Journal of Chemical Physics 92*, 2015.

[194] Izmitli, A., Schwartz, D. C., Graham, M. D., and de Pablo, J. J. (2008) The effect of hydrodynamic interactions on the dynamics of DNA translocation through pores. *The Journal of Chemical Physics 128*, 085102.

[195] Jendrejack, R. M., De Pablo, J. J., and Graham, M. D. (2002) Stochastic simulations of DNA in flow: dynamics and the effects of hydrodynamic interactions. *The Journal of Chemical Physics 116*, 7752.

[196] Milchev, A., Binder, K., and Bhattacharya, A. (2004) Polymer translocation through a nanopore induced by adsorption: Monte Carlo simulation of a coarse-grained model. *The Journal of Chemical Physics 121*, 6042.

[197] Podtelezhnikov, A., and Vologodskii, A. (1997) Simulations of polymer cyclization by Brownian dynamics. *Macromolecules 30*, 6668–6673.

[198] Usta, O. B., Ladd, A. J. C., and Butler, J. E. (2005) Lattice-Boltzmann simulations of the dynamics of polymer solutions in periodic and confined geometries. *The Journal of Chemical Physics 122*, 094902.

[199] Thomas, M., and Davis, R. W. (1975) Studies on the cleavage of bacteriophage lambda DNA with EcoRI Restriction endonuclease. *Journal of Molecular Biology 91*, 315–320.

[200] Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., and Petersen, G. B. (1982) Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology 162*, 729–773.

[201] Slater, G. W., Holm, C., Chubynsky, M. V., De Haan, H. W., Dubé, A., Grass, K., Hickey, O. A., Kingsburry, C., Sean, D., and Shendruk, T. N. (2009) Modeling the separation of macromolecules: A review of current computer simulation methods. *Electrophoresis 30*, 792–818.

[202] Maciejczyk, M., Spasic, A., Liwo, A., and Scheraga, H. A. (2010) Coarse-grained model of nucleic acid bases. *Journal of Computational Chemistry 31*, 1644–1655.

[203] Gopal, S. M., Mukherjee, S., Cheng, Y. M., and Feig, M. (2010) PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins: Structure, Function, and Bioinformatics 78*, 1266–1281.

[204] Villa, E., Balaeff, A., and Schulten, K. (2005) Structural dynamics of the lac repressor–DNA complex revealed by a multiscale simulation. *Proceedings of the National Academy of Sciences of the United States of America 102*, 6783.

[205] Jian, H., Vologodskii, A. V., and Schlick, T. (1997) A combined wormlike-chain and bead model for dynamic simulations of long linear DNA. *Journal of Computational Physics 136*, 168–179.

[206] Podtelezhnikov, A. A., and Vologodskii, A. V. (2000) Dynamics of small loops in DNA molecules. *Macromolecules 33*, 2767–2771.

[207] Schlick, T., and Perišić, O. (2009) Mesoscale simulations of two nucleosome-repeat length oligonucleosomes. *Physical Chemistry Chemical Physics 11*, 10729–10737.

[208] Rapaport, D. *The Art of Molecular Dynamics Simulation*; Cambridge University Press, 2004.

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

[210] Allen, M. P., and Tildesley, D. J. *Molecular Simulations of Liquids*; Oxford University Press, Oxford, 1987.

[211] Haile, J. *Molecular Dynamics Simulation: Elementary Methods*; John Wiley & Sons, Inc., 1992.

[212] Zeng, Q. H., Yu, A. B., and Lu, G. Q. (2008) Multiscale modeling and simulation of polymer nanocomposites. *Progress in Polymer Science 33*, 191–269.

[213] McQuarrie, D. A. *Statistical Mechanics*; Harper and Row, New York, 1976.

[214] Litvinov, S., Hu, X. Y., and Adams, N. A. (2011) Numerical simulation of tethered DNA in shear flow. *Journal of Physics: Condensed Matter 23*, 184118.

[215] Pan, H., Ng, T. Y., Li, H., and Moeendarbary, E. (2010) Dissipative particle dynamics simulation of entropic trapping for DNA separation. *Sensors and Actuators A: Physical 157*, 328–335.

[216] Zuo, C. C., Ji, F., Cao, Q. Q., and Sun, X. D. Simulating stretching dynamics of DNA with dissipative particle dynamics. 2008.

[217] Fan, X., Phan-Thien, N., Chen, S., Wu, X., and Ng, T. Y. (2006) Simulating flow of DNA suspension using dissipative particle dynamics. *Physics of Fluids 18*, 063102.

[218] Symeonidis, V., Em Karniadakis, G., and Caswell, B. (2005) Dissipative particle dynamics simulations of polymer chains: scaling laws and shearing response compared to DNA experiments. *Physical Review Letters 95*, 76001.

[219] Watari, N., Makino, M., Kikuchi, N., Larson, R. G., and Doi, M. (2007) Simulation of DNA motion in a microchannel using stochastic rotation dynamics. *The Journal of Chemical Physics 126*, 094902.

[220] Chinappi, M., and De Angelis, E. (2011) Confined dynamics of a single DNA molecule. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 369*, 2329–2336.

[221] Chelakkot, R., Winkler, R. G., and Gompper, G. (2011) Semiflexible polymer conformation, distribution and migration in microcapillary flows. *Journal of Physics: Condensed Matter 23*, 184117.

[222] Coffey, W. T., Waldron, J. T., and Kalmykov, Y. P. *The Langevin Equation*; World Scientific, Singapore, 1996.

[223] Doi, M., and Edwards, S. F. *The Theory of Polymer Dynamics*; Oxford University Press, USA, 1988; Vol. 73.

[224] Ottinger, H. C. *Stochastic Processes in Polymeric Fluids*; Springer, Berlin, 1996.

[225] Fixman, M. (1978) Simulation of polymer dynamics. *The Journal of Chemical Physics 69*, 1527.

[226] Russel, W. B., Saville, D. A., and Schowalter, W. R. *Colloidal Dispersions*; Cambridge University Press, 1992.

[227] Newman, M. E. J., and Barkema, G. T. *Monte Carlo Methods in Statistical Physics*; Oxford University Press, 1999.

[228] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*, 1087.

[229] Jayaraman, A., Hall, C. K., and Genzer, J. (2006) Computer simulation study of molecular recognition in model DNA microarrays. *Biophysical Journal 91*, 2227–2236.

[230] Carlon, E., Orlandini, E., and Stella, A. L. (2002) Roles of stiffness and excluded volume in DNA denaturation. *Physical Review Letters 88*, 198101.

[231] Marenduzzo, D., Bhattacharjee, S. M., Maritan, A., Orlandini, E., and Seno, F. (2001) Dynamical scaling of the DNA unzipping transition. *Physical Review Letters 88*, 28102.

[232] Eddington, A. S. New pathways in science. 1935.

[233] Tidor, B., Irikura, K. K., Brooks, B. R., and Karplus, M. (1983) Dynamics of DNA oligomers. *Journal of Biomolecular Structure & Dynamics 1*, 231.

[234] Causo, M. S., Coluzzi, B., and Grassberger, P. (2000) Simple model for the DNA denaturation transition. *Physical Review E 62*, 3958–3973.

[235] Bird, R. B., and Wiest, J. M. (1995) Constitutive equations for polymeric liquids. *Annual Review of Fluid Mechanics 27*, 169–193.

[236] Ririe, K. M., Rasmussen, R. P., and Wittwer, C. T. (1997) Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Analytical Biochemistry 245*, 154–160.

[237] Lipsky, R. H., Mazzanti, C. M., Rudolph, J. G., Xu, K., Vyas, G., Bozak, D., Radel, M. Q., and Goldman, D. (2001) DNA melting analysis for detection of single nucleotide polymorphisms. *Clinical Chemistry 47*, 635.

[238] Sambriski, E. J., Ortiz, V., and de Pablo, J. J. (2009) Sequence effects in the melting and renaturation of short DNA oligonucleotides: structure and mechanistic pathways. *Journal of Physics: Condensed Matter 21*, 034105.

[239] Einstein, A. (1934) On the method of theoretical physics. *Philosophy of Science* 163–169.

[240] Boland, T., and Ratner, B. D. (1995) Direct measurement of hydrogen bonding in DNA nucleotide bases by atomic force microscopy. *Proceedings of the National Academy of Sciences of the United States of America 92*, 5297.

[241] Araque, J. C., Panagiotopoulos, A. Z., and Robert, M. A. (2011) Lattice model of oligonucleotide hybridization in solution. I. Model and thermodynamics. *The Journal of Chemical Physics 134*, 165103.

[242] Noller, H. F. (1984) Structure of ribosomal RNA. *Annual Review of Biochemistry 53*, 119–162.

[243] Brown, T., Hunter, W. N., Kneale, G., and Kennard, O. (1986) Molecular structure of the GA base pair in DNA and its implications for the mechanism of transversion mutations. *Proceedings of the National Academy of Sciences 83*, 2402.

[244] Prive, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopka, M. L., and Dickerson, R. E. (1987) Helix geometry, hydration, and GA mismatch in a B-DNA decamer. *Science 238*, 498.

[245] Heus, H. A., and Pardi, A. (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science 253*, 191.

[246] Li, Y., Zon, G., and Wilson, W. D. (1991) NMR and molecular modeling evidence for a GA mismatch base pair in a purine-rich DNA duplex. *Proceedings of the National Academy of Sciences 88*, 26.

[247] Cheng, J. W., Chou, S. H., and Reid, B. R. (1992) Base pairing geometry in GA mismatches depends entirely on the neighboring sequence. *Journal of Molecular Biology 228*, 1037–1041.

[248] Santa Lucia Jr, J., and Turner, D. H. (1993) Structure of $(rGGCGAGCC)_2$ in solution from NMR and restrained molecular dynamics. *Biochemistry 32*, 12612–12623.

[249] Pley, H. W., Flaherty, K. M., and McKay, D. B. (1994) Three-dimensional structure of a hammerhead ribozyme. *Nature 372*, 68–74.

[250] Greene, K. L., Jones, R. L., Li, Y., Robinson, H., Wang, A. H. J., Zon, G., and Wilson, W. D. (1994) Solution structure of a GA mismatch DNA sequence, d $(CCATGAATGG)_2$, determined by 2D NMR and structural refinement methods. *Biochemistry 33*, 1053–1062.

[251] Katahira, M., Kanagawa, M., Sato, H., Uesugi, S., Fujii, S., Kohno, T., and Maeda, T. (1994) Formation of sheared G: A base pairs in an RNA duplex modelled after ribozymes, as revealed by NMR. *Nucleic Acids Research 22*, 2752–2759.

[252] Kalnik, M. W., Kouchakdjian, M., Li, B. F. L., Swann, P. F., and Patel, D. J. (1988) Base pair mismatches and carcinogen-modified bases in DNA: an NMR study of G. cntdot. T and G. cntdot. O4meT pairing in dodecanucleotide duplexes. *Biochemistry 27*, 108–115.

[253] Hunter, W. N., Brown, T., Anand, N. N., and Kennard, O. (1986) Structure of an adenine ·cytosine base pair in DNA and its implications for mismatch repair. *Nature*

[254] Holbrook, S. R., Cheong, C., Tinoco, I., and Kim, S. H. (1991) Crystal structure of an RNA double helix incorporating a track of non-Watson–Crick base pairs. *Nature*

[255] Sasisekharan, V., Zimmerman, S., and Davies, D. R. (1975) The structure of helical 5'-guanosine monophosphate. *Journal of Molecular Biology 92*, 171–174.

[256] Borden, K. L. B., Jenkins, T. C., Skelly, J. V., Brown, T., and

Lane, A. N. (1992) Conformational properties of the G. cntdot. G mismatch in d(CGCGAATTGGCG)$_2$ determined by NMR. *Biochemistry 31*, 5411–5422.

[257] Kang, C. H., Zhang, X., Ratliff, R., Moyzis, R., and Rich, A. (1992) Crystal structure of four-stranded Oxytricha telomeric DNA. *Nature 356*, 126–131.

[258] Smith, F. W. (1992) Quadruplex structure of Oxytricha telomeric DNA oligonucleotides. *Nature 356*, 164–168.

[259] Raszka, M. (1974) Mononucleotides in aqueous solution. Proton magnetic resonance studies of amino groups. *Biochemistry 13*, 4616–4622.

[260] Hare, D. R., and Reid, B. R. (1986) Three-dimensional structure of a DNA hairpin in solution: two-dimensional NMR studies and distance geometry calculations on d(CGCGTTTTCGCG). *Biochemistry 25*, 5341–5350.

[261] Hare, D., Shapiro, L., and Patel, D. J. (1986) Wobble dG. ·dT pairing in right-handed DNA: solution conformation of the d(CGTGAATTCGCG) duplex deduced from distance geometry analysis of nuclear Overhauser effect spectra. *Biochemistry 25*, 7445–7456.

[262] Cheng, Y. K., and Pettitt, B. M. (1992) Hoogsteen versus reversed-Hoogsteen base pairing: DNA triple helixes. *Journal of the American Chemical Society 114*, 4465–4474.

[263] Chou, S. H., Cheng, J. W., and Reid, B. R. (1992) Solution structure of [d (ATGAGCGAATA)] 2:: Adjacent G: A mismatches stabilized by cross-strand base-stacking and BII phosphate groups. *Journal of Molecular Biology 228*, 138–155.

[264] Hunter, C. A. (1993) Sequence-dependent DNA structure. The role of base stacking interactions. *Journal of Molecular Biology 230*, 1025–1025.

[265] Hunter, C. A., and Lu, X. J. (1997) DNA base-stacking interactions: a comparison of theoretical calculations with oligonucleotide X-ray crystal structures1. *Journal of Molecular Biology 265*, 603–619.

[266] Kneale, G., Brown, T., Kennard, O., and Rabinovich, D. (1985) G· T base-pairs in a DNA helix: the crystal structure of d (GGGGTCCC). *Journal of Molecular Biology 186*, 805–814.

[267] Brown, T., Leonard, G. A., Booth, E. D., and Chambers, J. (1989) Crystal

structure and stability of a DNA duplex containing A (anti)· G (syn) base-pairs. *Journal of Molecular Biology 207*, 455–457.

[268] Ebel, S., Lane, A. N., and Brown, T. (1992) Very stable mismatch duplexes: structural and thermodynamic studies on tandem G. cntdot. A mismatches in DNA. *Biochemistry 31*, 12083–12086.

[269] Goodman, L., and Ts' O, P. (1974) Basic Principles in Nucleic Acid Chemistry. *Academic Press, New York 1*, 93.

[270] Hunter, C. A., and Sanders, J. K. M. (1990) The nature of. pi.-. pi. interactions. *Journal of the American Chemical Society 112*, 5525–5534.

[271] Doi, K., Haga, T., Shintaku, H., and Kawano, S. (2010) Development of coarse-graining DNA models for single-nucleotide resolution analysis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368*, 2615.

[272] Prevost, C., Louise-May, S., Ravishanker, G., Beveridge, D. L., and Lavery, R. (1993) Persistence analysis of the static and dynamical helix deformations of DNA oligonucleotides: application to the crystal structure and molecular dynamics simulation of d(CGCGAATTCGCG)$_2$. *Biopolymers 33*, 335–350.

[273] Manzini, G., Yathindra, N., and Xodo, L. E. (1994) Evidence for intramolecularly folded i-DNA structures in biologically relevant CCC-repeat sequences. *Nucleic Acids Research 22*, 4634.

[274] Odijk, T. (1977) Polyelectrolytes near rod limit. *Journal of Polymer Scence B 15*, 477–483.

[275] Skolnick, J., and Fixman, M. (1977) Electrostatic persistence length of a worm-like polyelectrolyte. *Macromolecules 10*, 944–948.

[276] Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation 4*, 435–447.

[277] Huppert, J. L. (2010) Structure, location and interactions of G-quadruplexes. *FEBS Journal 277*, 3452–3458.

[278] Thomas, L. (1980) On Science and Certainty. *Discover*

[279] Turek, C., and Gold, L. (1990) Systematic evolution of ligands by exponential

enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science 249*, 505–510.

[280] Ellington, A. D., and Szostak, J. W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature 346*, 818–822.

[281] Malmstöm, B. G. *Nobel Lectures in Chemistry (1981- 1990)*; World Scientific, Singapore, 1992; Vol. 6.

[282] Nykypanchuk, D., Maye, M. M., Van Der Lelie, D., and Gang, O. (2008) DNA-guided crystallization of colloidal nanoparticles. *Nature 451*, 549–552.

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

[284] Bunka, D. H. J., and Stockley, P. G. (2006) Aptamers come of age–at last. *Nature Reviews Microbiology 4*, 588–596.

[285] Nowakowski, J., Shim, P. J., Prasad, G. S., Stout, C. D., and Joyce, G. F. (1999) Crystal structure of an 82-nucleotide RNA-DNA complex formed by the 10-23 DNA enzyme. *Nature Structural Biology 6*, 151–156.

[286] Nowakowski, J., Shim, P. J., Joyce, G. F., and Stout, C. D. (1999) Crystallization of the 10-23 DNA enzyme using a combinatorial screen of paired oligonucleotides. *Acta Crystallographica Section D: Biological Crystallography 55*, 1885–1892.

[287] Baum, D. A., and Silverman, S. K. (2008) Deoxyribozymes: useful DNA catalysts in vitro and in vivo. *Cellular and Molecular Life Sciences 65*, 2156–2174.

[288] Dorsett, Y., and Tuschl, T. (2004) siRNAs: applications in functional genomics and potential as therapeutics. *Nature Reviews Drug Discovery 3*, 318–329.

[289] Tian, Y., and Mao, C. (2005) DNAzyme amplification of molecular beacon signal. *Talanta 67*, 532–537.

[290] Todd, A. V., Fuery, C. J., Impey, H. L., Applegate, T. L., and Haughton, M. A. (2000) DzyNA-PCR: use of DNAzymes to detect and quantify nucleic acid sequences in a real-time fluorescent format. *Clinical Chemistry 46*, 625.

[291] Buchhaupt, M., Peifer, C., and Entian, K. D. (2007) Analysis of 2'-O-methylated nucleosides and pseudouridines in ribosomal RNAs using DNAzymes. *Analytical Biochemistry 361*, 102–108.

[292] Hengesbach, M., Meusburger, M., Lyko, F., and Helm, M. (2008) Use of DNAzymes for site-specific analysis of ribonucleotide modifications. *RNA 14*, 180.

[293] Kwok, P. Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics 2*, 235–258.

[294] Chen, Y., and Mao, C. (2004) Putting a brake on an autonomous DNA nanomotor. *Journal of the American Chemical Society 126*, 8626–8627.

[295] Tian, Y., He, Y., Chen, Y., Yin, P., and Mao, C. (2005) A DNAzyme that walks processively and autonomously along a one-dimensional track. *Angewandte Chemie International Edition 44*, 4355–4358.

[296] Bishop, J. D., and Klavins, E. (2007) An improved autonomous DNA nanomotor. *Nano Letters 7*, 2574–2577.

[297] Reif, J. H., and Sahu, S. (2009) Autonomous programmable DNA nanorobotic devices using DNAzymes. *Theoretical Computer Science 410*, 1428–1439.

[298] Cairns, M. J., and Sun, L. Q. (2004) Nucleic acid sequence analysis using DNAzymes. *Methods in Molecular Biology 252*, 291–302.

[299] Todd, A., Fuery, C. J., and Cairns, M. J. Catalytic nucleic acid-based diagnostic methods, US Patent # 6361941. 2009.

[300] Mei, S. H. J., Liu, Z., Brennan, J. D., and Li, Y. (2003) An efficient RNA-cleaving DNA enzyme that synchronizes catalysis with fluorescence signaling. *Journal of the American Chemical Society 125*, 412–420.

[301] Li, J., and Lu, Y. (2000) A highly sensitive and selective catalytic DNA biosensor for lead ions. *Journal of the American Chemical Society 122*, 10466–10467.

[302] Bruesehoff, P. J., Li, J., Augustine, I., and Lu, Y. (2002) Improving metal ion specificity during in vitro selection of catalytic DNA. *Combinatorial Chemistry & High Throughput Screening 5*, 327–335.

[303] Liu, J., and Lu, Y. (2003) A colorimetric lead biosensor using DNAzyme-directed assembly of gold nanoparticles. *Journal of the American Chemical Society 125*, 6642–6643.

[304] Brown, A. K., Li, J., Caroline, M. B. P., and Lu, Y. (2003) A lead-dependent DNAzyme with a two-step mechanism. *Biochemistry 42*, 7152–7161.

[305] Wang, D. Y., and Sen, D. (2001) A novel mode of regulation of an RNA-cleaving DNAzyme by effectors that bind to both enzyme and substrate1. *Journal of Molecular Biology 310*, 723–734.

[306] Stojanovic, M. N., de Prada, P., and Landry, D. W. (2001) Catalytic molecular beacons. *ChemBioChem 2*, 411–415.

[307] Levy, M., and Ellington, A. D. (2002) ATP-dependent allosteric DNA enzymes. *Chemistry & Biology 9*, 417–426.

[308] Richards, J. L., Seward, G. K., Wang, Y. H., and Dmochowski, I. J. (2010) Turning the 10–23 DNAzyme on and off with light. *ChemBioChem 11*, 320–324.

[309] Cairns, M. J., Saravolac, E. G., and Sun, L. Q. (2002) Catalytic DNA: a novel tool for gene suppression. *Current Drug Targets 3*, 269–279.

[310] Zhang, X., Xu, Y., Ling, H., and Hattori, T. (1999) Inhibition of infection of incoming HIV-1 virus by RNA-cleaving DNA enzyme. *FEBS Letters 458*, 151–156.

[311] Sriram, B., and Banerjea, A. C. (2000) In vitro-selected RNA cleaving DNA enzymes from a combinatorial library are potent inhibitors of HIV-1 gene expression. *Biochemical Journal 352*, 667.

[312] Unwalla, H., and Banerjea, A. C. (2001) Inhibition of HIV-1 gene expression by novel macrophage-tropic DNA enzymes targeted to cleave HIV-1 TAT/Rev RNA. *Biochemical Journal 357*, 147.

[313] Dash, B. C., and Banerjea, A. C. (2004) Sequence-specific cleavage activities of DNA enzymes targeted against HIV-1 Gag and Nef regions. *Oligonucleotides 14*, 41–47.

[314] Unwalla, H., Chakraborti, S., Sood, V., Gupta, N., and Banerjea, A. C. (2006) Potent inhibition of HIV-1 gene expression and TAT-mediated apoptosis in human T cells by novel mono-and multitarget anti-TAT/Rev/Env ribozymes and a general purpose RNA-cleaving DNA-enzyme. *Antiviral Research 72*, 134–144.

[315] Sood, V., Gupta, N., Bano, A. S., and Banerjea, A. C. (2007) DNA-enzyme-mediated cleavage of human immunodeficiency virus type 1 Gag RNA is signif-

icantly augmented by antisense-DNA molecules targeted to hybridize close to the cleavage site. *Oligonucleotides 17*, 113–121.

[316] Sood, V., Unwalla, H., Gupta, N., Chakraborti, S., and Banerjea, A. C. (2007) Potent knock down of HIV-1 replication by targeting HIV-1 Tat/Rev RNA sequences synergistically with catalytic RNA and DNA. *AIDS 21*, 31.

[317] Jakobsen, M. R., Haasnoot, J., Wengel, J., Berkhout, B., and Kjems, J. (2007) Efficient inhibition of HIV-1 expression by LNA modified antisense oligonucleotides and DNAzymes targeted to functionally selected binding sites. *Retrovirology 4*, 29.

[318] Wo, J. E., Wu, X. L., Zhou, L. F., Yao, H. P., Chen, L. W., and Dennin, R. H. (2005) Effective inhibition of expression of hepatitis B virus genes by DNAzymes. *World Journal of Gastroenterology 11*, 3504.

[319] Hou, W., Ni, Q., Wo, J., Li, M., Liu, K., Chen, L., Hu, Z., Liu, R., and Hu, M. (2006) Inhibition of hepatitis B virus X gene expression by 10-23 DNAzymes. *Antiviral Research 72*, 190–196.

[320] Trepanier, J., Tanner, J. E., Momparler, R. L., Le, O. N. L., Alvarez, F., and Alfieri, C. (2006) Cleavage of intracellular hepatitis C RNA in the virus core protein coding region by deoxyribozymes. *Journal of Viral Hepatitis 13*, 131–138.

[321] Roy, S., Gupta, N., Subramanian, N., Mondal, T., Banerjea, A., and Das, S. (2008) Sequence-specific cleavage of hepatitis C virus RNA by DNAzymes: inhibition of viral RNA translation and replication. *Journal of General Virology 89*, 1579.

[322] Takahashi, H., Hamazaki, H., Habu, Y., Hayashi, M., Abe, T., Miyano-Kurosaki, N., and Takaku, H. (2004) A new modified DNA enzyme that targets influenza virus A mRNA inhibits viral infection in cultured cells. *FEBS Letters 560*, 69–74.

[323] Wu, S., Xu, J., Liu, J., Yan, X., Zhu, X., Xiao, G., Sun, L., and Tien, P. (2007) An efficient RNA-cleaving DNA enzyme can specifically target the 5?-untranslated region of severe acute respiratory syndrome associated coronavirus (SARS-CoV). *The Journal of Gene Medicine 9*, 1080–1086.

[324] Zhou, J., Yang, X. Q., Xie, Y. Y., Zhao, X. D., Jiang, L. P., Wang, L. J., and

Cui, Y. X. (2007) Inhibition of respiratory syncytial virus of subgroups A and B using deoxyribozyme DZ1133 in mice. *Virus Research 130*, 241–248.

[325] Cairns, M. J., Hopkins, T. M., Witherington, C., Wang, L., and Sun, L. Q. (1999) Target site selection for an RNA-cleaving catalytic DNA. *Nature Biotechnology 17*, 480–486.

[326] Schubert, S., Furste, J. P., Werk, D., Grunert, H. P., Zeichhardt, H., Erdmann, V. A., and Kurreck, J. (2004) Gaining target access for deoxyribozymes. *Journal of Molecular Biology 339*, 355–363.

[327] Lu, Z. X., Ye, M., Yan, G. R., Li, Q., Tang, M., Lee, L. M., Sun, L. Q., and Cao, Y. (2005) Effect of EBV LMP1 targeted DNAzymes on cell proliferation and apoptosis. *Cancer Gene Therapy 12*, 647–654.

[328] Li, J., Zhu, D., Yi, Z., He, Y., Chun, Y., Liu, Y., and Li, N. (2005) DNAzymes targeting the icl gene inhibit ICL expression and decrease Mycobacterium tuberculosis survival in macrophages. *Oligonucleotides 15*, 215–222.

[329] Chen, F., Wang, R., Li, Z., Liu, B., Wang, X., Sun, Y., Hao, D., and Zhang, J. (2004) A novel replicating circular DNAzyme. *Nucleic Acids Research 32*, 2336.

[330] Chen, F., Li, Z., Wang, R., Liu, B., Zeng, Z., Zhang, H., and Zhang, J. (2004) Inhibition of Ampicillin-Resistant Bacteria by Novel Mono-DNAzymes and Di-DNAzyme Targeted to β-Lactamase mRNA. *Oligonucleotides 14*, 80–89.

[331] Hou, Z. (2007) Inhibition of β-lactamase-mediated oxacillin resistance in Staphylococcus aureus by a deoxyribozyme1. *Acta Pharmacologica Sinica 28*, 1775–1782.

[332] Schubert, S., and Kurreck, J. (2004) Ribozyme-and deoxyribozyme-strategies for medical applications. *Current Drug Targets 5*, 667–681.

[333] Mitchell, A., Dass, C. R., Sun, L. Q., and Khachigian, L. M. (2004) Inhibition of human breast carcinoma proliferation, migration, chemoinvasion and solid tumour growth by DNAzymes targeting the zinc finger transcription factor EGR-1. *Nucleic Acids Research 32*, 3065.

[334] Liang, Z., Wei, S., Guan, J., Luo, Y., Gao, J., Zhu, H., Wu, S., and Liu, T. (2005) DNAzyme-mediated cleavage of survivin mRNA and inhibition of the growth of PANC-1 cells. *Journal of Gastroenterology and Hepatology 20*, 1595–1602.

[335] Kuwabara, T., Warashina, M., Tanabe, T., Tani, K., Asano, S., and Taira, K. (1997) Comparison of the specificities and catalytic activities of hammerhead ribozymes and DNA enzymes with respect to the cleavage of BCR-ABL chimeric L6 (b2a2) mRNA. *Nucleic Acids Research 25*, 3074.

[336] Warashina, M., Kuwabara, T., Nakamatsu, Y., and Taira, K. (1999) Extremely high and specific activity of DNA enzymes in cells with a Philadelphia chromosome. *Chemistry & Biology 6*, 237–250.

[337] Wu, Y., Yu, L., McMahon, R., Rossi, J. J., Forman, S. J., and Snyder, D. S. (1999) Inhibition of bcr-abl oncogene expression by novel deoxyribozymes (DNAzymes). *Human Gene Therapy 10*, 2847–2857.

[338] Kabuli, M., Yin, J. A., and Tobal, K. (2004) Targeting PML/RARa transcript with DNAzymes results in reduction of proliferation and induction of apoptosis in APL cells. *Hematology Journal 5*, 426–433.

[339] Seifert, G., Taube, T., Paal, K., von Einsiedel, H. G., Wellmann, S., Henze, G., Seeger, K., Schroff, M., and Wittig, B. (2006) Brief communication: stability and catalytic activity of novel circular DNAzymes. *Nucleosides, Nucleotides and Nucleic Acids 25*, 785–793.

[340] Dass, C. R., Friedhuber, A. M., Khachigian, L. M., Dunstan, D. E., and Choong, P. F. M. (2008) Biocompatible chitosan-DNAzyme nanoparticle exhibits enhanced biological activity. *Journal of Microencapsulation 25*, 421–425.

[341] Ackermann, J. M., Kanugula, S., and Pegg, A. E. (2005) DNAzyme-mediated silencing of ornithine decarboxylase. *Biochemistry 44*, 2143–2152.

[342] Sioud, M., and Leirdal, M. (2000) Design of nuclease resistant protein kinase c [alpha] DNA enzymes with potential therapeutic application1. *Journal of Molecular Biology 296*, 937–947.

[343] Liu, C., Cheng, R., Sun, L., and Tien, P. (2001) Suppression of platelet-type 12-lipoxygenase activity in human erythroleukemia cells by an RNA-cleaving DNAzyme. *Biochemical and Biophysical Research Communications 284*, 1077–1082.

[344] Zhang, L., Gasper, W. J., Stass, S. A., Ioffe, O. B., Davis, M. A., and Mixson, A. J. (2002) Angiogenic inhibition mediated by a DNAzyme that targets vascular endothelial growth factor receptor 2. *Cancer Research 62*, 5463.

[345] Cieslak, M., Niewiarowska, J., Nawrot, M., Koziolkiewicz, M., Stec, W., and Cierniewski, C. (2002) DNAzymes to β1 and β3 mRNA down-regulate expression of the targeted integrins and inhibit endothelial cell capillary tube formation in fibrin and matrigel. *Journal of Biological Chemistry 277*, 6779.

[346] Cieslak, M., Szymanski, J., Adamiak, R. W., and Cierniewski, C. S. (2003) Structural rearrangements of the 10–23 DNAzyme to β3 integrin subunit mRNA induced by cations and their relations to the catalytic activity. *Journal of Biological Chemistry 278*, 47987.

[347] Dass, C. R. (2004) Deoxyribozymes: cleaving a path to clinical trials. *Trends in Pharmacological Sciences 25*, 395–397.

[348] Sun, L. Q., Cairns, M. J., Gerlach, W. L., Witherington, C., Wang, L., and King, A. (1999) Suppression of smooth muscle cell proliferation by a c-myc RNA-cleaving deoxyribozyme. *Journal of Biological Chemistry 274*, 17236.

[349] Dass, C. R., Saravolac, E. G., Li, Y., and Sun, L. Q. (2002) Cellular uptake, distribution, and stability of 10-23 deoxyribozymes. *Antisense and Nucleic Acid Drug Development 12*, 289–299.

[350] Zhang, G., Dass, C. R., Sumithran, E., Di Girolamo, N., Sun, L. Q., and Khachigian, L. M. (2004) Effect of deoxyribozymes targeting c-Jun on solid tumor growth and angiogenesis in rodents. *Journal of the National Cancer Institute 96*, 683.

[351] Fahmy, R. G., Waldman, A., Zhang, G., Mitchell, A., Tedla, N., Cai, H., Geczy, C. R., Chesterman, C. N., Perry, M., and Khachigian, L. M. (2006) Suppression of vascular permeability and inflammation by targeting of the transcription factor c-Jun. *Nature Biotechnology 24*, 856–863.

[352] Santiago, F. S., Lowe, H. C., Kavurma, M. M., Chesterman, C. N., Baker, A., Atkins, D. G., and Khachigian, L. M. (1999) New DNA enzyme targeting Egr-1 mRNA inhibits vascular smooth muscle proliferation and regrowth after injury. *Nature Medicine 5*, 1265.

[353] Khachigian, L. M. (2000) Catalytic DNAs as potential therapeutic agents and sequence-specific molecular tools to dissect biological function. *Journal of Clinical Investigation 106*, 1189–1196.

[354] Lowe, H. C., Fahmy, R. G., Kavurma, M. M., Baker, A., Chesterman, C. N.,

and Khachigian, L. M. (2001) Catalytic oligodeoxynucleotides define a key regulatory role for early growth response factor-1 in the porcine model of coronary in-stent restenosis. *Circulation Research* 200109786.

[355] Santiago, F. S., and Khachigian, L. M. (2001) Nucleic acid based strategies as potential therapeutic tools: mechanistic considerations and implications to restenosis. *Journal of Molecular Medicine 79*, 695–706.

[356] Bittker, J. A., Phillips, K. J., and Liu, D. R. (2002) Recent advances in the in vitro evolution of nucleic acids. *Current Opinion in Chemical Biology 6*, 367–374.

[357] Chaudhury, I., Raghav, S. K., Gautam, H. K., Das, H. R., and Das, R. H. (2006) Suppression of inducible nitric oxide synthase by 10-23 DNAzymes in murine macrophage. *FEBS letters 580*, 2046–2052.

[358] Dass, C. R., Choong, P. F. M., and Khachigian, L. M. (2008) DNAzyme technology and cancer therapy: cleave and let die. *Molecular Cancer Therapeutics 7*, 243.

[359] Yu, S. H., Wang, T. H., and Au, L. C. (2009) Specific repression of mutant K-RAS by 10-23 DNAzyme: sensitizing cancer cell to anti-cancer therapies. *Biochemical and Biophysical Research Communications 378*, 230–234.

[360] Yen, L., Strittmatter, S. M., and Kalb, R. G. (1999) Sequence-specific cleavage of Huntingtin mRNA by catalytic DNA. *Annals of Neurology 46*, 366–373.

[361] Hjiantoniou, E., Iseki, S., Uney, J. B., and Phylactou, L. A. (2003) DNazyme-mediated cleavage of Twist transcripts and increase in cellular apoptosis. *Biochemical and Biophysical Research Communications 300*, 178–181.

[362] Yuan, B. F., Xue, Y., Luo, M., Hao, Y. H., and Tan, Z. (2007) Two DNAzymes targeting the telomerase mRNA with large difference in Mg2+ concentration for maximal catalytic activity. *The International Journal of Biochemistry & Cell Biology 39*, 1119–1129.

[363] Sel, S., Wegmann, M., Dicke, T., Sel, S., Henke, W., Yildirim, A., Renz, H., and Garn, H. (2008) Effective prevention and therapy of experimental allergic asthma using a GATA-3-specific DNAzyme. *Journal of Allergy and Clinical Immunology 121*, 910–916.

[364] Khachigian, L. M., Fahmy, R. G., Zhang, G., Bobryshev, Y. V., and Ka-

niaros, A. (2002) c-Jun regulates vascular smooth muscle cell growth and neointima formation after arterial injury. *Journal of Biological Chemistry 277*, 22985.

[365] Grimpe, B., Dong, S., Doller, C., Temple, K., Malouf, A., and Silver, J. (2002) The critical role of basement membrane-independent laminin $\gamma 1$ chain during axon regeneration in the CNS. *The Journal of Neuroscience 22*, 3144.

[366] Grimpe, B., and Silver, J. (2004) A novel DNA enzyme reduces glycosaminoglycan chains in the glial scar and allows microtransplanted dorsal root ganglia axons to regenerate beyond lesions in the spinal cord. *The Journal of Neuroscience 24*, 1393.

[367] Kurreck, J., Bieber, B., Jahnel, R., and Erdmann, V. (2002) Comparative study of DNA enzymes and ribozymes against the same full-length messenger RNA of the vanilloid receptor subtype I. *Journal of Biological Chemistry 277*, 7099.

[368] Isaka, Y., Nakamura, H., Mizui, M., Takabatake, Y., Horio, M., Kawachi, H., Shimizu, F., Imai, E., and Hori, M. (2004) DNAzyme for TGF-β suppressed extracellular matrix accumulation in experimental glomerulonephritis. *Kidney International 66*, 586–590.

[369] Silverman, S. K. (2005) In vitro selection, characterization, and application of deoxyribozymes that cleave RNA. *Nucleic Acids Research 33*, 6151.

[370] Cairns, M. J., Hopkins, T. M., Witherington, C., and Sun, L. Q. (2000) The influence of arm length asymmetry and base substitution on the activity of the 10-23 DNA enzyme. *Antisense and Nucleic Acid Drug Development 10*, 323–332.

[371] Asanuma, H., Hayashi, H., Zhao, J., Liang, X., Yamazawa, A., Kuramochi, T., Matsunaga, D., Aiba, Y., Kashida, H., and Komiyama, M. (2006) Enhancement of RNA cleavage activity of 10–23 DNAzyme by covalently introduced intercalator. *Chemical Communications* 5062–5064.

[372] Zaborowska, Z. (2002) Sequence requirements in the catalytic core of the "10-23" DNA enzyme. *Journal of Biological Chemistry 277*, 40617.

[373] Zaborowska, Z., Schubert, S., Kurreck, J., and Erdmann, V. A. (2005) Deletion analysis in the catalytic region of the 10-23 DNA enzyme. *FEBS Letters 579*, 554–558.

[374] Nawrot, B., Widera, K., Wojcik, M., Rebowska, B., Nowak, G., and Stec, W.

(2007) Mapping of the functional phosphate groups in the catalytic core of deoxyribozyme 10–23. *FEBS Journal 274*, 1062–1072.

[375] Brooks, B. R., Bruccoleri, R. E., and Olafson, B. D. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry 4*, 187–217.

[376] Brooks, B. R., Bruccoleri, R. E., and Olafson, B. D. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry 4*, 187–217.

[377] MacKerell Jr, A. D., Brooks, B., Brooks III, C. L., Nilsson, L., Roux, B., Won, Y., and Karplus, M. CHARMM: the energy function and its parameterization. 1998.

[378] Brooks, B. R., Brooks III, C. L., Mackerell Jr, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., and Boresch, S. (2009) CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry 30*, 1545–1614.

[379] Banavali, N. K., and MacKerell Jr, A. D. (2002) Free energy and structural pathways of base flipping in a DNA GCGC containing sequence. *Journal of Molecular Biology 319*, 141–160.

[380] Geyer, C. R., and Sen, D. (1997) Evidence for the metal-cofactor independence of an RNA phosphodiester-cleaving DNA enzyme. *Chemistry & Biology 4*, 579–593.

[381] Fulbright, J. *Old Myths and New Realities, and Other Commentaries*; Random House, 1964; Vol. 264.

[382] Subcommittee on Human Genome of the Health and Environmental Research Advisory Committee, Report on the Human Genome Initiative for the Office of Health and Environmental Research. 1987.

[383] Bennett, S. T., Barnes, C., Cox, A., Davies, L., and Brown, C. (2005) Toward the $1000 human genome. *Pharmacogenomics 6*, 373–382.

[384] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. (2001) Initial sequencing and analysis of the human genome. *Nature 409*, 860–921.

[385] Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H.,

Ch'ang, L. Y., Huang, W., Liu, B., and Shen, Y. (2003) The international HapMap project. *Nature 426*, 789–796.

[386] Shaffer, C. (2007) Next-generation sequencing outpaces expectations. *Nature Biotechnology 25*, 149–149.

[387] Genetic Engineering and Biotechnology News, *NHGRI gives $1,000 Genome Project a $14M shot in the arm.* 23 August 2011.

[388] Pollack, A. *Company unveils DNA sequencing device meant to be portable, disposable, and cheap.* New York Times. 18 February 2012.

[389] Monegain, B. *Personalized medicine market in growth mode.* Healthcare IT News. 08 March 2012

[390] Hudson, K. L., Holohan, M. K., and Collins, F. S. (2008) Keeping pace with the times—the Genetic Information Nondiscrimination Act of 2008. *New England Journal of Medicine 358*, 2661–2663.

[391] Korobkin, R., and Rajkumar, R. (2008) The Genetic Information Nondiscrimination Act—A half-step toward risk sharing. *New England Journal of Medicine 359*, 335–337.

[392] Schlein, D. (2008) New frontiers for genetic privacy law: the Genetic Information Nondiscrimination Act of 2008. *George Mason University Civil Rights Law Journal 19*, 311.

[393] Rothstein, M. A. (2008) Putting the genetic information nondiscrimination act in context. *Genetics in Medicine 10*, 655.

[394] Slaughter, L. M. (2008) The Genetic Information Nondiscrimination Act: why your personal genetics are still vulnerable to discrimination. *Surgical Clinics of North America 88*, 723–738.

[395] Robertson, J. A. (2003) The $1000 genome: ethical and legal issues in whole genome sequencing of individuals. *American Journal of Bioethics 3*, 35–42.

[396] McGuire, A. L., Caulfield, T., and Cho, M. K. (2008) Research ethics and the challenge of whole-genome sequencing. *Nature Reviews Genetics 9*, 152–156.

[397] U.S. Census Bureau, *2012 Statistical Abstract. Health and Nutrition: Health Care Resources.* 2012.

[398] Hampton, T. (2006) Cancer Genome Atlas. *The Journal of the American Medical Association 296*, 1958–1958.

[399] Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., and Gerhard, D. S. (2010) International network of cancer genome projects. *Nature 464*, 993–998.