Mechanisms behind Protein-DNA Interactions Unveiled with Molecular Simulation and Atomic Force Microscopy

Elliot Weising Chan

Doctor of Philosophy

University of York
Physics

June 2023

Abstract

DNA specificity underlines the fundamental basis through which many different proteins perform their myriad roles within organisms, a key example being the genome structuring performed by the nucleoid-associated proteins (NAPs). This thesis presents results found through the usage of all-atom molecular dynamics (MD) and in-liquid atomic force microscopy (AFM) towards understanding DNA-protein interactions for a variety of these systems.

One of the most abundant NAPs is the histone-like protein from *E. coli* strain U93 (HU), which specifically binds to sites of DNA damage and creates sharp bends to aid in DNA repair. Here, simulations reveal how the protein diffuses along and between strands of DNA towards finding a binding site. Once a site of damage is found, we show a clear multimodality in the binding of HU to DNA, observed in both MD and AFM. AFM imaging also shows aggregation of DNA by HU, which simulations show to be highly energetically favourable, explaining how HU condenses the nucleoid, and why it is key to biofilm stability.

A pair of evolutionarily related proteins, ParB and Noc, have been used to study how specificity evolved in order to allow new regulatory functions to be fulfilled. The use of MD here explained the roles of key mutations unable to be sampled in experiments, and *in-silico* mutagenesis was applied to map out all potential pathways from changing amino acids and nucleotides.

Beyond the specific recognition of DNA by proteins, a mechanism through which proteins can be mechanically encapsulated by a DNA origami structure has also been studied. A new methodology to simulate such structures has been developed, and then applied to understand the interactions of GFP with the side of a structure, revealing strong, non-specific interactions between the protein and the DNA that would allow the protein to remain caught within a DNA box.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisors, Agnes and Mark, for their neverending support and guidance over these four years. I have been given the freedom to pursue paths and projects that steer far from my original project, things some might describe as “brave” or “unwise” for someone with my academic background, but you both not only allowed, but fully supported me in these endeavours. This all began when I stopped Agnes in a stairway as a second year undergraduate, and despite us having never met before, Agnes was immediately receptive and willing to work with me. I consider that amongst the most important moments I’ve had in seven years at York, and will be forever grateful for that, the unparalleled kindness, and everything Agnes has done for me, thank you!

This project would not have been possible without the extensive help of many, many people. I must thank Drs George Watson-Hyde and Sam Yoshua, both of whom helped me immensely and are the starting point of my project. In particular, I must thank George for first introducing me to biophysics and answering the many annoying questions from a keen undergraduate. Dr Jamieson Howard was also key to the project, having helped a theoretical physicist perform biology wetlab experiments. Also to Matteo Marozzi, learning AFM at the same time as you made the experience enjoyable (or at least bearable during the many times the machine refused to work), and I’m glad my first “independent” collaboration was with you.

To everyone I met in Sheffield, Drs Alice Pyne and Robert Moorehead, Eddie, Jean, Max, Libby, Sylvia, thank you for helping me so much throughout this whole project, for putting up with my continuous visits and questions, for the many interesting dicussions (both about work and not) and for all the encouragement over the last two years.

There are many friends, in and out of York, that have made this experience what it is, frankly too many to list. To Matt, Matteo and Emma, the P/C/011 antics were the perfect way to have started my PhD, and I will remember the jokes for a long time. To Seb and Monica, thanks for all the games, sorry for the salt! To Lara, thanks for our many shared rants over the last 4 years. To Devious Little Fella, Bestie and ™, the games nights are what kept me sane throughout writing and all I can say is I wish we had started sooner and had longer. You all, and everyone I reguarly interacted with in the Physics of Life group, are what made the last 4 years.

And lastly, to my parents, without your support (often financial!) this would not have been possible.

Declarations

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

- The construction of the DNA samples used was performed under the supervision of Dr. Jamieson Howard.
- The HU used in the experiments was provided by Dr. Michelle Hawkins.
- The code to calculate bend angles from AFM data was written by Mingxue Du.
- The AFM data was gathered in collaboration with Daniel Rollins.
- The crystal structures of the ParB-*parS* and Noc-*NBS* protein-DNA complexes were resolved and provided by Dr. Adam Jalal and Dr. Tung Le.
- DNA origami structures constructed via caDNAno and oxDNA simulations were provided by Matteo Marozzi.

Publications

EW Chan, JAL Howard, M Du, DE Rollins, ALB Pyne, MC Leake, A Noy,"The mechanism of action of the bacterial protein HU for DNA repair", *In Prep.*

GD Watson, EW Chan, MC Leake, A Noy, "Structural interplay between DNA-shape protein recognition and supercoiling: the case of IHF", *Computational and Structural Biotechnology Journal.* **20**, 5264-5274 (2022)

ASB Jalal, NT Tran, CE Stevenson, EW Chan, R Lo, Xiao Tan, A Noy, DM Lawson, TBK Le, "Diversification of DNA-binding specificity via permissive and specificity-switching mutations in the ParB/Noc protein family", *Cell Reports.* **32**, 107928 (2020)

Chapter 1

Introduction

1.1 Nucleic acids

Nucleic acids are biological macromolecules which are responsible for the storage and transmission of genetic information in all known life. The "central dogma" of molecular biology [1] is the process through which nucleic acids cause the production of proteins, which thus leads to the functioning of living organisms. There are two different types of nucleic acids, *ribonucleic acid* (RNA) and *deoxyribonucleic acid* (DNA), though both are of similar composition, being a sequence of nucleotides comprising of a pentose sugar, a nitrogenous base and a negatively-charged phosphate group. The difference is in the sugar, in which RNA has ribose, $C_5H_{10}O_5$ whilst DNA has deoxyribose, $C_5H_{10}O_4$.

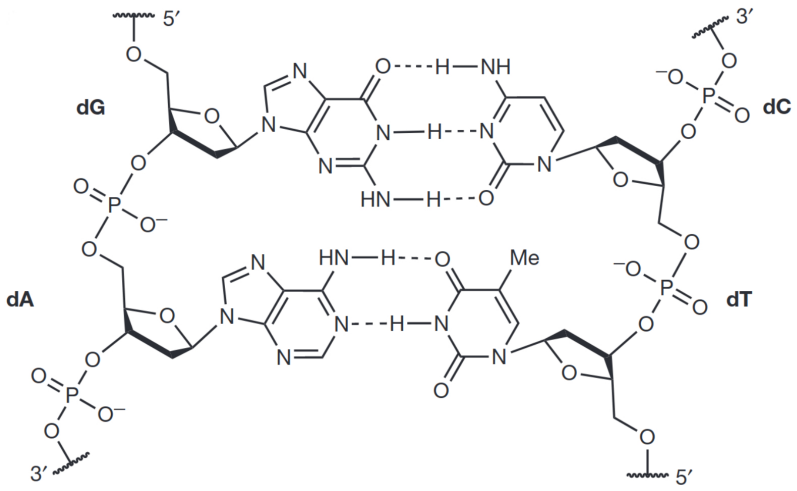


Figure 1.1: The structure of DNA, showing base pairing and the phosphate backbone (image from [2])

There are 5 nucleobases, in DNA these are adenine (A), cytosine (C),

guanine (G) and thymine (T), whilst in RNA thymine is swapped for uracil (U). These nucleobases are the only difference between nucleotides, with the deoxyribose (or ribose for RNA) and phosphate group being the same for all nucleotides. As seen in figure 1.1, thymine and cytosine are six-membered rings (known as pyrimidines) whilst adenine and guanine are fused aromatic rings (known as purines).

Of more relevance in grouping the nucleobases is the number of hydrogen bonds each is able to form. Adenine and thymine are capable of forming two hydrogen bonds, whilst guanine and cytosine can form three, and their donor and acceptor atom locations are reversed. Due to this, base pairing is limited to the A-T and C-G bonds, which results in base pairs each containing one purine and one pyrimidine and so being roughly equal in size. As it has fewer hydrogen bonds, the A-T bond is weaker than the C-G bond, and thus AT-rich sequences of DNA can be broken apart more easily, though it has been shown that base stacking interactions contribute more to the stability of the DNA double helix than base pairing does [3].

Two key functions of DNA are transcription and replication. Transcription is the process through which genes coded in the DNA are copied into messenger-RNA by the enzyme RNA polymerase. The produced RNA strand contains the template for the synthesis of proteins by the ribosome. Replication is the process through which DNA is cloned into new and identical molecules — a crucial process in cell division to ensure the integrity of the daughter cell.

1.1.1 DNA structure

As DNA is a double helix, a large number of parameters must be defined for the accurate description of it geometry. These exist for each base pair (with this set being known as the base-pair parameters) and base-pair step (the base-step parameters) [4] and are shown in figure 1.2.

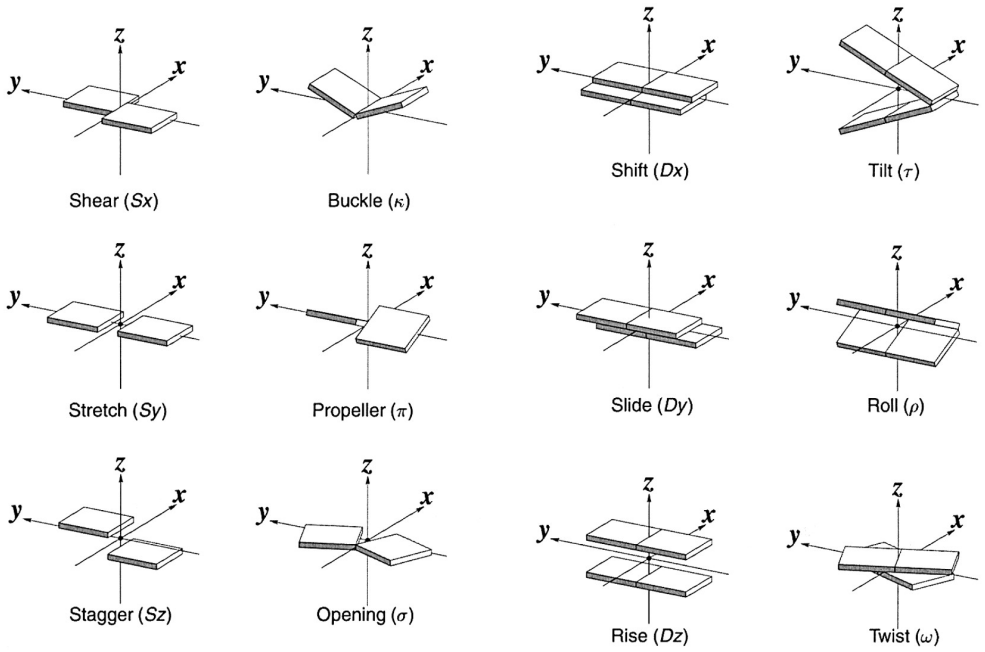


Figure 1.2: The base pair and base-step parameters (adapted from [5])

Both sets of parameters are composed of 3 translations and 3 rotations. In the base pair parameters, shear, stretch and stagger translations describe movements in x, y and z respectively, with buckle, propeller twist and opening being the equivalent for rotations. For base-step parameters, shift, slide and rise represent the displacements whilst tilt, roll and twist the rotations, again in x, y and z respectively.

DNA in cells is typically divided into two forms, called A-DNA and B-DNA. These form right-handed helices, though left-handed helices such as in Z-DNA are also capable of forming [6]. B-DNA is the usual conformation found *in vivo*, however Z-DNA, despite being notably rarer, does have biological importance, with proteins existing which bind to Z-DNA yet not B-DNA [7]. A-DNA is typically formed during dehydrating conditions, whilst Z-DNA can appear in high salt or during negative supercoiling (though this is often energetically unfavourable). The key properties of these three forms of DNA are described in table 1.1.

Parameter	A-DNA	B-DNA	Z-DNA
Base pairs per turn	11	10.5	12
Rotation / $^{\circ}\text{bp}^{-1}$	32.7	34.3	30
Rise along axis / Å bp^{-1}	2.3	3.32	3.8
Diameter / Å	23	20	18

Table 1.1: Structural parameters of A-, B- and Z-DNA.

Two interesting parameters to note are the major and minor groove width of DNA — as the strands of the double helix aren't directly opposing one another the result is that the grooves between strands are of different sizes. In B-DNA the average major and minor groove widths are 22Å and 12Å wide [8]. Due to this, the bases at the major groove are easier to access, which promotes readout for many DNA-binding proteins [9], though the minor groove is a common target for many small ligands such as DNA-binding dyes [10].

1.1.2 DNA damage

DNA itself isn't immune to changes or damage, it is quite sensitive to its surroundings. Since before even the composition of DNA was known, it was shown that X-rays had negative effects on chromosomes [11]. Different ways in which DNA can be damaged are classified as either endogenous or environmental — either from internal or external activities [12].

Replication errors are an example of endogenous damage, whereby the wrong bases may be incorporated during replication. Another examples of endogenous damage is that induced by reactive oxygen species, chemical compounds which contain reactive oxygen atoms such as hydroxyl radical. Hydroxyl radicals are produced by immunological responses and have been shown to be capable of causing double strand breaks in DNA [13]. Spontaneous changes in pH or temperature is another source, causing the nucleobase cytosine to lose the amino groups [14].

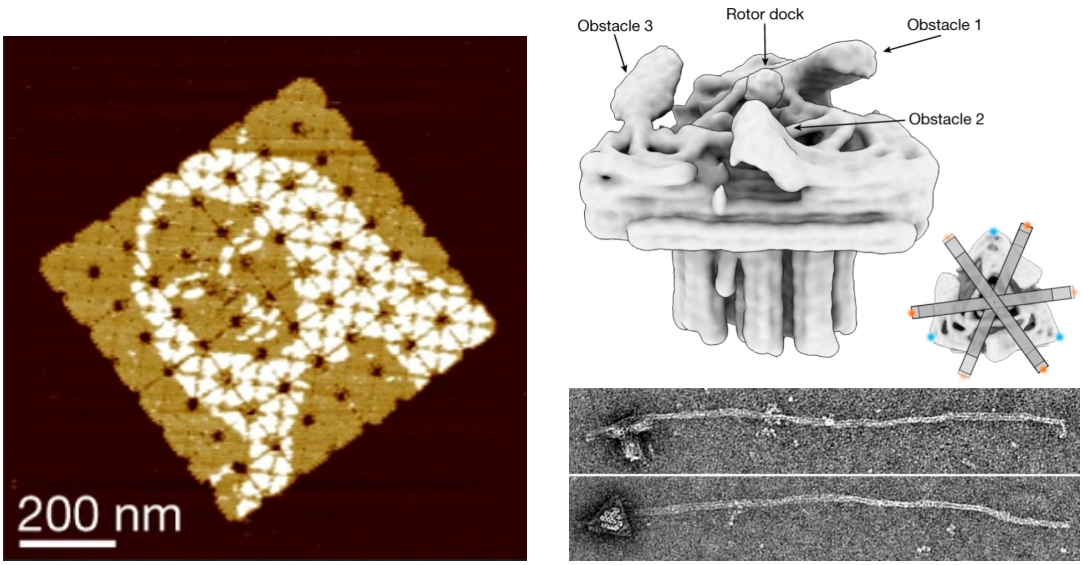
Meanwhile extracellular agents can also deal serious damage to DNA and are a threat to genomic stability. Prominently, drugs and ionising radiation are key examples [15]. Radiation particularly has been shown to induce a large variety of lesions in DNA [16]. Types of damage range of base mismatches to single-strand breaks, and flipped or unpaired bases in the double helix, to double-strand breaks from lesions.

1.1.3 DNA origami

By using the rules of DNA pairing, one can create 2D and 3D structures through a technique known as DNA origami. This has advanced the field of DNA nanotechnology, which takes the molecule out of its biological context to assemble structural motifs through a 'bottom-up' approach [17] for technological uses. This has applications in a wide range of fields, initially thought up to allow easier X-ray crystallography and nuclear magnetic resonance of proteins (via creating highly ordered lattices of DNA containing gaps with protein recognition sites, creating a periodic structure of DNA and proteins), there has recently been a lot of attention for nanomedicine and the use of DNA boxes for smart drug delivery [18].

In 1982, Nadrian Seeman suggested that DNA's specific pairing could be exploited to create lattices from DNA junctions [19], and later experimentally verified that lattices [20] and cubes [21] could be created. However, this method required 1:1 stoichiometry between strands, with even small deviations causing a low yield which reduced its efficiency and effectiveness for DNA nanotechnology. DNA origami as is used today was truly developed in 2006, when Paul Rothmund published methods taking advantage of Watson-Crick base pairing that allows two complementary DNA strands to bind with high specificity, with the use of "staple" strands which are complementary to spatially separated regions of a "scaffold" strand [22]. By designing a large number of staple and scaffold strands, a large scale structure can be formed. Fortunately, the development of new tools like caDNAno [23] has made the designing of these structures significantly easier, no longer needing to be done by hand.

DNA origami has represented an exciting new avenue of DNA nanotechnology, which as originated a diverse set of structures such as highly detailed sub-1μm [24] designs, encapsulation of proteins via DNA boxes [25], and a rotary ratchet motor [26].



(a) Atomic force microscopy image of a DNA origami construct designed with a pattern in mind (The Mona Lisa here) [24].

(b) Cryo-EM of the "motor block" of a DNA origami rotor, with TEM images of the motor with a rotor arm attached [26].

Figure 1.3: Examples of recent developments using DNA origami as a tool for designing DNA nanostructures for nanotechnology.

1.2 Nucleoid-Associated Proteins

Proteins are biomolecules made up of chains of amino acid residues. These molecules are responsible for a vast number of biological functions including responding to stimuli, DNA replication, providing structure to cells and the transport of molecules. Chief amongst these roles is catalysing biochemical reactions, performed by a subsection of proteins known as enzymes.

Whilst eukaryotes have compartmentalised cells, prokaryotes do not. They lack a cell nucleus, instead having a irregularly shaped region known as the nucleoid, which houses the cell's DNA. There are many forces at play within the nucleoid, without which the DNA would not fit within the cell — for example, *E. coli* has a genome of 4.6 Mbp [27], with a contour length of 1.5 mm, yet the cell itself is only 2 µm in length. It is thus clear that a large amount of compaction must take place. However, it must be done in a controllable manner, such that key parts of the genome remain accessible.

There are various means through which this compaction occurs. Firstly, in most prokaryotes, DNA is a covalently closed circle [28] (as opposed to linear as in eukaryotes) known as a plasmid, which allows the DNA to form higher-order helical structures through a process called supercoiling [29].

In addition, a group of DNA binding proteins, known as nucleoid-associated proteins (NAPs), help to stabilise these structures, through the bending, bridging and wrapping of DNA. As well as this genome structuring, NAPs also aid in a wide repertoire of biological functions like gene regulation and expression. NAPs can perform their role with varying degrees of specificity depending on their concentrations in cells, which changes throughout the cell cycle and environmental conditions. At low concentrations, highly specific interactions are seen, whereby the NAP acts as a regulator in DNA transactions (such as replication, transcription, or recombination). Meanwhile, at high NAP concentrations, weak non-specific interactions dominate which act to condense the bacterial genome [30, 31].

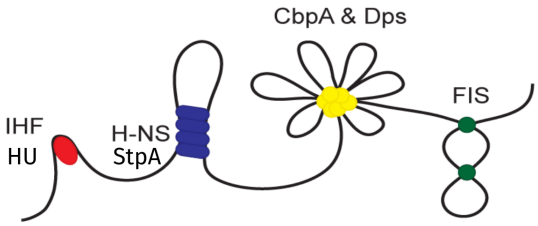


Figure 1.4: Cartoon representation of different ways NAPs interact with DNA (adapted from [32]). IHF and HU cause sharp bends, H-NS and StpA form closed loops of DNA, CbpA and Dps form aggregates with DNA, and Fis cooperatively form microloops.

1.2.1 IHF

Integration host factor (IHF) is one of the most abundant proteins associated with the bacterial chromosome. A 22 kDa protein, it binds to a consensus sequence WATCARNNNNTTR[1] [33] with high specificity, though non-specific binding has also been observed. This binding is known to cause sharp bends of up to 160° [34] in DNA, though a recent study using molecular dynamics and atomic force microscopy has confirmed multiple binding modes inducing angles at $\sim 70°$ and $\sim 110°$ too [35].

As seen in figure 1.5, IHF exists as a heterodimer, made up of two structurally similar subunits, each consisting of an α-helical core with β-ribbon loops attached. At the tips of both subunit's β-ribbon loops lies a proline. These facilitate the binding of the protein to DNA, with the β-ribbon "arms" wrapping along the groove of the DNA, with the prolines then intercalating between two base pairs. This intercalation disrupts the local DNA structure, causing a flexible hinge to be formed which is theorised to allow the

[1]W is A or T; R is A or G; N is any nucleotide

protein's strong bending of the DNA [34].

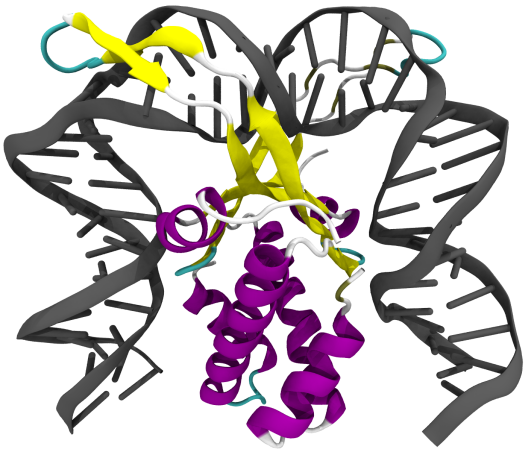


Figure 1.5: IHF is a heterodimer with a α helical core and β ribbon arms which wrap around the DNA to allow prolines at the apices of these arms to intercalate into the DNA. This facilitates the wrapping of DNA around the protein, causing the known sharp bends to form. (PDB entry 1IHF [36]).

Primarily, IHF acts to compact prokaryotic DNA via the creation of sharp bends. However, the full list of functions of this protein is more extensive, particularly The DNABII family of proteins (IHF and HU, a structural homologue) have been implicated in DNA replication [37], recombination [38], and gene regulation — in *E. coli* IHF and HU are known to regulate around 120 genes[39].

IHF has also been implicated in biofilms, a type of microbial community made up of polysaccharides, proteins and extracellular DNA (eDNA) which are present in $\sim 80\%$ of chronic infections and show high resistance to antibiotics [40]. This eDNA forms interwoven lattices, and IHF has been imaged at the crossing points of these lattices [41]. It has been shown that of all the NAPs, only the DNABII proteins are relevant in biofilms [42, 43], and anti-IHF and anti-HU antibodies have been shown to disrupt biofilms [44, 45] with a reduction of up to $\sim 50\%$ having been seen [43].

It has been shown that IHF binding is a two-step process. Initially, a non-specific step which occurs on a ~ 100 µs timescale, followed by a site-specific step on the order of milliseconds [46]. These steps are thought to be initially IHF binding to the DNA, followed by it bending the DNA [47]. This binding-bending mechanism occurs via the protein wrapping the β-ribbon arms along the DNA and intercalating prolines between bases. It has been shown that once intercalated, one of three things can occur in the IHF-DNA complex, as can be seen in figure 1.6. The most energetically favourable outcome is that the protein bridges with a second strand of DNA

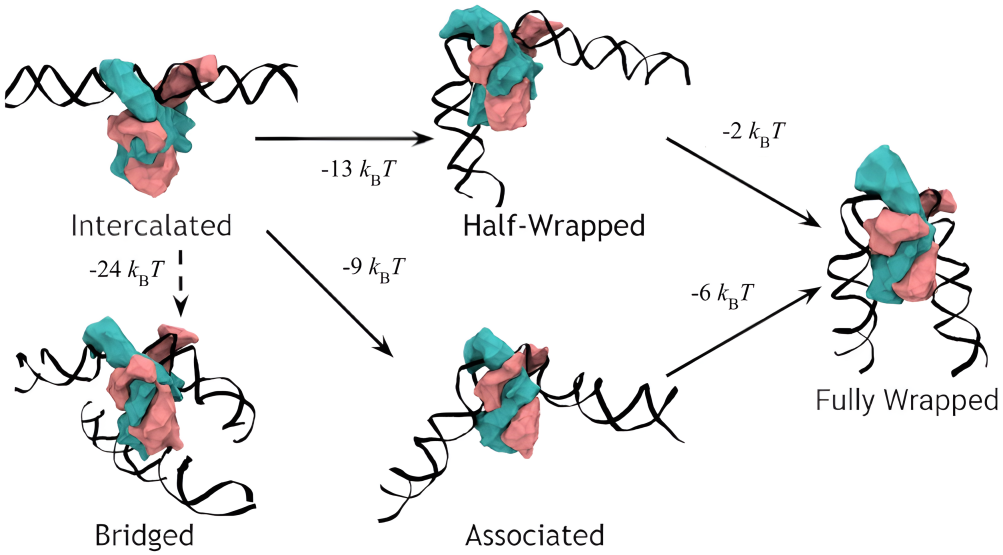


Figure 1.6: A complete model for IHF binding, bending and bridging of DNA. Initially, prolines intercalate between DNA bases to form flexible hinges. If a second strand of DNA is sufficiently close, a bridge will form, otherwise the protein begins to bend the DNA, entering either the associated or half-wrapped state, with slight favourability to the half-wrapped. From here the protein will enter the fully wrapped mode, though only if the consensus sequence is present. (Figure adapted from [35].)

provided one is nearby enough. This occurs via non-specific binding with the bottom of the α-helical core interacting with the phosphate backbone of a DNA strand, and is a possible explanation for why IHF is so relevant at the vertices of the eDNA in biofilms.

If this bridging does not occur, the system instead passes into either the "associated" state (where both sides of the DNA begin wrapping around the protein loosely) or the "half-wrapped" state (where one side of the DNA remains fully unwrapped whilst the other side tightly wraps around the protein), with a slight preference toward the half-wrapped state. Both of these states are either stable or metastable, though will lead to a fully wrapped state (as seen in crystal structures) provided the consensus sequence is present. Interestingly, atomic force microscopy imaging has shown that without the consensus sequence, the protein-DNA complex never enters into the fully wrapped state, only showing results aligning with the associated and half-wrapped states predicted in simulation [35].

Whilst simulations shows the bridging capabilities of IHF combining the initial binding of the β-ribbon arms and non-specific binding to the backbone, experimental results indicate the bridging abilities of the pro-

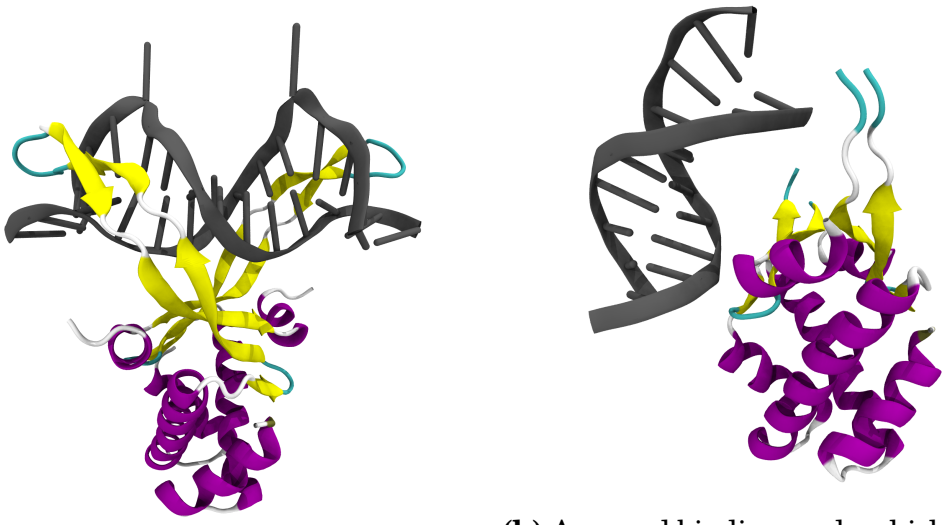
tein go beyond just this state. It was found that at high concentrations of IHF, some aggregation of DNA strands occurred even without the consensus sequence [48, 35]. This suggests that at large cellular concentrations of IHF — such as during the stationary phase in the cell cycle — genome compaction via bridging will occur generally rather than just at points where the consensus sequence appears. In DNA constructs with multiple binding sites close by, bridging was preferred to bending with moderate concentrations of DNA. This indicates that positively-charged architectural proteins screen the electrostatic repulsion between neighbouring DNA molecules which drives the bridging and overall compaction behaviour. It has also been found that even if aggregation is possible, the bending activity of IHF is stronger, allowing it to act as a 'fluidizer' of the genome [49].

An interesting possibility unveiled by these recent results is that the IHF-DNA complex may act as a mechanical switch, capable of occupying multiple distinct states moderated by the position of the DNA to its left, indicating that structural influences (such as tension or supercoiling) upstream of the binding site modulates the bending enforced by the protein.

1.2.2 HU

The histone-like protein from *E. coli* strain U93[2] (HU) is one of the most abundant NAPs, approaching up to 50,000 dimers per cell during exponential growth [50]. HU binds to and bends DNA, though whilst IHF is typically associated with a bend of 160°, HU has had a range of bends angles reported, 70° [51] to 140° [52]. As opposed to IHF, HU binds to DNA with no sequence specificity, but has a strong preference to damaged DNA [34] — such as that with nicks or kinks.

[2]Also known as the heat unstable protein



(a) The canonical HU binding mode (PDB entry 1P78 [52])

(b) A second binding mode which occurs via the α-helical core (PDB entry: 4YFT [53])

Figure 1.7: Two binding modes have been found for HU, one in the same fashion as IHF that occurs at sites of damage (a) and a second non-specific mode where the DNA interacts with the α-helix body — the mode that allows for non-specific compaction of DNA in the nucleoid.

As seen in figure 1.7, HU is a structural homologue to IHF, with a similar dimeric make up of a α-helical body with β-ribbon arms. Though it should be noted that whilst IHF is an obligate heterodimer, HU can exist as either a heterodimer or a homodimer, made up of two subunits (HUα and HUβ). HUαα appears most during the lag and exponential growth phase, whilst the heterodimer HUαβ is most present during the early and late stationary phase. HUββ only appears in small quantities, and only close to the end of a cell's life [54].

HU also binds to DNA in a similar fashion to IHF, with the β-ribbon arms lying in the minor groove of the DNA, allowing apical prolines to intercalate between base pairs and induce and/or stabilise DNA bending [34]. This is the canonical binding mode that occurs specifically at sites of DNA damage. A second binding mode, in which DNA is bound non-specifically along the α-helix body rather than through the arms has also been found, and is the mode that promotes nucleoid compaction and organisation [53].

HU is a member of the DNABII family of proteins (alongside integration host factor (IHF)), of which most prokaryotic genomes encode at least one member. Typically, if there is just one member of the DNABII proteins, it has more similarities to *E. coli* HU than *E. coli* IHF, with those bacteria which do encode a IHF-like protein often also encoding a HU-like

protein [34]. Thus, HU is an almost universal protein, playing important roles in gene regulation. For example, the *gal* operon, which switches the production of enzymes necessary for metabolism of sugar galactose in *E. coli* on or off, uses HU as a transcriptional regulator. Here, HU facilitates DNA loops which block access to promoter regions from RNA polymerase, thus preventing transcription [55, 56].

HU has also been shown to play roles *ex vivo*, capable of acting as a 'molecular glue' that holds biofilms together via non-specific binding of the eDNA and the bacteria [50]. However, the mechanism through which HU plays this role is unknown. Whilst it has been shown that IHF is capable of bridging DNA via a combination of the canonical binding mode and using the bottom of the α-helical core, it is unknown whether HU is capable of using this binding mode.

1.2.3 Other proteins

1.2.3.1 ParB

Faithful inheritance of genetic information in daughter cells is vital in all cell types, and requires accurate DNA partitioning at cell division [57]. Chromosome segregation in bacteria is non-trivial as the DNA must remain in a compacted state to fit within the limited volume of cells whilst DNA replication occurs, as opposed to being separated temporally as occurs in eukaryotes.

Approximately two-thirds of bacterial species encode the ParABS system [58], which mediates DNA segregation at cell division [57]. The ParABS system consists of three components, an ATPase protein ParA, the DNA-binding protein ParB, and a centromere-like sequence *parS* [58]. In this system, one or more *parS* site are located near the origin of replication (*oriC*) [58]. ParB nucleates onto *parS*, following which additional ParB molecules bind to adjacent DNA non-specifically, forming a network of protein-DNA complexes [59].

This ParB-DNA complex stimulates the ATPase activity of ParA, which drives the movement of the *parS* locus (and subsequently, the whole chromosome) towards the opposite pole of the cell [60, 61]. ParB also recruits the Structure Maintenance of Chromosomes (SMC) complex onto the chromosome, reducing DNA entanglement and thus promoting the individualisation of replicated chromosomes [62].

Whilst sequence conservation between ParB proteins is relatively low,

the overall structure of the proteins remains consistent. ParB is widely distributed in bacteria and so must have occurred early in evolution [58], which likely explains the low sequence conservation — as different species evolved, random non-damaging mutations would occur independently. The protein is typically divided into 3 domains: the helix-turn-helix (HTH) DNA-binding domain, which is connected to a C-terminal domain (CTD) and an N-terminal domain (NTD) via flexible linkers [59]. The CTD, the least conserved domain amongst ParB homologs, contains a leucine zipper which allows the protein to homodimerise [63]. It can also play a role in allowing non-specific DNA binding to occur. Conversely, the NTD is the most highly conserved domain, containing an arginine-rich motif, which is used in interactions with ParA, whilst also playing a role in the spreading on *parS*-adjacent DNA [64].

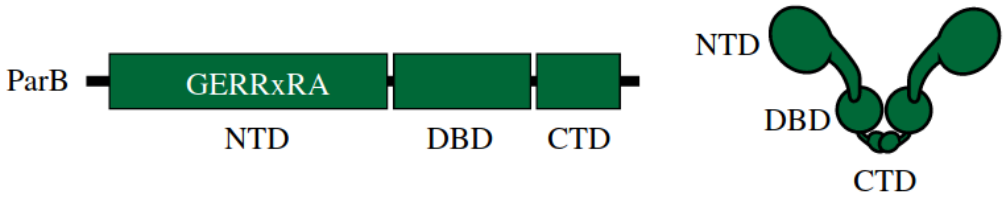


Figure 1.8: Chromosomal ParB proteins share similar structures, with an N-terminal domain, a DNA-binding domain (DBD) and a C-terminal domain all linked with flexible linkers (Figure from [65].)

1.2.3.2 Noc

Noc, another DNA-binding protein, is highly homologous to ParB and binds to a similar DNA sequence (*NBS*) [66], as seen in figure 1.9. Much like ParB, Noc has a three-domain structure, an NTD for protein-protein interactions, a central DNA-binding domain, and a CTD [67]. Whilst *parS* sites occur solely near *oriC*, *NBS* is distributed widely around the genome, barring the terminus of replication (*ter*) [68, 69].

Noc plays an important role in maintaining nucleoid integrity, functioning to prevent cell division machinery from assembling in the vicinity of the nucleoid. By doing this, Noc prevents the nucleoid from being guillotined, which would otherwise damage the DNA [67, 69] — an effect known as nucleoid occlusion [70]. In bacteria where the protein was deleted, cell division was uninhibited, whereas overproduction of Noc lead to longer cells [70]. This suggests that without Noc, the cell division machinery could assemble all along the cell. With too much Noc, instead cells could not be

divided normally, as it would block the division machinery.

Due to their genomic proximity and sequence similarity, it has been suggested that Noc resulted from a gene duplication event from ParB [71]. X-ray crystallography and deep mutational scanning work has revealed the specificity between *parS* and *NBS* is encoded via four residues at the protein-DNA interface and that simple mutations in these residues is enough to switch the protein's DNA binding specificity.

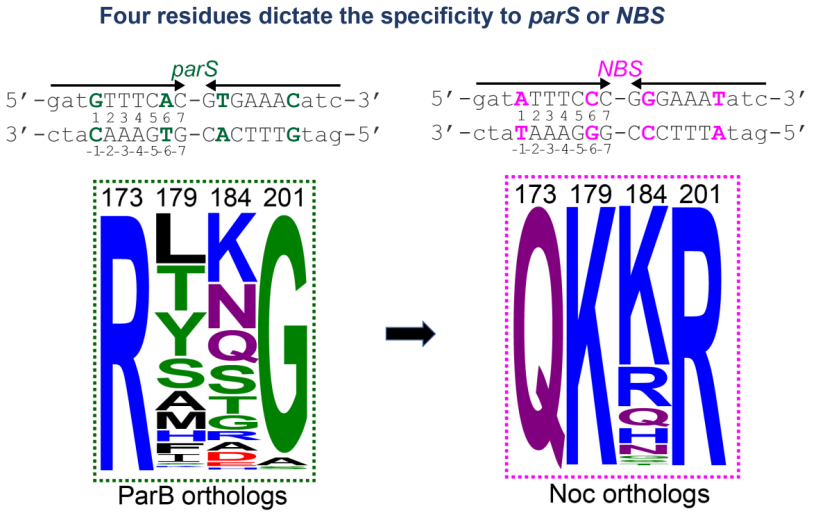


Figure 1.9: The *parS* and *NBS* DNA sites are highly homologous, and four key amino acids have been found that switch specificity between them. This sequence logo shows how frequently amino acids appear at these residues based on their height, so ParB 173 and 201 are almost always arginine and glycine and in Noc they are glutamine and arginine, and residues 179 and 184 are frequently lysine. Amino acid sequences from 21 bacterial species were retrieved and aligned with MUSCLE [72]. (Figure from [73].)

Figure 1.9 shows the 4 key amino acids that switch specificity. A key result was that the only variants that could bind to *NBS* always had a lysine in either or both of the 179 and 184 positions. A single lysine was enough to allow the other key residues to bind to the DNA, but lacking a lysine in these positions caused at least one to fail to bind. Meanwhile, other variants of these 4 amino acids which contained a lysine in the 179, 184 or both residues were found to bind promiscuously to multiple different DNA sites. Hence, it was suggested that these lysine likely have a permissive effect that allows the other two amino acids to bind. These findings show that studying these proteins is a good example for understanding how protein-DNA recognition evolves as new regulatory functions become necessary.

1.2.3.3 Fluorescent Proteins

Coined by George Stokes when he observed a fluorite sample emitting a blue light when exposed to ultraviolet light, fluorescence is the process in which photons are emitted by a molecule after being excited by a photon of higher energy. Fluorescence has become a key microscopy tool for studying biological structures and dynamics. One of the most key advancements in this field came in 1962, when Osamu Shimomura discovered the Green Fluorescent Protein (GFP) [74].

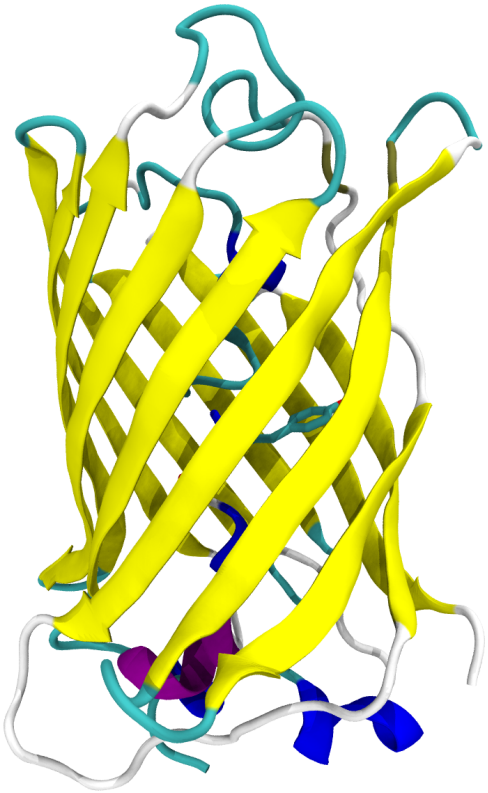


Figure 1.10: GFP (PDB 1EMA [75]), has a β-barrel structure, with an internal chromophore.

As can be seen in figure 1.10, GFP is comprised of an internal chromophore surrounded by a β-barrel (eleven beta sheets in a barrel-like structure). It is approximately 40Å in length by 25Å in width, and is made up of 238 amino acids. GFP absorbs light with an excitation maximum of 395 nm and fluoresces with an emission maxima of 510 nm [76, 77]. The primary use of GFP is that is can be expressed directly attached to target proteins in different organisms [78]. By then shining a laser at the excitation wavelength of GFP at the cell, the GFP will then fluoresce, which allow the localisation of the protein. Through this expression, proteins can be studied in their physiological conditions without perturbing the cell via

introducing new labelled molecules into the system (As it has been shown that adding GFP to the cell has minimal effect on the cell life cycle).

There exists a rather large number of variants of GFP [79] which have their own advantages. For example, some may fluoresce more intensely than the wild-type whilst others are designed for specific experiments, such as the enhanced yellow fluorescent protein (eYFP) which is used for blinking assisted localisation microscopy due to its photoblinking nature [80]. A commonly used variant known as monomeric enhanced green fluorescent protein (meGFP), which contains a mutation from an alanine to a lysine at the GFP dimerization domain. This mutation introduces a positive charge, preventing the aggregation of the protein which would otherwise occur [81].

1.3 DNA-Protein specificity

The binding between proteins and DNA is of particular interest as the nature of these interactions forms the fundamental basis on how many proteins perform their key roles in organisms. Despite this, there is a relative lack of knowledge on the specifics of these interactions. Even though attempts have been made to create a general code to explain these interactions through the study of atomic interactions between amino acids and nucleobases, no general explanation exists [82, 83]. Thus, it's clear that work must be done within individual protein families to understand their own specificities, rather than attempting to find a general, catch-all explanation [84].

The myriad ways through which a protein can bind to DNA furthers the complexities that prevent a standard model being truly developed. From initial X-ray structures of nucleic acid duplexes, it was realised that the major groove of the DNA helix unveiled base-specific hydrogen bond donors and acceptors that could be recognised by the amino acid side chain's donors and acceptors, and thus the idea of direct readout was developed [85]. In direct readout, hydrogen bonding between nucleobase and amino acid occurs, typically at the major groove though it does also occur at the minor groove [9]. However, despite the relevance of direct readout (which is present in almost all protein-DNA models deposited on the Protein Data Bank), if it were the only method of recognition then a general code between amino acid and base would theoretically be possible. The structure of the DNA itself contributes to the binding activity of proteins, with deviations in the B-form helix such as a bend unveils other bases which enable

interactions to occur — this was coined "indirect readout" [86].

It is also difficult to understand to what level these methods of recognition are being used however. DNA shape is a function of its sequence, meaning it is difficult to determine whether a protein is binding due sequence or shape. It has also been shown that interaction with the phosphate backbone of DNA, as opposed to the base itself, is a method of indirect readout. Previous studies have shown that by mutating backbone-binding amino acids, the overall binding preferences of the protein changed and it no longer preferred the same sequence [87].

If one were to look only into structural families of proteins as opposed to individual proteins, there are three overall classifications for how a family recognises DNA. These are non-specific, highly specific and multi-specific. Respectively, these are that they bind promiscuously with no sequence specificity, that the whole family binds only to a specific sequence, or that different members in a family bind specifically but to different sequences [82].

Another factor of interest is in how specificity evolved in proteins. Whilst most studies considered how changes in DNA occur to allow a new target gene to interact with an existing transcription factor (a protein that controls the rate of transcription of genetic information), fewer consider how the transcription factor's specificity evolved, or how they could have evolved new regulatory modules if they were bound to conserve essential ancestral functions [88].

To solve this, the evolutionary transition for the binding of proteins to DNA must be dissected fully to identify why certain pathways wouldn't work in order to figure out which pathway did. Previous work has shown that permissive mutations play an important role in enabling a protein to tolerate other mutations that would otherwise negatively impact the capabilities of the protein to bind to DNA [89]. Enzyme engineering has previously suggested that these beneficial mutations occur in a series of single amino acid substitutions at a time [90].

1.4 Facilitated Diffusion

An additional question in DNA-protein interactions is how do these proteins locate their recognition sites within DNA. In 1970, it was found that the Lac repressor associated to its target site 100-times faster than would be expected of a 3D diffusion-limited search [91]. This ultimately lead to the idea of facilitated diffusion, a model introduced in 1981 by Berg, Win-

ter and von Hippel [92]. Rather than pure diffusion, the target site search instead occurs via a combination of 3D diffusion in the solution, and 1D diffusion along DNA [92, 93]. This reduces the search time significantly, as the probability of finding a binding site substantially increases when compared to if the search was purely 3D bulk diffusion.

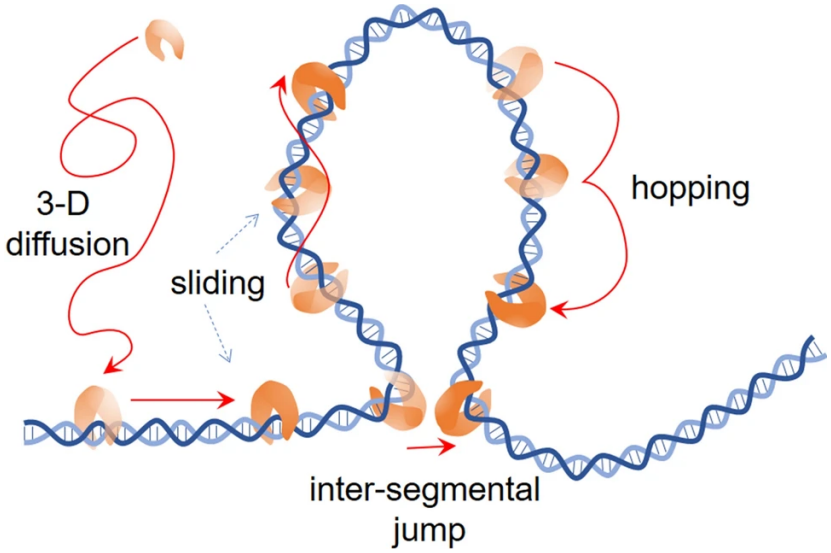


Figure 1.11: A representation of the different 3D and 1D diffusion methods [94]. The facilitated diffusion model assumes that, once 3D diffusion allows a protein to interact with a DNA strand, a combination of sliding, hopping and inter-segmental jumping allows said protein to search the DNA for a target site. The protein will dissociate from the DNA and the process is repeated until the site is found.

The first 1D diffusion method is sliding along the DNA. Here, the DNA-binding protein associates with the DNA via non-specific interactions, typically electrostatic interactions between the DNA phosphates and the basic residues of the protein [95]. This allows the protein to repeatedly sample the DNA sequences around its initial landing site, and was first observed at a single molecule level in 1993 [96]. An important thing to note is that this sliding can be either linear (such that the protein moves along the DNA staying on one side) or helical (such that it rotates along the DNA as it slides). This can vary depending on the protein's architecture, or environmental factors such as salt concentration [97].

The second 1D diffusion mechanism is hopping along the DNA. As seen in figure 1.11, this method involves the protein repeatedly dissociating and reassociating with the DNA at different sections on the order of tens of bases, dependent on factors such as salt concentration or how positively charged the surface of the protein is. Experimentally, this is hard to differentiate from sliding due to spatial and temporal resolution limitations [98].

A key advantage of hopping over sliding is that it allows for the obstacles (such as other bound proteins) to be avoided, which has defined how the hopping mechanism is studied experimentally, placing obstacles along the DNA and viewing if the protein still diffuses along a strand [99]. Another test is the dependence on salt concentration — as these interactions are mostly electrostatic, an increase in salt concentration would be expected to increase the dissociation rate of the protein.

As DNA is highly compacted in cells, the formation of loops and plectonemes causes distant segments to be within close proximity to one another. Therefore, it is possible for proteins to perform an intersegmental jump from one of these segments to another during its diffusion along the DNA. This also gives an advantage in that there will be minimal revisiting of a site that might otherwise occur in Brownian 1D sliding as the protein may unbind and then rebind to the DNA facing the opposite direction, thus reducing the oversampling and improving the rates to finding an appropriate site [100].

1.5 Scope of this Thesis

Protein-DNA interactions are key for life, and also present a promising avenue for DNA nanotechnology. The work presented here aims to understand the key underlying mechanics in some of the most abundant proteins that allow their roles to occur. The following chapters describe:

- The methodology applied in this work. All-atom molecular dynamics simulations will be reviewed in full, followed by an in-depth explanation of atomic force microscopy. These two techniques give us atom-level detail into key protein-DNA interactions with complementary experimental verification.
- An application of combining molecular dynamics and atomic force microscopy to characterise the interactions between HU and DNA, to study how HU recognises sites of damage, provide atomistic insight into the mechanisms of facilitated diffusion of architectural proteins like HU, how the non-specific interactions of HU aid to compact and stabilise biofilms, and how the protein is similar and yet different to IHF, for example why does IHF act as a mechanical switch to control thermal fluctuations whilst HU seems to allow more flexible breathing of DNA?

- Work to understanding how the specificity in protein-DNA recognition evolved over time, using the related proteins ParB and Noc to view how single base pair or amino acid mutation can cause changes in the specific binding.
- Development of a simulation methodology for DNA origami, and then applying it to understand the mechanical encapsulation of fluorescent proteins within a DNA box.

Chapter 2

Methods

2.1 Molecular Dynamics

Molecular dynamics (MD) is a computational approach that allows the changes in a system to be viewed over time. The system is comprised of multiple components, each with their own positions and velocities, and these are then evolved through time based on physical laws (at a most basic level, these would be Newton's laws of motion and electrostatic interactions). This can be done on multiple levels depending on the time and length scales desired to be studied, starting at the smallest level, *ab initio*, allowing influences on electronic structure, to scales simulating on the atomic level, whilst larger ones would be coarse-grained by the combination of multiple atoms into a single group for calculation.

At a base level, all MD simulations follow the same simple steps outlined here. However, these belie the complexities of each step, which require individual in-depth discussion for a full understanding. But these are sufficient for a general understanding before such a discussion.

1. Determine initial atomic positions for atoms and assign velocities
2. Calculate the force (and from which, the acceleration) experienced by each atom in the system based on a predefined potential
3. Move the atom's positions based on their velocities and adjust the velocity based on the acceleration
4. Repeat steps 2 and 3 until the desired number of steps is completed

2.1.1 Simulation initialisation

To perform a simulation one needs a system to simulate, requiring the atoms with which the dynamics will occur. Whilst this may seem simple, this is a task which requires the utmost care, as the initial step it is key towards running a successful simulation [101].

For DNA, this is a relatively simple task. As it has the known double-helical structure, made up of repeating units with known geometries, the atomic positions of an ideal B-DNA structure can be easily calculated, and programs such as the Nucleic Acid Builder (NAB) have been developed to do so [102]. NAB has also been extended to work to create A-DNA, but more complex forms of DNA (such as Z-DNA, or supercoiled DNA) require further modifications to the generated structure.

For proteins, this becomes a more challenging situation. Whilst, like DNA, amino acids have known atomic positions relative to one another, proteins aren't just straight chains of amino acid and so a base structure cannot be created in the same way. Rather, they are folded into complex structures for which there lacks a simple method to predict — a problem known as the 'protein folding problem' [103]. Instead, typically structures are generated experimentally using techniques such as X-ray crystallography, nuclear magnetic resonance or electron cryomicroscopy. In order to share these structures, an online database known as the "Protein Data Bank" (PDB) [104] was created and is used by researcher worldwide. In recent years, the machine learning approach to protein structure prediction has lead to the AlphaFold program [105], which has seen unprecedented success for the highly accurate prediction of structures from amino acid sequence. However, there is limited reported data on the quality of these structures for simulation, and it should be noted that this is not in itself a solution to the protein folding problem. AlphaFold alone also only captures a single state, telling nothing about about how a protein switches between states. For this, models that are at least partly physical are necessary.

When an appropriate crystal structure has been selected for MD, it is often necessary to manipulate it to have the desired system for simulation. The most basic example would be of simply extracting a segment of interest (for example, one may simply be interested in the dynamics of the DNA-binding domain of a protein), however other cases would require a much higher level of editing of the structure. For example, crystal structures of DNA bound to proteins typically only have few base pairs, whereas the dynamics of a much larger system are of interest. In this case, a longer

strand of DNA would be created (such as with NAB) and would then have to be aligned with that of the PDB in such a way that the root-mean-square deviation (RMSd) of the initial to final structure are minimised.

These manipulations will often create a final structure that has unrealistic bond lengths or atom positions. If two atoms are too close together, they would immediately experience an incredibly strong repulsive force which would result in an unusable trajectory. Similarly, bond lengths or angles that go beyond the ranges that force fields have been parametrised for would also result in unphysical trajectories. Hence, before an MD simulation is actually performed it is wise to first perform an energy minimisation on the starting structure.

Energy minimisation is the process in which the initial positions of the atoms are taken and moved in such a way that the total potential energy of the resultant system finds a minimum. This is not a global minimum as these algorithms are applied locally, and there is no knowledge of the full configuration space.

2.1.2 The AMBER force field

Once an initial structure has been created, the rules that govern how the atoms in the system will interact must be defined. To do this, a potential field (or "force field") that considers the interactions between each pair of atoms is constructed. The Assisted Model Building with Energy Refinement (AMBER) [106, 107] software suite provides a set of force fields, which is comprised of several terms:

$$\begin{aligned} V_{\text{total}} &= \sum_{\text{bonds}} V_{\text{L}} + \sum_{\text{angles}} V_{\text{A}} + \sum_{\text{torsions}} V_{\text{T}} + \sum_{\text{pairs}} (V_{\text{LJ}} + V_{\text{C}}) \\ &= \sum_{i=1}^{n_{\text{B}}} K_r (r_i - r_{0i})^2 + \sum_{i=1}^{n_{\text{A}}} K_\theta (\theta_i - \theta_{0i})^2 + \sum_{i=1}^{n_{\text{T}}} \sum_n V_{in} [1 + \cos(n\phi_i - \gamma_i)] \\ &+ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 R_{ij}} \right) \end{aligned} \tag{2.1}$$

With n_{B}, n_{A} and n_{T} being the number of covalent bonds, bond angles and bond torsions respectively. This section will go into detail explaining each of these terms.

2.1.2.1 Bond lengths term

The first term of Eq.(2.1) is the bond length term, which represents the bond energy between covalently bonded atoms. These covalent bonds vibrate due to their thermal energies, and if the temperature are not extremely hot or cold (i.e. room temperature), this bond can be modelled as a harmonic spring to good approximation, and is thus represented as a Hookean function [108].

$$V_{\mathrm{L}} = K_r(r - r_0)^2 \tag{2.2}$$

Where K_r is the spring constant, r is the instantaneous bond length and r_0 is the equilibrium bond length.

2.1.2.2 Bond angles term

Supposing three atoms (i, j and k) which are covalently bonded as *i–j–k*, the angle between these bonds, θ, will also fluctuate due to thermal energy. Thus, like the bond lengths term, the bond angles term is also represented as a Hookean function.

$$V_{\mathrm{A}} = K_\theta(\theta - \theta_0)^2 \tag{2.3}$$

With the variables being equivalents to those in the bond lengths term, but being based for the bond angles instead.

2.1.2.3 Torsional term

Whilst single bonds can generally rotate freely, double and triple bonds generally cannot. Instead, there is a torsional cost that must be paid due to that bond order, neighbouring bonds, or nearby lone electron pairs. This is represented by the third term in the force field, which accounts for the potential of the torsion of each bond. As it is due to the relative rotation of bonds *i–j* and *k–l* about bond *j=k*, it must be valid over the full 2π angle and is generally expressed as a Fourier series.

$$V_{\mathrm{T}} = \sum_n V_n \cos(n\phi - \gamma) \tag{2.4}$$

Where V_n is the amplitude of the n-th term of the Fourier series, with the first maximum of the periodic function at the phase angle γ. n is the periodicity of the function, with ϕ being the torsional angle subtended by the covalently bonded atoms.

2.1.2.4 Nonbonded interactions

Even atoms with no bonds between them have interactions, these being the short-ranged electron orbital repulsive force, the long-ranged attractive London dispersive force, and the Coulomb potential.

$$V_{\mathrm{LJ}} = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{2.5}$$

$$V_{\mathrm{C}} = \frac{q_1 q_2}{4\pi\varepsilon_0 r} \tag{2.6}$$

Equation 2.5 represents the Lennard-Jones potential, which itself models the attractive London dispersion force (via the r^{-6} term) and repulsive effect due to the Pauli exclusion principle (via the r^{-12} term). ε is the depth of the potential well, r is the distance between the interacting particles, and σ is the value of r at which the potential equals zero.

Whilst the attractive term has a physical basis for choosing a power of six with the attractive dispersive interactions occurring due to fluctuating partial charges which have been shown to decay with $1/r^6$ [109] (The London dispersion force, where the interactions between two induced dipoles approximately falls as $1/r^6$ as the potential of a single dipole decays $1/r^3$), the $1/r^{12}$ term is primarily chosen for the computational efficiency. Whilst it does approximate the Pauli repulsion with reasonable accuracy, it allow the potential to be formulated as $V_{\mathrm{LJ}} = 4\varepsilon(a^2 - a)$ where $a = (\sigma/r)^6$, allowing the more computationally expensive operations to be performed just once.

Equation 2.6 adds the Coulomb potential to the force field. Whilst on it's own the Lennard-Jones potential is sufficient for many simulations, such as a fluid of uncharged argon atoms, a system of biological molecules is more complicated. Even if the system is electrically neutral, many individual atoms in the system will still carry partial charges which must be accounted for in their interactions with one another. Here, q_1 and q_2 are the charges of atoms, r is the distance between them, and ε_0 is the permittivity of free space. This is representative of an ideal free-space simulation, often the term is represented as ε_{out} as the relative dielectric permittivity which is system dependent (e.g. about 80 for water).

2.1.3 Solvent models

An additional factor that must be accounted for when simulating any biological molecule is that the interior of a biological cell is an aqueous envi-

ronment, and that charged molecules such as DNA require counterions to be stable. Here, two methods of incorporating solvents into simulation are discussed.

2.1.3.1 Explicit solvents

The method that might come to mind first would simply be to simulate each water molecule individually, which forms the basis of the "explicit solvent model". In this model, the molecule of interest is effectively placed in a box such that it is entirely immersed in water. However, this seemingly menial task is deceptively simple, with a number of parameters which must be considered for accurate simulation. The water box must be large enough such that it captures the molecule regardless of any conformations it may enter, which can increase the number of atoms in the system exponentially. This is unfortunate, as the dynamics of the water itself is generally of little interest, yet it makes up a vast number of the atoms simulated, far overshadowing the molecule of interest.

The dynamics of the box itself are also an area of interest. If the box were kept fully shut, the water atoms would simply hit the edges of the box and bounce back, however this is not an accurate representation of real life. If the system instead had no hard boundaries, the solvent would continuously diffuse into the surrounding vacuum, until it all leaves the molecule of interest.

Instead, periodic boundary conditions are employed to approximate a large system. Under this scheme, when an atom hits the edge of the box, it enters on the other side of the box. In this way, the pressure and volume of the box is kept constant whilst approximating the dynamics of an infinitely large system. To ensure more realism in the simulation, it is important that molecules can interact even across the boundary, so atoms on opposite edges of the boxes will interact as though they were right next to each other.

There is no requirement that the box be a cube, the only constraint on the shape of the box is that periodic boundary conditions can be applied, thus it must be space-filling. This allow more compact shapes (such as a truncated octahedron) to be used as opposed to a cube, thus reducing the number of solvent atoms in the distant corners.

There are various water models available, thus giving the user choice depending on the accuracy required and how quickly the simulation must be run. The simplest of these are rigid models which rely solely on non-bonded interactions. As such, in these models each water molecule is sim-

ply a triangle of O-H covalent bonds whose side lengths and angles are fixed. One of the most frequently used models, the TIP3P (transferable intermolecular potential with 3 points) model [110], works on this basis, with a potential of form:

$$V_{\mathrm{H_2O}} = V_{\mathrm{LJ}} + V_{\mathrm{C}} \tag{2.7}$$

Where the Lennard-Jones applies only to the oxygen atoms whilst the hydrogen atoms interact purely on the Coulomb force. The TIP3P model uses bond lengths $r(\mathrm{OH}) = 0.9572\mathring{A}$ and angle $\mathrm{HOH} = 104.52^{\circ}$ which were derived empirically from experimental data. There are also 4 and 5 site models, in which the charge (or charges) are instead placed on dummy atoms that better replicates the electrostatic distribution around the molecule. However this leads to an increase in the computational cost (5 site models, by virtue of adding in the two dummy atoms, increase the total number of atoms in a system by approximately two-thirds). Whilst there are specific simulations in which these more accurate models are necessary (in particular, simulations of ice), for a typical simulation the cost of these models are not worth the minimal gains that would be achieved when compared to the 3 site model.

Lastly, whilst the computational cost of these solvents are a large factor in the slow down that occur, it must also be noted that water molecules introduces viscosity into the system, thus slowing down the dynamics of the molecule of interest.

2.1.3.2 Implicit solvents

An alternative approach to solvation in simulations is to model the solvent as a continuum with dielectric properties, rather than explicitly modelling it. This is the basis of the implicit solvent, with ion effects being incorporated into the models itself, such that neither water molecules or solvent ions are necessary in the simulation. The starting point for these models is to consider the solvation free energy, which is split into electrostatic and non-electrostatic parts:

$$\Delta G_{\mathrm{solv}} = \Delta G_{\mathrm{el}} + \Delta G_{\mathrm{nonpolar}} \tag{2.8}$$

$\Delta G_{\mathrm{nonpolar}}$ represents the free energy of solvating a molecule from which all charges are neutralised. This originates from a combination of the solvent-solute van der Waals forces and the cost incurred by disrupting the structure of the solvent around the solute — $\Delta G_{\mathrm{nonpolar}}$ is often approximated

as proportional to the solvent-accessible surface area of the solute with an empirically derived constant of proportionality.

ΔG_{el}, in contrast, is the energy required to remove all charges from the solute and add them into the solvent.

$$\Delta G_{\mathrm{el}} = \frac{1}{2}\sum_{i=1}^{N} \mathrm{q_i}\Psi(\mathbf{r_i}) \tag{2.9}$$

Where N is, again, the number of atoms, $\mathrm{q_i}$ is the charge of the i-th solute atom with position r_i, and Ψ is the electric potential, the bulk distribution of which can in principle be described by the Poisson equation [111].

$$\nabla^2\Psi = -\frac{\rho_e}{\varepsilon\varepsilon_0} \tag{2.10}$$

Where ρ_e is the local electric charge density and ε is the dielectric constant of the solvent. This can then be combined with the Boltzmann equation to account for the free movement of ions in solution, thus giving the local ion density c in terms of the bulk ion concentration c_0:

$$c = c_0 \exp\left(\frac{-\mathrm{W}}{k_B T}\right) \tag{2.11}$$

where W is the work required to move an ion from an infinite distance. This can be reformulated to:

$$c_{\pm} = c_0 \exp\left(\frac{\mp e\Psi}{k_B T}\right) \tag{2.12}$$

as $W = \pm e\Psi$ where e is the charge of an electron. Hence, the local electric charge density is:

$$\begin{aligned}\rho_e &= e(c_+ - c_-)\\ &= c_0 e\left[\exp\left(\frac{-e\Psi}{k_B T}\right) - \exp\left(\frac{e\Psi}{k_B T}\right)\right]\\ &= -2c_0 e \sinh\left(\frac{e\Psi}{k_B T}\right)\end{aligned} \tag{2.13}$$

By substituting this into equation 2.10, the Poission-Boltzmann equation is created,

$$\nabla^2\Psi = \frac{2c_0 e}{\varepsilon\varepsilon_0}\sinh\left(\frac{e\Psi}{k_B T}\right) \tag{2.14}$$

however this is a nonlinear differential equation which, in most circumstances, can only be solved numerically. Nevertheless, it can be solved analytically in certain geometries, most notably as a planar surface which is

infinite in both the y and z coordinates, such that the potential can only change in x. When the potential is small ($e|\Psi| << k_B T$) in this geometry, equation 2.14 is solved by

$$\Psi(x) = \Psi_0 \exp\left(-\sqrt{\frac{2c_0 e^2}{\varepsilon_{out}\varepsilon_0 K_B T}}x\right) \tag{2.15}$$

as this low potential approximation is valid for typical molecular dynamics simulations. This linearisation at small potentials is known as the Debye-Hückel approximation.

Whilst specialised Poisson-Boltzmann solvers do exist [112], these remain inefficient (requiring Fourier transform-based algorithms which are computationally expensive). Thus, for use in large scale molecular dynamics, further approximations must be applied. A widely used approximation is the implicit Born model, beginning with the Born equation

$$\Delta G_{solv}(R_i) = -\left(1 - \frac{1}{\varepsilon}\right)\frac{q^2}{2\mathrm{A}} \tag{2.16}$$

where A is the ion radius and q is its charge.

This is an example of an exact solution to the Poisson-Boltzmann equation for spherical geometries with a single charge centre [113, 114], however neither of these criteria are fulfilled by biomolecules.

Therefore, a more general model is necessary. Consider a molecule to consist of N charges $q_1 \cdots q_N$ embedded in spheres of radius $a_1 \cdots a_N$. Should the separation r_{ij} between any two spheres be sufficiently large compared to their radii, the free energy can be approximated by a set of Born terms and pairwise Coulomb terms [115].

$$\Delta G_{el} \simeq \sum_i^N -\frac{q_i^2}{2a_i}\left(1 - \frac{1}{\varepsilon_{out}}\right) + \frac{1}{2}\sum_i^N\sum_{j\neq i}^N \frac{q_i q_j}{r_{ij}}\left(\frac{1}{\varepsilon_{out}} - 1\right) \tag{2.17}$$

A key limitation occurs with this representation — atomic spheres are not necessarily far from one another, hence necessitating further expansion to this equation 2.17. ΔG_{solv} will be quadratic in the source charges due to the linearity of the Poisson-Boltzmann equation. Therefore, equation 2.17 can be generalised to

$$\Delta G_{el} \simeq \left(1 - \frac{1}{\varepsilon_{out}}\right)\frac{1}{2}\sum_{ij}\frac{q_i q_j}{f_{ij}^{\mathrm{GB}}} \tag{2.18}$$

where f^{GB} is some simple function. The diagonal $(i = j)f^{\mathrm{GB}}$ terms are considered the effective Born radii, whilst off-diagonal terms are the effective

interaction distance. These generalised models are, fittingly, known as the generalised Born. Still *et al* [116] proposed an effective, commonly used function for f^{GB}:

$$f_{ij}^{\mathrm{GB}} = \left[r_{ij}^2 + R_i R_j \exp\left(\frac{-r_{ij}^2}{4R_i R_j} \right) \right]^{\frac{1}{2}} \tag{2.19}$$

Where the R_i terms are the effective Born radii of atoms, dependent on the intrinsic radius of atom i, a_i and the relative positions of all other atoms. The shielding of an atom from the solvent by surrounding atoms determines the effective Born radius. As R_i must be calculated at each timestep, an efficient method to compute this is necessary. The AMBER implementation is to use the Coulomb field approximation (CFA) [117]. The CFA replaces the true electric displacement around the atom by the Coulomb field, allowing the following expression to be derived:

$$\begin{aligned} R_i^{-1} &= \rho_i^{-1} - \frac{1}{4\pi} \int \theta(|\mathbf{r}| - \rho_i) r^{-4} d^3\mathbf{r} \\ &= \rho_i^{-1} - \mathrm{I}_i \end{aligned} \tag{2.20}$$

where the integral is over the solute volume surrounding atom i. Computationally, this is an expensive integral to solve, and thus further approximations are made to obtain a closed-form analytical expression via a pairwise descreening approach by Hawkins, Cramer and Truhlar [118] — the GB-HCT model. However, for macromolecules this underestimates the effective radii for buried atoms, and so the Born radii of buried atoms must be scaled up using an empircally derived set of parameters α, β, γ, as found by Onufriev, Bashford, and Case (in their GB-OBC model [119]),

$$R_i^{-1} = \tilde{\rho}_i^{-1} - \rho_i^{-1} \tanh(\alpha\Phi - \beta\Phi^2 + \gamma\Phi^3) \tag{2.21}$$

where $\tilde{\rho}_i$ is ρ_i minus some offset, and $\Phi = \tilde{\rho}_i \mathrm{I}_i$. Both models use the van der Waals surface to define the solvent-solute boundary, allowing much more efficient computation at the cost of some accuracy. Mongan *et al* partially corrected for this by additionally integrating over the "neck" regions formed by molecular surfaces between pairs of nearby atoms, in the GB-neck model [120]. The last effect that needs to be incorporated for a complete implicit solvent model is that of electrostatic screening performed by monovalent ions. This is done by adding the Debye-Hückel screening

parameter κ [121]

$$\kappa = \sqrt{\frac{8\pi \mathrm{I}}{\varepsilon k_B T}} \tag{2.22}$$

for solution of ionic strength I, such that when incorporated into equation 2.18 results in

$$\Delta G_{el} \approx -\frac{1}{2} \sum_{ij} \frac{q_i q_j}{f_{\mathrm{GB}}} \left(1 - \frac{\exp(-\kappa f_{\mathrm{GB}})}{\varepsilon}\right) \tag{2.23}$$

The generalised Born model has become the standard for implicitly solvated MD simulations due to its reasonable efficiency and accuracy to the linearised Poisson-Boltzmann equation. It is simple to integrate the Gibbs free energy into a force field provided the Born radii is estimated accurately, with the parameters have been refined numerous times to yield accurate results for the majority of systems [122].

Despite the amount of work that has gone into developing generalised Born models for simulation, ultimately they remain less mature than explicit solvent simulations. It's been noted that salt bridges are overstabilised [123] and protein and peptide secondary structures are often not properly reproduced [124]. Despite these issues, the speed up from implicit solvent has resulted in its wide use for simulation systems with large length scales due to the removal of water and ion atoms which also in turn reduces solvent friction. This allows a wider sampling of the conformational space in any given system — typically the larger the conformational space, the higher the speed up from implicit solvent.

2.1.4 Integration methods

Once a system is suitably prepared for simulation, one must actually simulate it. At a glance this may appear to be trivial, where one might think you just solve the movement over time continuously. Alas, this is not the case, there is no general analytical solution existing for three-bodies, let alone the many hundreds of thousands which may be involved in a typical MD simulation. Instead, one must instead evolve the system through discrete time steps Δt. There are multiple schemes to compute this, well known ones include the Euler [125], leapfrog [126] and velocity Verlet [127].

2.1.4.1 Euler method

The Euler method is the most straightforward method for numerical integration. Here, the position, x, of any given particle at the next time-step is calculated as

$$x_{n+1} = x_n + v_n \Delta t \tag{2.24}$$

with the velocity, v, also being recalculated each time-step by,

$$v_{n+1} = v_n + a_n \Delta t \tag{2.25}$$

and acceleration, a, being calculated from the potential. Whilst this simplicity makes the Euler method an attractive choice when implementing an integration scheme, there are two key issues that result in it often being unsuitable for most applications — errors and stability. As the Euler method is a first-order method, global error (the error made on the whole time interval of the integration) scales linearly with step size, whilst local error (the error that occurs on a single step) scales quadratically with step size. On the other hand, the stability (that is to say, how the growth of round-off errors and small fluctuations in initial data affects the final result) is also poor. For many linear equations, a step size must be chosen such that it is within the stability region, which is often poorly defined. If a step size outside of this region is chosen, then the global error will approach infinity as the integration continues. The ill-definition of the stability region is ultimately a key limitation which results in the Euler method seeing limited use beyond a simple example of numerical integration. Further issues of the Euler method is that it is not symplectic, and is not time reversible.

2.1.4.2 Leapfrog method

Whilst the Euler method acts to solve for position and velocity simultaneously, the leapfrog method instead staggers these calculations as follows

$$v_{i+\frac{1}{2}} = v_{i-\frac{1}{2}} + a_i \Delta t \tag{2.26}$$

$$x_{i+1} = x_i + v_{i+\frac{1}{2}} \Delta t \tag{2.27}$$

$$v_i = \frac{1}{2}\left(v_{i+\frac{1}{2}} + v_{i-\frac{1}{2}}\right) \tag{2.28}$$

with a_{i+1} being calculated using the known force. This staggered calculation allows the position and velocity calculations to "leapfrog" over one another, hence the name. Provided Δt is kept constant and $\Delta t \leq \frac{2}{\omega}$, the

leapfrog method is stable for oscillatory motion whilst keeping the same number of steps per calculation as the Euler method [128]. A key advantage to the leapfrog method is that it is a second-order scheme, hence the global error scales quadratically with Δt. In addition, the stability of this algorithm is far superior to that of the Euler method, as there are fewer ill-defined stability regions.

However, there are disadvantages to the scheme. Firstly, it is clear that to determine $v_{\frac{1}{2}}$, $v_{-\frac{1}{2}}$ is necessary yet it is poorly defined. Whilst this can be circumvented by assuming $v_{\frac{1}{2}} = \frac{1}{\Delta t}(x_1 - x_0)$, one must then estimate x_1 as it is yet to be calculated and it itself depends on $v_{\frac{1}{2}}$. Therefore, the overall accuracy is dependent on how well x_1 is predicted. MD simulations that use the leapfrog method solve this issue in the equilibration phase, in which the system is gradually heated from 0K (where all atoms have 0 velocity) to the desired temperature. This allows the system to explore the conformational space until an equilibrium structure is found. If done correctly, all atoms will take on appropriate velocities, and a Maxwell-Boltzmann distribution of velocities will arise naturally. Secondly, as velocity and position are calculated asynchronously, the velocity-dependent kinetic energy and position-dependent potential energy are out-of-sync as well, causing the total energy of the system to be incorrect at all times, though by using a small enough time-step the difference in the real and calculated energy will be minimal.

2.1.4.3 Velocity Verlet

The velocity Verlet method is very similar to the leapfrog method, to the point that some authors consider it to be in a class of "leapfrog-type" integrators. It uses higher-order equations of motion as follows:

$$x_{i+1} = x_i + v_i \Delta t + \frac{1}{2} a_i \Delta t^2 \tag{2.29}$$

$$v_{i+1} = v_i + \frac{1}{2}(a_i + a_{I+1})\Delta t \tag{2.30}$$

where the acceleration terms make this a half-step method. This allows the position and velocity to be synchronous, hence allowing the accurate calculation of total energy. This is at the cost of requiring these higher-order calculations, thus making the velocity Verlet method less computationally efficient in comparison to the Euler and leapfrog methods.

One can't simply define the integrator in MD suites like AMBER as leapfrog or velocity Verlet, as the addition of SHAKE restraints and ther-

mostats and barostats adds a layer of complexity that makes the distinction between the techniques more technical and ill-defined.

2.1.5 Thermostats and barostats

2.1.5.1 Berendsen thermostat and barostat

There's more to a physical system than the positions of atoms. A usual experiment is performed on the isothermal-isobaric ensemble (NPT) in which the number of particles N, pressure P and temperature T are fixed. However, basic MD simulations instead use the microcanonical ensemble (NVE) where volume V and total energy E is fixed as opposed to pressure and temperature. There is not a one-to-one map between total energy of a system and the temperature — if the initial potential energy is high, as it is partially converted into kinetic energy and back the temperature would noticeably fluctuate. The time-average kinetic energy is related to the equilibrium temperature (known as the equipartition theorem) such that:

$$\langle E_{\mathrm{k}} \rangle = \frac{3}{2} \mathrm{N} \mathrm{k}_{\mathrm{B}} \mathrm{T}_0 \tag{2.31}$$

Where $\langle E_{\mathrm{k}} \rangle$ is the time-average kinetic energy and T_0 is the equilibrium temperature of a system.

To suppress fluctuations in the kinetic energy (and thus those in temperature), the system can be weakly coupled to a heat bath with constant temperature T_0 and force the effective temperature to decay exponentially toward T_0 with a time constant τ.

$$\frac{\Delta \mathrm{T}}{\Delta \tau} = \frac{\mathrm{T}_0 - \mathrm{T}}{\tau} \tag{2.32}$$

Whilst temperature might not be a parameter of the simulation, the factor λ can be obtained from this equation, which gives the amount to which the velocities of the particles in the system should be rescaled such that,

$$v \to \lambda v \tag{2.33}$$

where

$$\lambda = \left[1 + \frac{\Delta t}{\tau} \left(\frac{T_0}{T} - 1 \right) \right]^{\frac{1}{2}} \tag{2.34}$$

This is the weak-coupling method, also known as the Berendsen thermostat [129]. Due to the suppression of kinetic energy, the Berendsen ther-

mostat does not produce results consistent with the canonical ensemble. However, for sufficiently large systems in which the particles collide and transfer kinetic energy, the result roughly converge on the canonical ensemble. Thus, it is often used on explicitly solvated systems.

It is also key that the pressure, P, of the system remain at a constant; in particular for explicitly solvated systems using periodic boundary conditions. Here, the properties of the system should agree with those of the isothermal-isobaric ensemble (NPT). In order to control pressure, the Berendsen barostat [129]. Similar to the Berendsen thermostat, here the lengths in the system are scaled by a factor μ, such that the position of each coordinate is adjusted as

$$r \to \mu r \tag{2.35}$$

$$\mu = \left[1 + \frac{\Delta t}{\tau_{\mathrm{p}}}(\mathrm{P} - \mathrm{P}_0)\right]^{\frac{1}{3}} \tag{2.36}$$

where τ_{P} is the "rise time" of the barostat, a time constant, and P_0 is the target pressure of the system. P is the instantaneous pressure, approximated via the box volume V as

$$\mathrm{P} = \frac{1}{V}\left(Nk_BT + \frac{1}{3}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbf{F}_{ij}\cdot\mathbf{r}_{ij}\right) \tag{2.37}$$

Where $\mathbf{F}_{ij}$ is the force particles i and j exert on one another, and $\mathbf{r}_{ij}$ is the vector between them.

2.1.5.2 Langevin dynamics

Lacking the solvent, implicit solvent simulations will have far fewer collisions between atoms to aid in thermalisation, and hence the Berendsen thermostat is a poor choice here. Instead, the Langevin dynamics [130] approach is often chosen. As the generalised Born model only considers electrostatic screening, there are two degrees of freedom lacking in the system's dynamics. Firstly, friction with the solvent molecules is not considered, which normally exerts a force

$$\mathbf{F}_{\mathrm{F}} = -\gamma\mathbf{v} \tag{2.38}$$

on a particle with velocity $\mathbf{v}$, with γ being the Langevin friction constant, typically defined in terms of the frequency in solvent-solute collisions. As some of these collisions are of particularly high energy which cause perturbations to the system, there is an additional random force (due to the

fluctuation-dissipation theorem),

$$\mathbf{F}_{\mathrm{R}}(t) = \sqrt{2\gamma k_B T}\mathbf{R}(t) \tag{2.39}$$

where $\mathbf{R}(t)$ is a delta-correlated stationary Gaussian process with zero-mean, which satisfies

$$\langle \mathbf{R}(t) \rangle = 0 \tag{2.40}$$

$$\langle \mathbf{R}(t) \cdot \mathbf{R}(t') \rangle = \delta(t - t') \tag{2.41}$$

with δ being a Dirac delta function. This means that $\mathbf{R}(t)$ is an uncorrelated stochastic process in which at random intervals, individual "kicks" to the system occur which result in no net acceleration. Hence, the total force experienced by particle i at time t is given by

$$\mathbf{F}_i(t) = -\nabla V_i(t) - \gamma \mathbf{v}_i(t) + \sqrt{2\gamma k_B T}\mathbf{R}(t) \tag{2.42}$$

where V_i is the particle interaction potential. This stochastic differential equation accounts for solvent viscosity, which has a direct dependence on temperature, and thus can be used as a thermostat. The solvent viscosity should be kept as a small, non-zero value, such that the system doesn't become Brownian instead of inertial, which would allow no net acceleration to take place.

2.1.6 Constraints and restraints

In a typical biological system, the highest-frequency oscillation that occurs is covalent bonds that involve hydrogen atoms, as these atoms are particularly light. However, these are insignificant for the majority of systems, and thus it is often best to constrain these bonds — that is to keep them absolutely fixed — using the SHAKE algorithm, which removes these oscillations [131]. If these hydrogen bonds were allowed to oscillate freely, they would necessitate the use of an incredibly small time step to ensure the simulation remains stable. Hence, by fixing these bonds, larger time steps can be used, allowing longer simulations to occur.

Restraints are similar to constraints, but rather than holding an absolutely fixed value, they instead act to bias a coordinate towards a particular value whilst still allowing some fluctuation about this value via adding an additional term to the potential. These can be applied to restrain the lengths or angles of certain bonds to prevent undesirable behaviour occurring in the system, or to lead the system into a specific desired state. As a

simple and convenient method to modify the system's potential, restraints are applied often in advanced sampling techniques.

2.1.7 Umbrella sampling

Umbrella sampling is a method for calculating the change in free-energy between two states in a system. It was initially derived for Monte-Carlo simulations in the canonical ensemble [132], however it has since been extended and applied in molecular dynamics simulations in other thermodynamic ensembles.

The Helmholtz free energy, F, is a thermodynamic potential that measures the amount of reversible work performed by a system at constant temperature, T, and volume, given by [133]:

$$F = U - TS \tag{2.43}$$

Where U is the internal energy and S is the entropy. In a system, in order for a spontaneous transition between two states to occur, the free energy of the second state must be lower than that of the first.

Free energy can also be formulated in terms of the partition function, Z_{NVT}, so long as said system has constant volume and number of particles. In this case, the free energy has the form:

$$F_{\mathrm{NVT}} = -k_{\mathrm{B}} T \ln(Z_{\mathrm{NVT}}) \tag{2.44}$$

Where k_{B} is the Boltzmann constant.

However, the partition function is a multiple integral over 3N degrees of freedom and requires sampling the entire conformation landscape, thus making its calculation challenging:

$$Z_{\mathrm{NVT}} = \int \cdots \int_{3N} \exp\left(\frac{-U_{(x_1,\cdots,x_{3N})}}{k_{\mathrm{B}} T}\right) dx_1 \cdots dx_{3\mathrm{N}} \tag{2.45}$$

Fortunately, often the property of interest in a system can be condensed into a single reaction coordinate, be it a torsional angle or a distance between atoms. Thus, one can extract the partition function purely of the reaction coordinate of interest by multiplying the integral through a Dirac

delta function $\delta(x-x_0)$:

$$\begin{aligned} Z_{\mathrm{NVT}} &= \int \cdots \int_{3N} \exp\left(\frac{-U_{(x_1,\cdots,x_{3N})}}{k_{\mathrm{B}}T}\right) \delta(x-x_0) dx_1 \cdots dx_{3\mathrm{N}} \\ &= \int \exp\left(\frac{-U(x)}{k_{\mathrm{B}}T}\right) dx \end{aligned} \tag{2.46}$$

Thus, the Helmholtz free energy as a function of the reaction coordinate is:

$$F_{\mathrm{NVT}}(x) = -k_{\mathrm{B}}T \ln(\mathrm{Z_{NVT}(x)}) \tag{2.47}$$

Which is referred to as the potential of mean force (PMF) which simply means the free energy across a single degree of freedom.

To calculate the partition function, the internal energy of the system at every point of the reaction coordinate must be known. However, the free energy of any given state is related to said state's probability, P, such that:

$$\mathrm{P(x)} \propto \exp\left(\frac{-\mathrm{F(x)}}{\mathrm{k_B T}}\right) \tag{2.48}$$

Thus allowing the determination of the PMF by finding the probability of each value along the reaction coordinate.

This is theoretically possible by running a normal MD simulation and extracting all the values of x. However, frequently there will be an energy barrier in the PMF which prevents the system from adequately sampling the conformational landscape in any realistic time frame.

To overcome this, one can add in an potential $U'(x)$ which allows the system to overcome the potential barrier — a technique known as umbrella sampling. As the strength of the umbrella potential is known, it is trivial to extract it from the overall energy landscape calculated, giving the unbiased PMF,

$$F(x) = -k_{\mathrm{B}}\mathrm{T} \ln(\mathrm{P'(x)}) - \mathrm{U'(x)} + \mathrm{C} \tag{2.49}$$

Where C is an arbitrary constant.

The simplest scenario is one in which there is a known energy barrier, and thus a single simulation with the umbrella potential is required, allowing C to be discarded. Unfortunately this is rarely the case, as the free-energy surface is unknown before the simulations are performed. Therefore, it becomes necessary to perform multiple simulations, each with their own biased umbrella potential, leading from one another which are then

combined to construct the whole PMF.

Here, there would be numerous restraints along the entire reaction coordinate to pull the system along it. These are harmonic restraints based on Hooke's law:

$$U'(x) = (x_0 - x)^2 k \tag{2.50}$$

In which the only change made along each restraint is that of the equilibrium of said restraint, x_0.

Were the system undergoing purely Brownian motion, a normal distribution of x about x_0 would be expected. However, in these simulations, the underlying unbiased potential will affect this, resulting in the observed distribution deviating from the expected distribution. Hence, by performing multiple simulations following from one another in which x_0 is varied, with k being strong enough to sufficiently sample the whole energy landscape yet weak enough that each distribution overlaps, the whole range in a reaction coordinate can be sampled. Combining this data with equation 2.49 allows for the change in PMF over reaction coordinate to be extracted.

However, the value of the offset, C, in each window is different and each window should be weighted according to the favourability of the free energy landscape within its range. Fortunately, the optimal value of each can be determined using the weighted histogram analysis method (WHAM) [134]. Here, the values of x are divided into bins, such as when constructing a histogram, and a pair of coupled equations are iteratively solved

$$P(x) = \frac{\sum_{i=1}^{N} n_i(x)}{\sum_{i=1}^{N} N_i \exp\left([F_i - U_i'(x)]/k_B T\right)} \tag{2.51}$$

$$C_i = -k_B T \ln\left(\sum_{\text{bins}} P(x) \exp\left(\frac{-U_i'(x)}{k_B T}\right)\right) \tag{2.52}$$

Where N is the number of windows, N_i is the number of frames within the i-th window, $n_i(x)$ is the number of frames in which the value of the reaction coordinate is within the bin associated with x, and C_i is the offset of each window.

This techniques allows the PMF along the entire sample reaction coordinate to be computed. There are a number of other techniques that can be applied with molecular dynamics to study free energies such as metadynamics, replica exchange, and forward-flux sampling. These each have their own positives and drawbacks so which method is used will depends

on the system and interactions that are being studied.

2.2 Atomic Force Microscopy

Atomic force microscopy (AFM) is a type of scanning probe microscopy (SPM), a type of microscopy in which a probe moves across a surface which allows it to record an image of said surface [135]. Through careful design and accurately moving the probe, images on the nanometer can be obtained, enabling even deformations of the major and minor grooves of DNA to be viewed [136].

2.2.1 AFM operation

As opposed to optical microscopy, AFM "visualises" via a sharp tip attached below a flexible lever which scans over a surface, either the material of interest or a sample attached to the surface. As the tip is raster scanned across the surface, the surface topography can be constructed. This can occur in-air or in-liquid, ensuring that samples can be imaged in similar conditions to their native environments. As this is not an optical technique, the diffraction limit is not an issue, thus making AFM a single-molecule technique that is capable of sub-molecular imaging [136, 137, 138].

2.2.2 Tip-Sample interaction forces

The main forces which apply onto the tip during scanning are the long range attractive van der Waals interactions, electrostatic interactions, short-range repulsive interactions and, in air, capillary forces and the resultant adhesive forces. The long range van der Waals and electrostatic forces both act over tens of nanometres, and are caused by dipole interactions between atoms, and by Coulomb interactions between surface charges respectively. The strong repulsive force occurs at the Ångstrom range, as the overlap of atomic orbitals begins to dominate — where the tip is considered to be in contact with the sample. In liquid, it's possible to mitigate these electrostatic forces by using salt solutions, which screen these effects for tip-sample distances larger than a few Å.

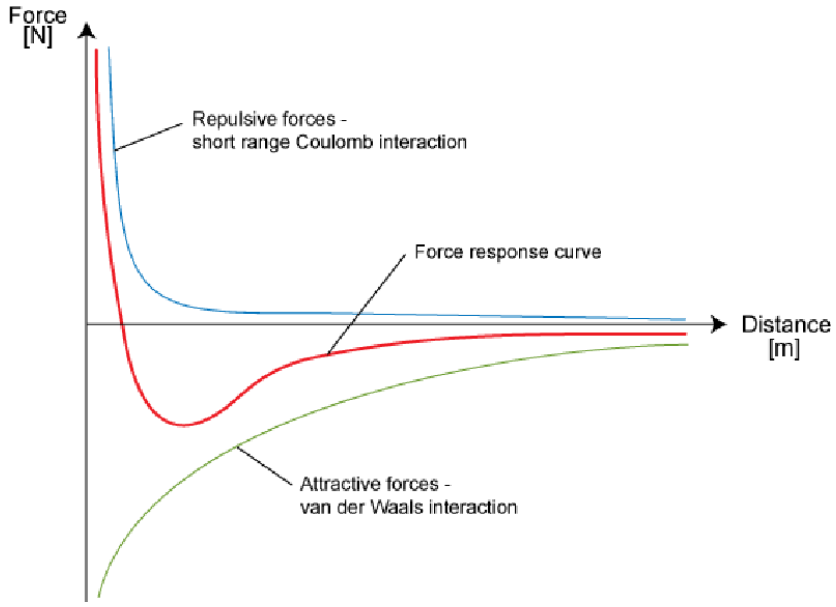


Figure 2.1: The distance-dependent forces which act on the AFM tip, ignoring electrostatics. At higher separations the attract van der Waals dominate, whilst the repulsive short-range forces dominate at lower separations. (image from [139])

2.2.2.1 AFM in liquid

As stated before, AFM is capable of imaging in fluid, thus allowing the probing and visualisation of biomolecules in their natural hydrated state - for example, DNA has a measured height in air of half that of in fluid (primarily due to the tip and sample/surface interactions). Beyond the biological advantage, the addition of fluid results in the tip-sample capillary forces experienced in air being avoided. These capillary forces occur when the humidity of the environment generates water layers on the tip and sample, causing a large force pulling the cantilever towards the sample and so increases the contact area resulting in lower spatial resolution [140].

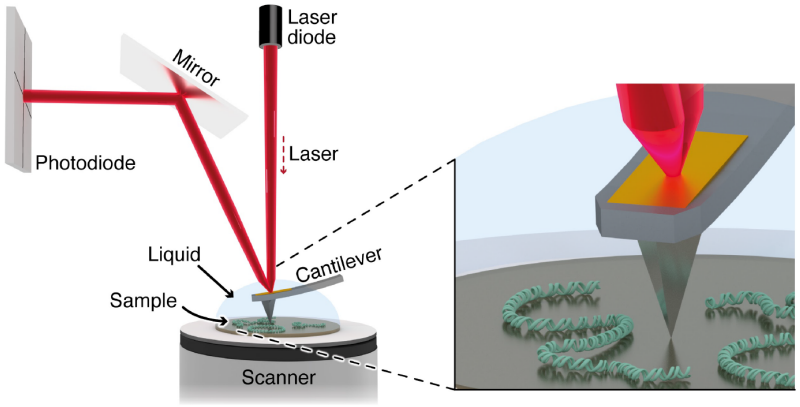


Figure 2.2: A schematic showing the operation of AFM in liquid. A sharp tip scans line-by-line across a surface with a sample (represent as DNA here), allowing an image of the surface topography to be built up. The topography is a function of the tip-sample interaction monitored via the bending of the cantilever, which is detected as a laser deflected off of the cantilever onto a quadrant-photodiode (QPD). The sample is mounted on a piezoelectric scanner for three-dimensional positioning with sub-nanometer accuracy. (image from [141])

Figure 2.2 shows an AFM setup for in liquid imaging. An AFM probe has a cantilever which itself has a small tip at the end that contacts the surface. This cantilever is moved by a piezostage and has a reflective top. A laser is pointed onto the reflective surface onto a quadrant photodiode (QPD) sensor which is zeroed for the cantilever at rest. Hence when the tip moves across a sample on the surface, there is a deflection in the reflected laser which is measured as a translation on the QPD.

The use of a fluid allows the electrostatic forces between tip and sample to be tuned and reduced, which allows for better force control and imaging resolution [142]. This follows the equation:

$$\kappa^{-1} = \sqrt{\frac{\varepsilon_0 \varepsilon_b k_B T}{2e^2 I}} \tag{2.53}$$

The Debye screen length (κ^{-1}) defines the distance over which the long range electrostatic force decays, based on the permittivity of a vacuum and the bulk solution (ε_0 and ε_b), absolute temperature (T), the ionic strength of the solution (I) and the electron charge (e).

2.2.3 AFM imaging modes

AFM has a number of different imaging modes available, the most commonly used being contact and tapping mode. More recently, a new imaging mode called peak force tapping has been developed which allows for high resolution imaging of biomolecules and is the main mode used to gather the AFM data throughout this thesis.

2.2.3.1 Contact Mode

Contact mode is the simplest and fastest method of AFM imaging, in which the tip of the cantilever is kept in constant contact with the surface. As seen in figure 2.2, a laser which is pointed at the back of a cantilever is reflected onto a split quadrant photodiode, allowing the deflection signal to be measured. This is then used as an input for a feedback loop that adjusts the height of the cantilever with respect to the sample, keeping the deflection signal at a predefined value called the setpoint. This means the interaction force between tip and sample can remain constant [143]. However, this results in large lateral forces applied to the sample. The deflection of the cantilever will also drift over time, resulting in much larger forces than intended being applied. Typically applied forces approximately range from 100 to 500 pN when using cantilevers with a spring constant of sub-0.1 nm[1].

Contact mode has been used to image biomolecules, such as purple membrane [144] but the large applied forces means the samples must be prepared such that they aren't an issue, such as being prepared as a flat surface or molecular array.

These issues subsequently lead to dynamic imaging modes being developed which reduce the forces applied to the sample, which ended up especially powerful for single molecule studies [145].

2.2.3.2 Tapping Mode

Tapping mode AFM is the most commonly used dynamic imaging mode. In this mode, the cantilever is moved into and out of contact with the sample, thus minimising the lateral forces applied. This technique is particularly useful for biomolecule imaging where the binding to the substrate is weak. In these situations, contact mode imaging would result in the sample being damaged or moved. If the sample is moved when imaged, there is a

loss of resolution and the tip can become damaged, which further reduces resolution.

The tip is oscillated in a sinusoidal manner above the sample at a frequency dependent on the cantilever — close to its the natural resonance f_0. As this occurs, the cantilever is raster scanned over the surface. The resonance frequency of a given cantilever is calculated using the effective mass m and stiffness k:

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{2.54}$$

The cantilever is usually driven mechanically by a piezoactuator — a device that converts electrical signals into precise, controlled physical movements. This motion is monitored as the RMS value of the deflection on the detector. Upon repulsive tip-sample forces occurring, the effective stiffness of the cantilever increases, resulting in its resonance frequency changing and thus energy is dissipated. With an actuation frequency f close to yet below the resonance frequency, both effects result in a reduction of the detected oscillation amplitude.

The tip-sample interaction will thus cause a reduction δA in the amplitude of the oscillation. In order to maintain the amplitude of oscillation A, a feedback loop is applied which adjusts the position of the cantilever with respect to the sample (or vice versa), which will yield traces of approximately constant tip-sample distance. The amplitude of oscillation in intermittent contact with the sample is defined by the setpoint A_{sp}, usually in liquid this is about $70-80\%$ of the free amplitude of oscillation. This allows softer biomolecules to be imaged as it reduces the lateral forces which would otherwise be applied.

However, there are disadvantages to using tapping mode imaging in fluid. The quality factor, Q, of the cantilever is reduced due to the viscosity of the fluid itself. This leads to a reduced sensitivity to changes in amplitude, as small shifts in f_0 produce much smaller δA than in air.

2.2.3.3 Peak Force Tapping Mode

Peak force tapping is a relatively recent imaging mode developed by Bruker (Bruker LTD, Santa Barbara) which performs a series of force curves at a significantly lower frequency than the cantilever's resonance frequency. In fluid, the tip is 'tapped' sinusoidally at amplitudes typically sub 10 nm, and frequencies of 1–8 kHz. When the probe interacts with the sample surface, the tip-sample interaction is controlled by keeping the maximum

force ('peak force') constant, see figure 2.3 for a schematic of tip motion in peak force tapping.

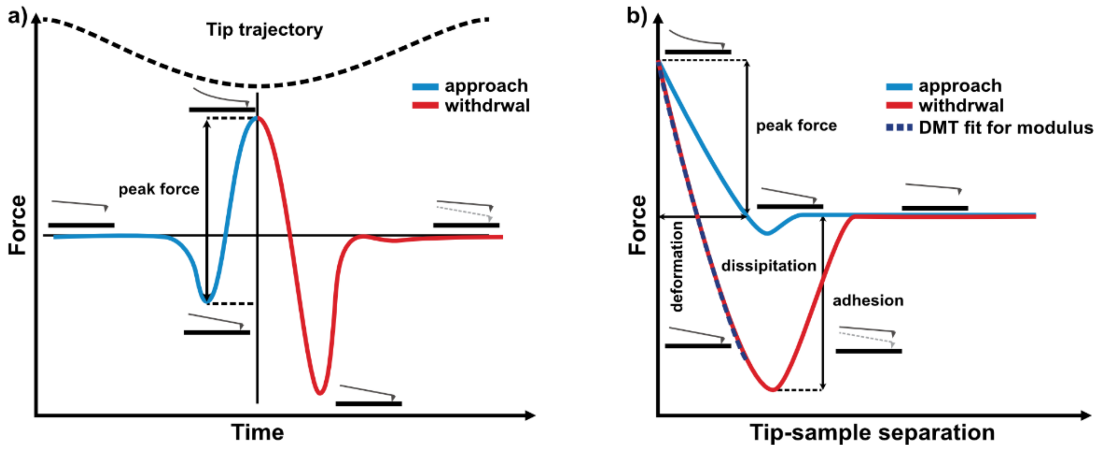


Figure 2.3: **a**) Scheme of the force curves for a cantilever operating in peak force tapping showing force (blue-approach and red-retract) and z-piezo (dashed) as functions of time. **b**) Force vs tip-sample separation plot showing the different parameters calculated in peak force QNM. (image from [146])

By considering the motion of the probe in terms of the Z position, a force curve at every pixel position on the sample surface is performed. This reduces the presence of higher-harmonic components in the deflection signal which would otherwise cause lower imaging quality due to ringing of the cantilever. Peak force tapping also contains algorithms which correct the hydrodynamic effects that occur on the tip.

Owing to the use of sinusoidal movements on the tip, the tip velocity as it approaches the sample is reduced. The force curves taken at each point on the surface allow precise control over tip-sample interaction forces, enabling imaging at peak forces as low as 30 pN in fluid. As the velocity and forces are kept low, both the tip and the sample are at a lower risk of getting damaged and is the key in achieving high-resolution imaging. Peak force tapping is also capable of recording nanomechanical information whilst recording topographical information, including adhesion, dissipation, deformation and elastic modulus [147]. Figure 2.3b shows a typical peak force curve and how these properties are calculated. Whilst not used in high resolution imaging, these can give other important properties of biomolecules [148, 149, 150].

2.2.4 Limitations of AFM

Under ideal conditions AFM could achieve atomic-level resolution in liquid, typically on flat surfaces such as calcite or muscovite mica. Mica in particular can be cleaved to reveal atomically flat planes with well defined lattices, which can be images with AFM to reveal individual atom locations [151], as can be seen in figure 2.4.

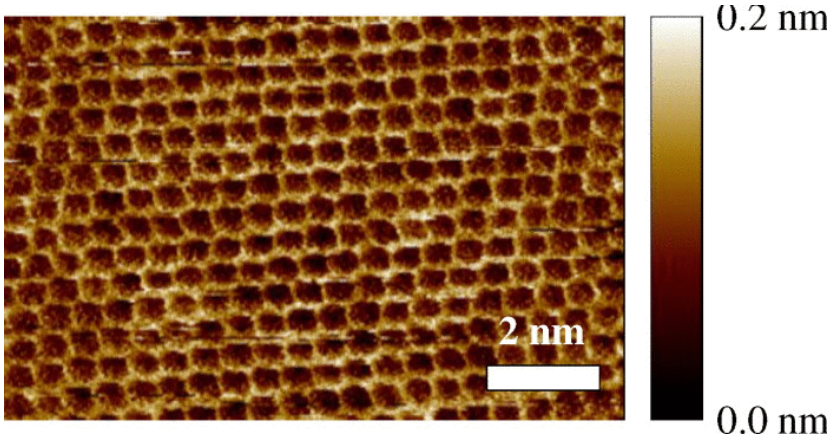


Figure 2.4: AFM imaging revealing the lattice structure of flat mica (image from [151])

However, whilst this is possible on mica, there are numerous complications which prevent this resolution being achieved on biomolecules. These issues include the binding of the molecule to a substrate, movement of the molecule, contamination within the sample, the effect of forces exerted on the sample, and the challenge in following the contours of complex and highly corrugated molecules whilst controlling the tip sample interaction and imaging force. The tip used will also have a large influence on the imaging quality [152] — the sharper the tip the better quality imaging of corrugated surfaces. In fluid, the viscosity of the liquid can also lower the imaging quality. The damping force exerted by the viscous liquid can cause the resonant frequency (f_0) and quality factor (Q) of the cantilever to decrease.

It should also be noted that as the molecule must be immobilised on a surface, there is limited capability of imaging dynamic events, and the surface chemistry may alter the results. Furthermore, AFM is an incredibly sensitive technique (which becomes even more so when performed in-liquid): disruptions such as air turbulence, electromagnetic interference and thermal noise will all distort the imaging. Whilst these issues can be addressed individually — the development of high-speed AFM allows for dynamics such as binding events to be viewed [153], and modifying the

microscope such as with acoustic isolation will dampen the noise on the machine [154] — these can have trade-offs in resolution or complexities in imaging setup.

2.2.5 Computational and experimental setup

2.2.5.1 Computational setup

All simulations used the AMBER software suite versions 18 to 20 [107, 155, 156]. Simulations were performed on local group clusters, the University of York HPC cluster Viking, and national Tier 2 clusters JADE2 and Bede.

Simulations of large constructs were solvated implicitly using the generalized Born model [157] at a salt concentration of 0.2M with GBneck2 corrections [158], mbondi3 Born radii set [159] with no cutoff which allows for molecular surfaces, salt bridges and solvation forces to be reproduced better [122]. Langevin dynamics were used to regulate temperature at 300 K with a collision frequency of 0.1 ps^{-1}, allowing greater sampling of conformational space via reducing effective viscosity. Shorter constructs were solvated explicitly, using a truncated octahedral TIP3P box, neutralised using either K or Na with Cl ions [160] (specific concentrations mentioned in each chapter).

In all simulations, DNA was represented by PARMBSC1 [161] whilst protein by the ff14SB [162] force fields. Whilst these force fields were developed independently, their consistent use of electrostatics allows their use to be combined. For PMF computation, the WHAM implementation by Alan Grossfield [163] was employed.

One of the main ways in which proteins interact with DNA is via hydrogen bonds, in which a donor atom (typically oxygen or nitrogen) that is covalently bonded to a hydrogen atom interacts with an electronegative acceptor atom. This can be easily defined when analysing a trajectory, one must simply locate potential acceptor-donor pairs which are within a certain distance from one another — in this work this distance cutoff was selected as 3.5Å. Hydrogen bonds are further restricted via the arrangement of the lone electron pair relative to the donor-H covalent bond, such that the donor-hydrogen-acceptor angle is required to fall within a range close to its maximum possible angle of 180° — here any angle above 120° was considered.

A second way in which DNA-protein interactions can occur is a salt

bridge. Salt bridges are defined in two ways, the solvent-separated ion-pair (SIP) where two like charges both form hydrogen bonds with the solvent, and the contact ion-pair (CIP) where two oppositely charged atoms make direct contact. CIP is similar to hydrogen bonds and do can be determined using the same 3.5Å cutoff, but with no angle requirement. SIP has been found to occur at distances between 3.8Å and 6Å [164]. For both hydrogen bond and salt bridge determination, the AMBER analysis tool cpptraj [165] was used, using either it's inbuilt hbond function, or distance function respectively.

2.2.5.2 Experimental setup

A 6mm mica surface was cleaved to be atomically flat with adhesive tape and initially prepared with 20 μL of imaging buffer with a composition of 3 mM $NiCl_2$ and 20 mM HEPES. This is because at neutral pH both mica and DNA are negatively charged, so the divalent cation functionalises the surface to allow for the DNA to stick. This was followed by adding the sample to the buffer and mixing it with a pipette. For bare DNA, 1–1.5 μL at 5 ng/μL was used as this results in good coverage without overcrowding the mica.

However, divalent ions cause HU and IHF to disassociate with the DNA. Hence, another method to functionalise the mica was chosen. Instead of using the NiCl2 imaging buffer, the mica was coated with 20μL 0.01% poly-L-lysine for one minute, which was then washed off under a stream of milli-Q. A 20μL solution of DNA and protein (with varying concentrations) was left to incubate for one hour at either room temperature or 37.5° C. Once prepared, the samples were imaged using either a Bruker Bioscope Resolve or a Bruker MultiMode 8 AFM using Peakforce QNM in fluid with Peakforce HIRS-F-B tips.

Images were then analysed using the TopoStats [166] AFM image analysis program. This performs automated image processing, skeletonises the DNA based on parameters given, which can then be analysed for statistics such as end-to-end distance, curvature, contour length and bend angle.

Chapter 3

HU-DNA interactions

Molecular modelling can be applied which allows for atomic-resolution studies on key DNA-HU interactions at length scales which can be probed experimentally. Whilst the functions of HU are well known, it is not well understood how the protein fulfils these roles. Here, molecular dynamics simulations were used which identified the existence of multiple binding modes resulting in varying bend angles, which were then quantified via umbrella sampling. Furthermore, the capability of HU to hold two strands of DNA was confirmed, and again the energy landscape of this interaction was calculated. These results were then experimentally verified through the use of atomic force microscopy (AFM), chosen as it allows for imaging of short DNA constructs, such as those simulated with implicit solvent. Additionally, simulations suggest a hopping mechanism via the β-ribbon arms of the protein to search along and between DNA strands.

Others' Contributions

- HU was provided by Michelle Hawkins and Jamieson Howard
- The code to interpolate and plot the free energies as a 2D and 3D heat map was provided by George Watson.
- The code to calculate bend angles on AFM data was implemented into TopoStats by Mingxue Du
- Daniel Rollins also performed AFM imaging on DNA and HU to provide extra data which is shown

The same DNA sequences for damaged and B-DNA in the implicit solvent simulations were used for AFM imaging.

3.1 Creating the HU structure

In order to characterise interactions between HU and DNA using molecular simulations, an initial structure consisting of *E. coli* HU$\alpha\beta$ (to match experiments) bound to damaged DNA had to be created. To do this, the PDB 1P78 [52] was taken, which contains HU bound to a short segment of damaged DNA (the damage being two flipped bases, two stacked bases and a T-T mismatch in the centre). As the crystal structure is from *Anabaena*, the protein had to be converted to that of *E. coli* HU$\alpha\beta$. PDB entry 4YEW [53], which contains a structure of the α-helical body of HU$\alpha\beta$ but not the β-ribbon arms, was taken, with the protein body being superimposed onto the model of 1P78 via rms fitting and connected to the remaining arms. The protein and central 15 bp were extracted, such that the DNA had minimal initial bending whilst remaining as close to the crystal structure as possible. Individual amino acids in the β-ribbon arms were then than corrected to those in the *E. coli* HU$\alpha\beta$sequence (a change of 9 amino acids per arm), and the resultant protein structure minimised. The protein and DNA segment were then embedded in longer 63 and 305 bp structures for explicit and implicit solvent simulation respectively. The sequences were designed based on previous work on IHF, which binds with an A-tract on the left and its consensus sequence on the right. In this work, the left hand side was used, and mirrored about the binding site to minimise the effect differing sequences may have for this work. Systems with B-DNA were also prepared in which the flipped and unpaired nucleotides were removed, and the mismatch corrected.The simulation parameters used were described in 2.2.5.1 with a 0.2M concentration of K and Cl ions for the explicit solvent simulations, with implicit solvent simulations also using a 0.2M salt concentration. The explicit simulations were used as a baseline, providing the most accuracy of the key protein-DNA interface, whilst the implicit solvent simulations played a two-fold role — firstly to provide much greater conformation sampling, and secondly as a direct comparison point for AFM experiments.

3.2 HU hops along DNA

In order to investigate if the facilitated diffusion model described in chapter 1 applies to HU, a structure was created in which the protein was placed ~ 30Å away from a 61 bp segment of DNA (as can be seen in figure 3.1) and this was left to simulate for 100 ns. This was done in explicit solvent using

a truncated octahedron water box.

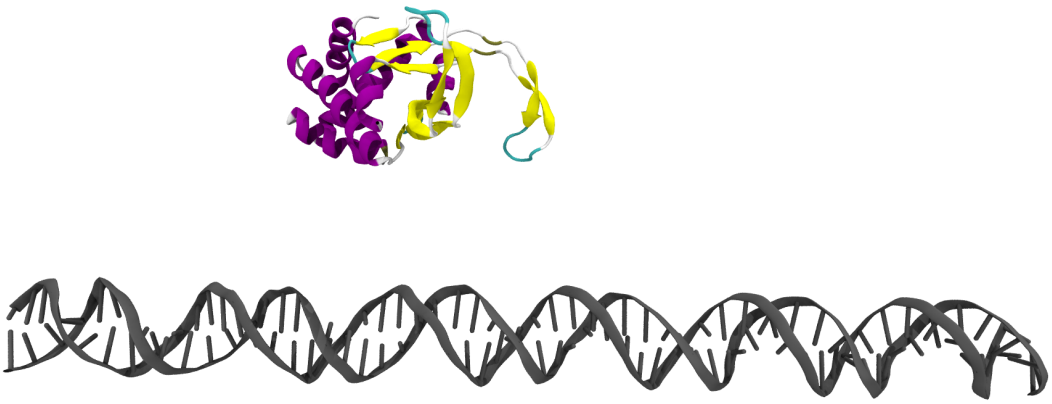


Figure 3.1: The starting structure to investigate the diffusion of HU near to DNA. Here the protein is ~ 30Å away from the DNA.

When the simulation was started, the protein immediately began to attract towards the DNA, with a β-ribbon arm approaching a major groove in the DNA. This interaction was short lived (within 1 ns) and the protein then diffused along the DNA with no apparent interactions for approximately 20 ns before again one of the β-ribbon arms interacted with the DNA. This interaction was with the minor groove in this instance, though it should be noted it occurs near the end of the DNA strand and begins close to where a major groove would exist — with this being about 20 bp away from the initial interaction site. To quantify this further, the simulation was analysed for hydrogen bonds, and these were plotted in figure 3.2 as the base pair that the interaction occurred at against time.

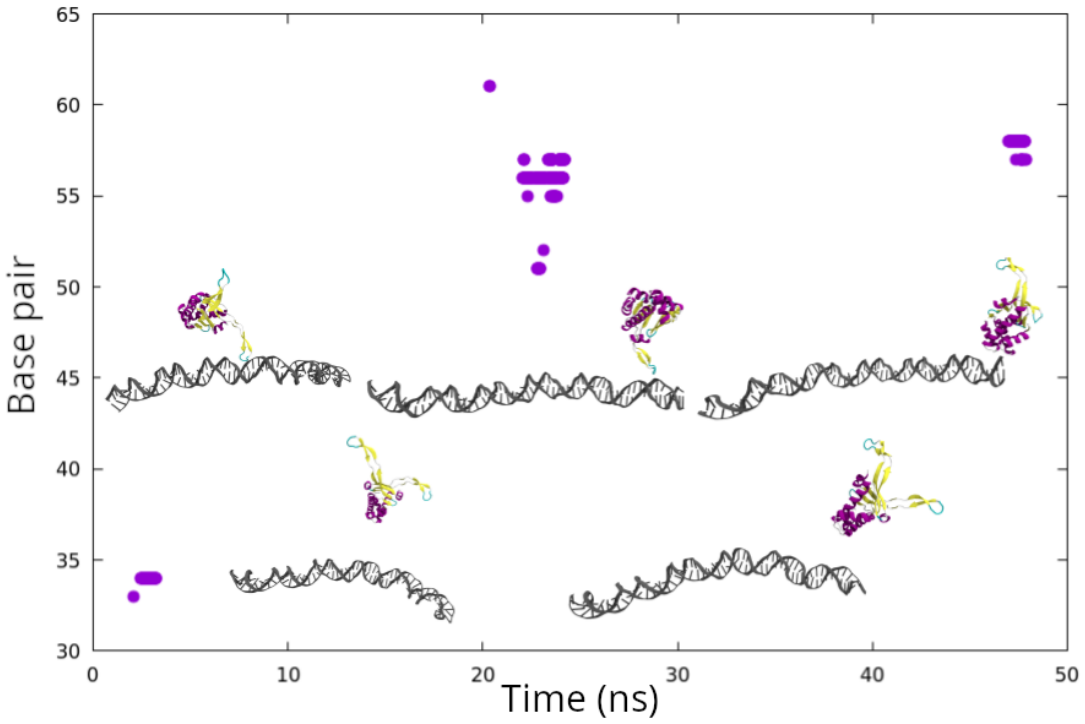


Figure 3.2: Base pair against time plot showing points at which hydrogen bonds between the protein and DNA occurred. It can be seen that at ~ 20 ns intervals the protein moves along where the interaction with the DNA is by about 20 bp. Example images of the simulation at these approximate points in time can be seen for visualisation of the hopping mechanism of HU.

A system was then created using a 150 bp segment of DNA with the protein. As the DNA was significantly longer, a long box was used in place of the truncated octahedron for the water box. To ensure the DNA did not cross over into the next water box and cause large scale self-interactions, the ends of the DNA were fixed in place with restraints. An example of the created structure (with ions to represent the water box) can be seen in figure 3.3.

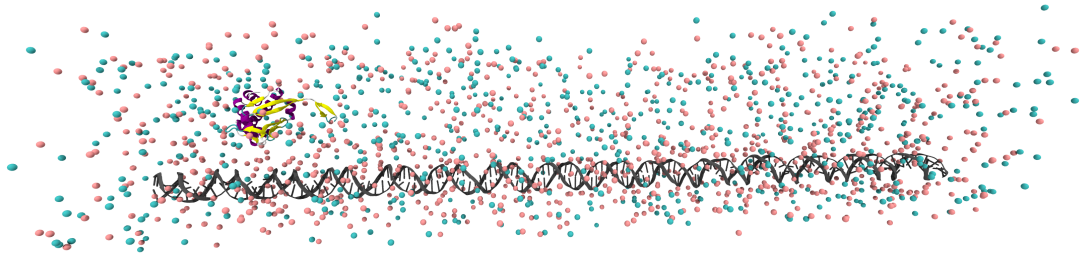


Figure 3.3: A 150 bp long strand of DNA was placed, again ~ 30Å away from a HU. The ions were left in this image to represent the long water box, with the K ions being in pink and the Cl being in blue.

Whilst there was an initial contact between the protein and the DNA (again, based around the β-ribbon arms), as the boundaries on the water box was quite small (15Å) the protein soon crosses the periodic boundary and interacts with the DNA on either side before entering a state resembling the initial binding state. The evolution of the system over time can be seen in figure 3.4.

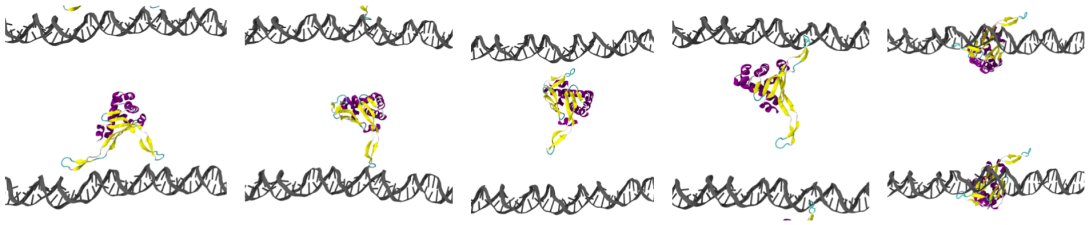
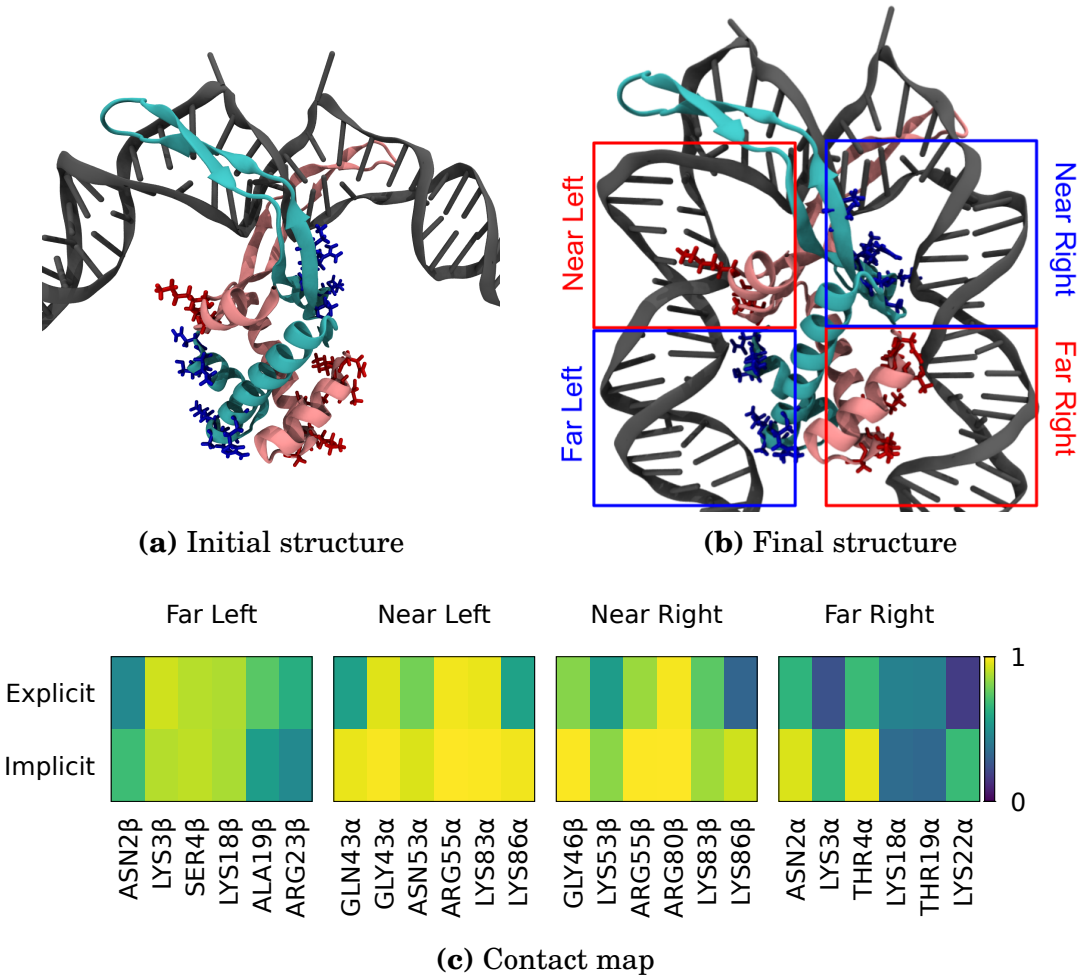


Figure 3.4: Initially the protein interacts with a strand using its β-ribbon arms, before one arm diffuses away toward the next strand along. The protein then moves towards the second strand, with the arms beginning to wrap around the grooves in the DNA. In these images, the periodic boundaries are represented using VMD.

3.3 Simulating HU bending DNA

Once the initial structures of HU bound to 63 bp and 305 bp DNA at a site of damage had been created, they were explicitly and implicitly solvated respectively. Simulations parameters are as described in section 2.2.5.1.



(a) Initial structure **(b)** Final structure

(c) Contact map

Figure 3.5: An initial structure in which HU was bound to kinked DNA (a), from which a fully wrapped final structure is produced (b). In these images, the HUα subunit is represented in pink whilst the HUβ subunit is in blue. The amino acids which interact with the DNA have also been shown using an atomic representation, with the DNA in dark grey. A contact map (c) showing new DNA-protein interactions beyond what could be seen from the crystallographic structure, separated into four sites. Note that this is the time-average number of hydrogen bonds formed by each amino acid, capped at 1.

In both simulations, a structure of the DNA fully wrapping around the protein was formed as can be seen in figure 3.5. However as there lacks a crystal structure of the DNA fully wrapping around the HU, the relevant hydrogen bonds using the α-helix core cannot be tested. Regardless, the fact that the system does evolve towards a wrapped state as found with IHF before whilst following a similar methodology, suggests that this result is valid.

As done previously in molecular modelling studies of IHF, the DNA-protein interactions were divided into four regions based on which side of

the protein the DNA interacts with. Here, the HUα and HUβ subunits were used to define which side is considered left and right, with the α helical section of HUα representing the "far right" section and so for HUβ the "far left". The near sites oppose this, with the "near right" using amino acids from the HUβ subunit and so "near left" using HUα. The previous IHF study used the original crystal structure in defining the interactions, however for this system the DNA captured in the structure was too short to capture any interactions with the protein's α-helical body.

3.4 HU exhibits multiple binding modes

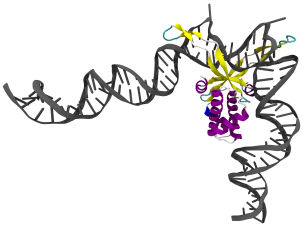
Whilst one implicit solvent simulation did confirm the existence of the fully wrapped DNA state, a wider variety of states were also found - aligning with previous studies on the possibility of a wider range of DNA bending modes. These were unveiled using the original implicit solvent simulation and an additional 3 replicas, each run for 100 ns.

To further define these potential states, hierarchical agglomerative clustering was applied to the central 60 base pairs and protein of the implicit solvent simulations (as the effect the protein has on the DNA is highly localised to this region). It was found that there are four distinctive states as the protein goes from fully unbound to fully wrapping the DNA, as can be seen in figure 3.6.

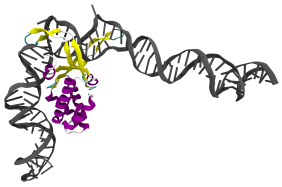
From these clusters, the bend angle of every frame in each cluster was calculated to generate a mean and standard deviation for each binding mode. To do this, the WrLINE molecular contour of the DNA was projected onto a best-fit plane using SerraLINE. From this, two vectors were found from the central point of the DNA and points 30 bp away, with the bending angle being calculated from these vectors.

(a) Three-quarters wrapped

(b) Fully wrapped



(c) Right-wrapped



(d) Left-wrapped

(e) Associated

Figure 3.6: Representative structures of different modes of DNA bending by HU. Hierarchical agglomerative clustering classified 5 different modes which showed different amounts of bending. It should be noted that the left-wrapped mode isn't the representative frame, but is instead a selected frame that most shows the fully wrapping of the left side of the DNA with no wrapping on the right.

	Associated	Half	Three-quarters	Fully wrapped
Bend angle / °	70 ± 12	97 ± 11	133 ± 12	161 ± 7
Percentage appearance	1%	27%	27%	45%

Table 3.1: Mean bend angle of different binding modes exhibited by HU

As the modes are similar to those found for IHF, they were named in accordance with the previous naming convention. The loosest wrapped state, termed the "associated" state, has the smallest bend angle at $\sim 70°$ with the interactions only occurring in the near subunits on each side. Whilst found to only exist in $\sim 1\%$ of the analysed data, it is notable as one of mul-

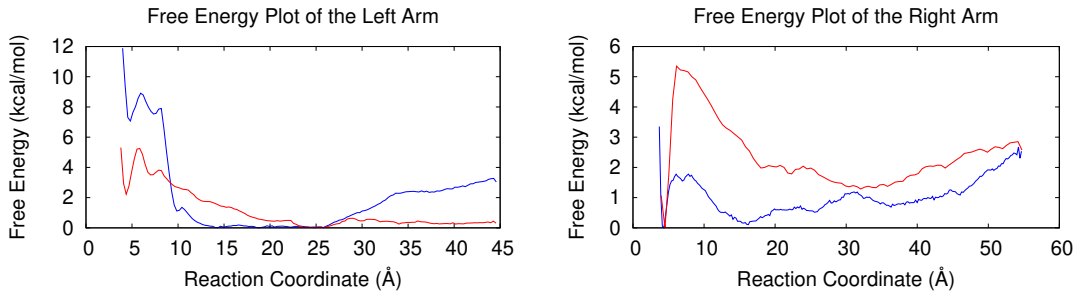
tiple possible pathways for the protein to enter from the unbound state. There were then two modes found, a mode where the right arm of the DNA fully bound and a mode where the left mode fully bound (respectively referred to as "right-wrapped" and "left-wrapped"). For bend angle analysis, these modes were considered as one, as it would not be possible to differentiate between them in the AFM imaging and the angle imposed on the DNA by the protein would be the same, and so these are analysed together under the term "half-wrapped". This mode was found to have a bend angle of $\sim 97^{\circ}$. An intermediate state, coined the "three-quarters" state, exists which wraps the DNA on all four subunits of the protein except the bottom right. This mode has a bend angle of $\sim 133^{\circ}$ and in the implicit solvent simulations performed, is the most common state at $\sim 45\%$ of the analysed data being in this cluster. Lastly, the largest bending angle ($\sim 161^{\circ}$) is "fully wrapped" state where all four of the defined subunits interact with the DNA.

An interesting observation from the explicit solvent simulation, is that the protein spends a long time in the three-quarters state before fully wrapping the DNA around itself. This is due to the position of the major groove of the DNA, where the major groove has to be pulled inwards to allow the DNA backbone to come into contact with the protein.

To further understand these binding possibilities, the underlying energetics of the system can be sampled. Umbrella sampling simulations were performed to obtain potentials of mean force of each arm of the DNA relative to their distance from the protein. In order to do this, reaction coordinates were defined as the distance between an atom from the base of the protein and on the DNA backbone on either side of the complex that interact when in the fully wrapped state, and NMR restraints were used to bias the distance between these reaction coordinates. For each side, the reaction coordinates used a phosphorus atom from the DNA backbone, and the Cα atom of the amino acids Ser 17α (for the right side) and Ser17β (for the left side). These were specifically chosen as the unbiased simulations found strong hydrogen bonding between these amino acids and the paired DNA base. The Cα atoms are chosen as they make up part of the amino acid backbone and hence should be less flexible than side-chain atoms, whilst P atoms from the DNA backbone were chosen as the interactions were predominantly betwen the DNA backbone as opposed to the base itself. Each reaction coordinate was reduced in 2 Å steps, with each window lasting for 5 ns and spring constants of 2 kcal mol^{-1}Å^{-2}. This occurred for distance of 3-43 Å for the left arm and 3-53Å for the right arm, with the final frame of

each window being used as the starting structure for the next.

For each arm, two sets of simulation occurred, representing the two extremes possible in the binding. In the first, as one of the DNA arms was pulled towards the protein, the other arm was held away. In the second, as an arm was pulled in, the other arm was already fully bound (using the final frame of the 5 ns window from the first set of simulations). Comparison between these should show the effect each arm had on the other during binding.

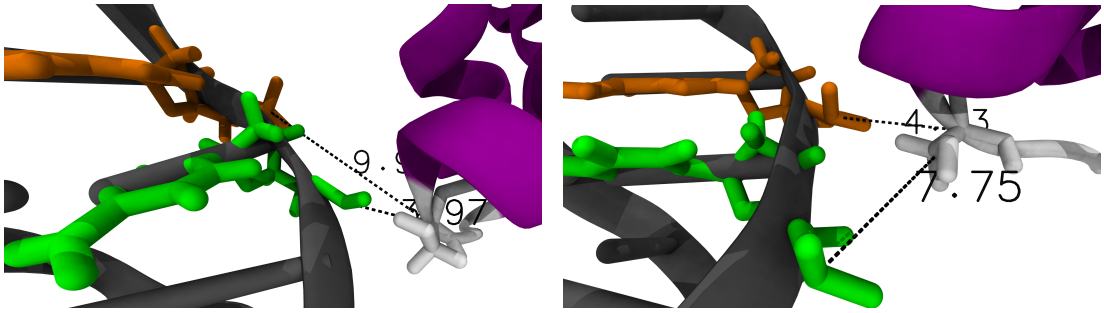


(a) Left **(b)** Right

Figure 3.7: Free-energy landscapes for the binding of DNA arms to HU. The left hand side (a) shows a very flat potential regardless of whether the right arm was bound (red) or held away (blue). Of note, when the right arm is held away, there is a large energy peak at ~ 9Å away. This occurs because a different DNA base interacts first and this must be broken to allow the reaction coordinate to come closer. The second trough then doesn't have adequate sampling to show its a further energy minima. The right arm (b) shows a similarly flat energy landscape, when the left arm is held away (blue) there is a barrier towards before hitting the final interaction.

There is a high peak before the final interaction in the case where the left arm is binding when the right arm is held away. This is because in the initial approach of the DNA to the protein, there is first an interact ion with a different base, and this must then be broken before the base used as a reaction coordinate is able to bind. This is similarly seen in the unbiased explicit solvent simulation, initially the left arm binds and the first DNA base binds. The right arm of DNA begins to pull towards the protein, and the electrostatic repulsion between each arm of the DNA causes the left arm to slightly unbind. It then rebinds using the next base (the one chosen as the reaction coordinate). This effect can be seen in figure 3.8. In the umbrella sampling, as the right arm of DNA does not get bought in, that knock pushing the left away does not occur, hence the large barrier to pulling the reaction coordinate in. Conversely, when the right arm binds

first, there is a flat potential dominated by fluctuations on the order of k_BT before a slight increase due to the electrostatic repulsion between the two DNA arms.



(a) First interaction **(b)** Second interaction

Figure 3.8: The unbiased explicit solvent simulation, where the initial interaction between HU and one of the DNA bases (green) can be seen (a), whilst (b) shows the second interaction which is formed with the second base (orange) which was used as the reaction coordinate in the umbrella sampling.

For the right arm, when the left arm is held away there is a relatively low free-energy landscape again, before an initial peak at $\sim$ 15Å away from the protein. This is representative of the DNA groove having to be pulled in to allow the fully wrapped state to occur, as seen in the unbiased simulations. In the case where the right arm binds after the left, the peak seen increases to $\sim$ 5 kcal/mol, which occurs again due to the electrostatic repulsion between the arms. This is clearly also seen in the unbiased simulations, in which the system gets caught in the three-quarters state for a notable amount of time before it is able to traverse the energy landscape into the fully-wrapped mode.

In order to fully represent any possible binding state along the reaction coordinates, a two-dimensional free-energy landscape would be required. However, it would not be feasible to simulate every possible combination of each arm's positions. Instead, the free-energy landscape can be estimated by considering each reaction coordinate to be orthogonal axes, with the two PMFs for each arm being taken to represent the extremes, and interpolating between them. Whilst this won't be entirely accurate, the positions of minima and maxima should be sufficiently reproduced in spite of inaccuracies in the exact values calculated for the free energy. Due to the issue with the left arm's reaction coordinate when the right arm is not bound, for the construction of free-energy landscape, anything lower than $\sim$ 10 Å in that PMF was taken as 0 and considered as "bound".

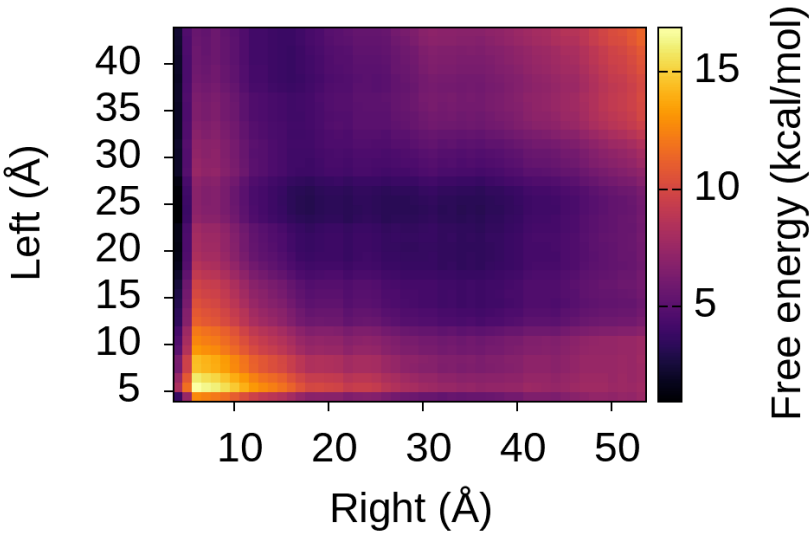


Figure 3.9: A 2D free-energy landscape of the HU-DNA interactions. It can be seen that the fully wrapped state does contain a minima, but it has a large barrier before. There is a general smoothness in the central area indicative of DNA breathing. Left and Right refer to the distance of each side of the protein and the DNA based on the reaction coordinates.

Figure 3.9 clearly shows the shape of such landscape. Though the suboptimal reaction coordinate choice for the left arm does bias the system somewhat to not be entirely representative of the true landscape, key details can still be gleaned from these. The most glaring features include the energetic peak blocking the fully wrapped state, but that the majority of the landscape is flat. This is indicative of a "breathing" mechanism through which the protein allows the DNA to go in and out as opposed to imposing a harsh lock on the DNA such as IHF. Whilst this doesn't occur in unbiased simulation, this may simply due to sampling limitations within the simulations, and once the fully wrapped state is formed there is still a strong barrier to leaving this configuarion. It would be possible to estimate the probability that each state would occur from this, and from that estimate the time scales in which these vibrations occurs. However, the created free-energy landscape will contain errors due to the reaction coordinate and so any estimates created would be inaccurate and so this wasn't performed.

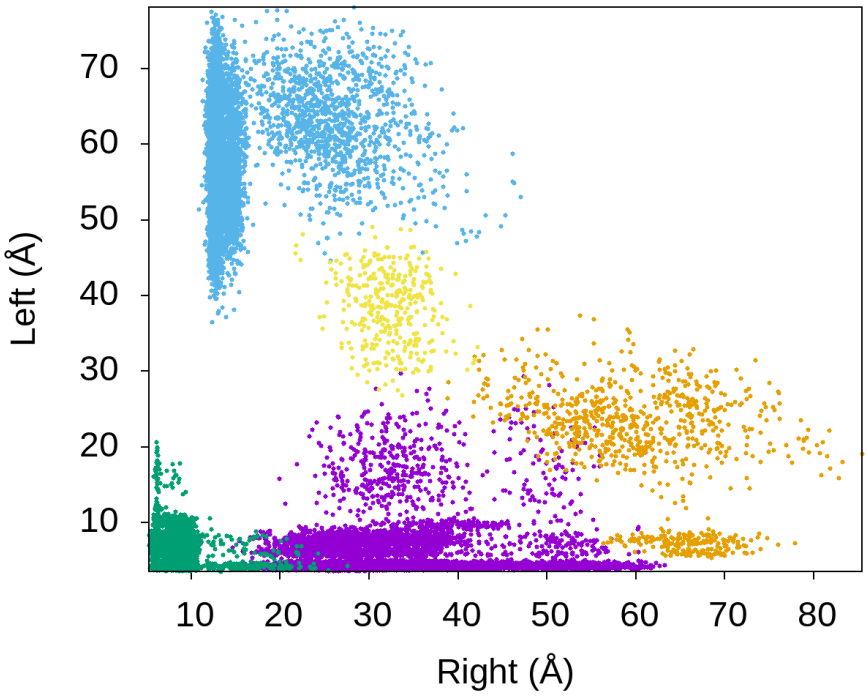


Figure 3.10: Positions of binding modes from clustering of the implicit solvent simulations on the conformational landscape. The fully wrapped mode (green) can be seen in the bottom left and is confined to a tight region. The left-wrapped (orange) and right-wrapped (blue) can be see in the bottom right and top left regions respectively, both occupying areas of tightly bound but also having points where the arm is only bound in the top half of the protein. The three-quarters state (purple) dominates most of the area along the right arm binding where the left arm is already bound, and the associated state (yellow) has a small region in the center of the conformation landscape.

To further validate the constructed free-energy landscape, the clustered populations from the implicit solvent simulations were taken and the distances of the reaction coordinate in each population was measured. These were then plotted as a conformation landscape as can be seen in figure 3.10. These clusters roughly line up with the minima that occur in the free-energy landscape, providing validation for both methods as results were generated using different simulation and solvation methods, on different systems.

3.5 Experimentally verifying HU-DNA interactions

3.5.1 DNA constructs ligation

In order to experimentally verify the simulations performed, the same 303 and 305 bp DNA sequences were constructed. Oligonucleotide sequences (Integrated DNA Technologies, Inc.) used to construct each sequence are given in appendix A.1. These were resuspended in TE buffer (10 mM Tris pH8 at room temperature, 1 mM EDTA) to a concentration of 1 mM. Due to the complementarity of the sequences, each sequence was annealed sequentially to create each construct at 95°C and left to cool overnight. These were ligated in 10X T4 DNA ligase buffer and left for 15 minutes at room temperature. This was then purified using a QIAquick PCR Purification Kit (QIAGEN), and some DNA was then run through a 1.5% agarose gel, as seen in figure 3.11.

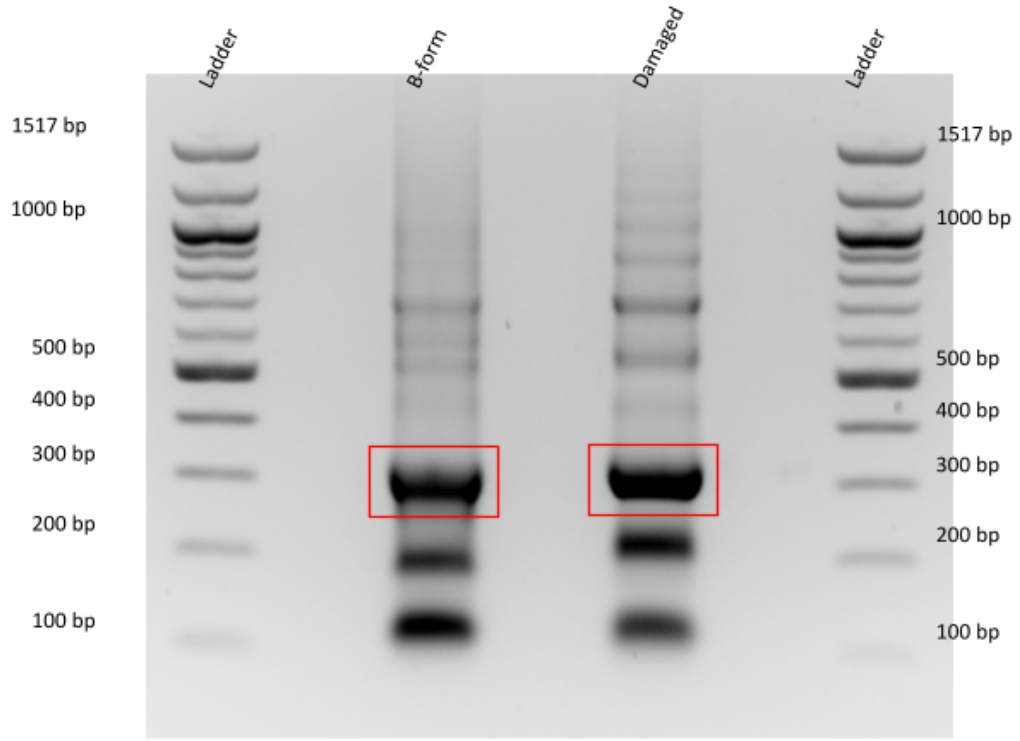
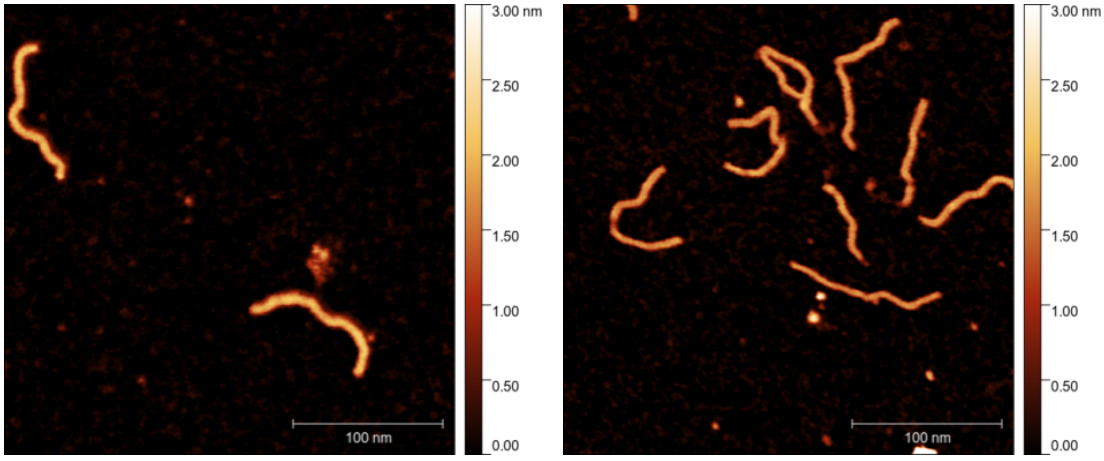


Figure 3.11: Gel electrophoresis of the two short constructs that were ligated to have either a corrected sequence or the damaged section. The two constructs are at their expected sizes.

As the bulk of the DNA was at the correct size, the rest of the DNA

was then run through another 1.5% agarose gel, and the ~ 300 bp band was taken and purified using a QIAquick Gel Extraction Kit (QIAGEN) and stored, giving a final yield of 225 ng/µL and 260 ng/µL of the damaged and B-DNA respectively. Finally, to ensure the site of damage had been correctly formed, each sample was imaged under an AFM using a $NiCl_2$ imaging buffer (which gives higher resolution imaging to PLL, which was used for all other imaging). As can be seen in figure 3.12, both the B-DNA and damaged samples formed at roughly the right lengths, and the damage can be clearly seen in the form of either a lower section or a kink, typically in the centre of a DNA strand.



(a) B-DNA **(b)** D-DNA

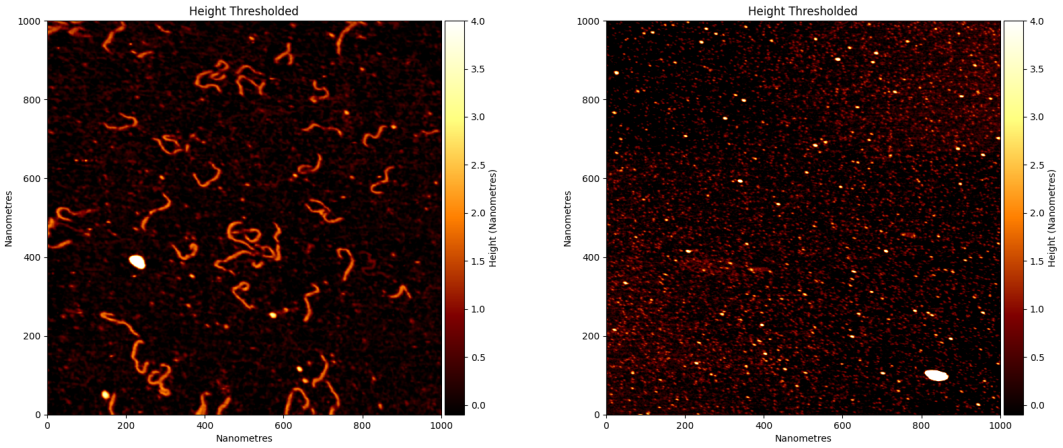
Figure 3.12: Once the DNA had been purified from the gel, they were imaged on an AFM using a $NiCl_2$ imaging buffer to ensure they were as desired. (a) shows how the B-DNA is clean and at the correct length, whilst (b) shows there are relevant damaged sections in the middle of the DNA.

The constructed DNA was incubated with HU and imaged using AFM. As stated in 2.2.5.2, PLL was used to functionalise the mica and a 10 mM Tris 250 mM KCl buffer was used to image the DNA and HU, as the $NiCl_2$ would cause the HU to dissociate from the DNA.

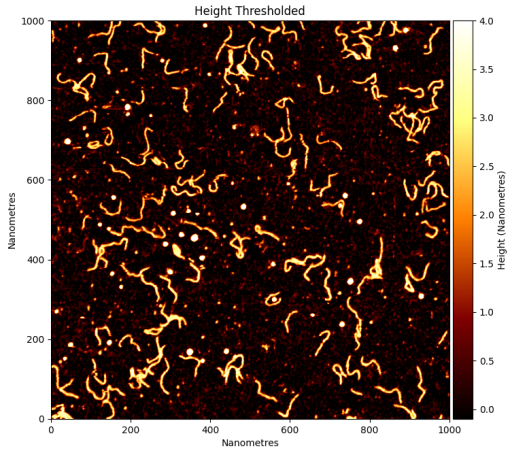
TopoStats was used which traces the DNA and can then calculated various statistics about the molecules tracked. Table 3.2 shows a drop in each property upon the addition of HU, suggesting that the conformational change induced in the DNA by the protein acts to compact the DNA. This is most clearly seen in the area, which has a 56.79% reduction.

	n	Contour Length (nm)	End-to-End Distance (nm)	Area (nm^2)
Damaged DNA	123	101.59±16.89	55.99±23.84	1099.66±388.62
With HU	318	86.99±6.78	53.95±23.32	624.57±94.37

Table 3.2: Properties of the damaged DNA constructs with and without HU. The data suggests a general compaction of the DNA by HU, however the effect on end-to-end distance is minimal. This can partly be attributed to having far fewer molecules without the protein than with.



(a) 5 ng damaged DNA alone **(b)** 1 ng HU alone

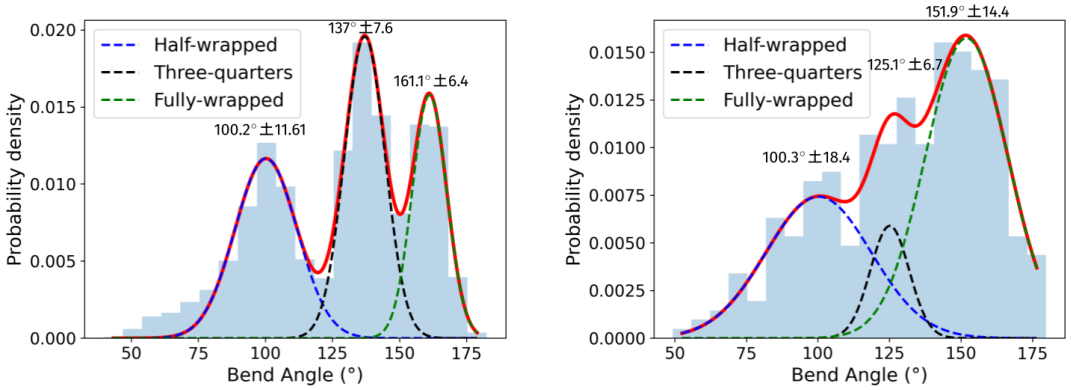


(c) 7.5 ng damaged DNA with 1ng HU

Figure 3.13: Example AFM images of the DNA (a), HU (b) and both incubated together (c). Typically when imaged together, less DNA will stick to the mica surface so the concentration of DNA is increased to ensure there is still sufficient molecules imaged.

As the protein is too small to appear in the AFM imaging when bound

to the DNA, the bend angle was calculated using the point of highest curvature in the central 20 nm of the DNA, and creating two vectors outwards from this point. Whilst this has a disadvantage of including unbound DNA in the sampling, as well as missing out potential non-specific interactions, the high specificity towards the damaged section should still be indicative of the protein interacting with the DNA.



(a) Molecular dynamics **(b)** Atomic force microscopy

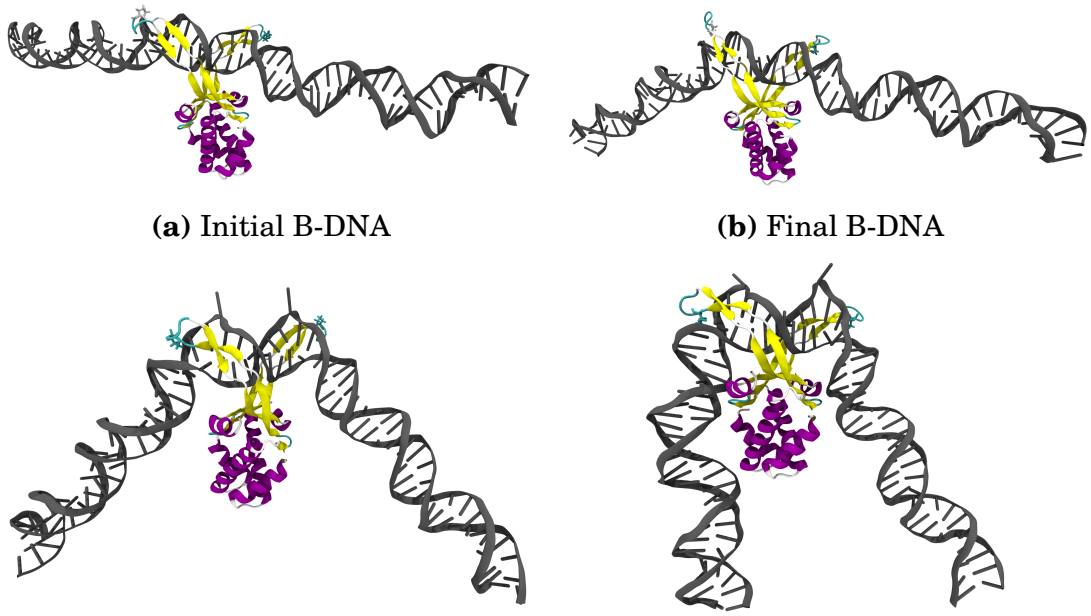
Figure 3.14: A comparison between the bend angles found in the implicit solvent simulations with the first 5 ns removed (a), and those found in the AFM imaging (b). The substates shown as dashed lines where each state was found with a Gaussian fit.

As can be seen in figure 3.14, there is a reasonable agreement between the simulations and the experiments. The simulations show a strong three mode distribution matching to those of the different binding modes found through the clustering, with AFM imaging suggests a similar three modes. One thing to note is that whilst the peaks on each plot are different, for the simulations the bend angle for each frame after the first 5 ns of each replica was taken and plot in the histogram, whilst for the AFM it was each individual molecule so they cannot be compared in this way, but instead that the location of each peak is relatively similar. Also to note is that whilst the bend angle is lower in the AFM than the MD, this could be due to the effect where large bend angles are underestimated in AFM due to the limited resolution [167].

3.6 Proline intercalation

A key question with the DNABII proteins is whether the DNA bends to allow the prolines to intercalate, or whether the intercalation causes the DNA to bend. In particular for HU, why the protein binds specifically to

damaged DNA and whether the canonical mode is possible for normal B-DNA is also unknown. To investigate these, structures were created of HU bound to the DNA, but the prolines were pushed out using NMR restraints.



(a) Initial B-DNA **(b)** Final B-DNA

(c) Initial damaged DNA **(d)** Final damaged DNA

Figure 3.15: Structures were create where the prolines were pushed out of the DNA using NMR restraints. For undamaged B-DNA, the prolines were unable to intercalate after 500 ns and the overall structure didn't change much. However for the damaged DNA (with the extra flipped and unpaired bases and a mismatch), the prolines intercalated and the DNA wrapped around the protein.

These systems were simulated for 500 ns in explicit solvent. For the undamaged B-DNA, it was found that the prolines were unable to intercalate in-between the DNA. Hence, whilst there was still some small bending of the DNA, the DNA remained in a mostly linear, as can be seen in figure 3.15b.

Meanwhile, for the damaged DNA, the prolines did intercalate into the DNA within the first 5 ns of the simulation. Beyond this, the DNA did then wrap around the protein as expected. However this binding was notably less strong than when starting with the perfect canonical structure based on the crystal structure. Here, it was seen that the DNA would interact with the far sites of the protein, but this was unstable, instead binding and unbinding throughout the simulation. This suggests that the intercalation is what allows the hinge-like mechanism to be formed for the DNABII proteins. Here, the fact the intercalation occurs so quickly in damaged DNA suggests that there does not need to be much initial bending and instead

the intercalation produces the bending.

3.7 Switching between non-specific and specific binding

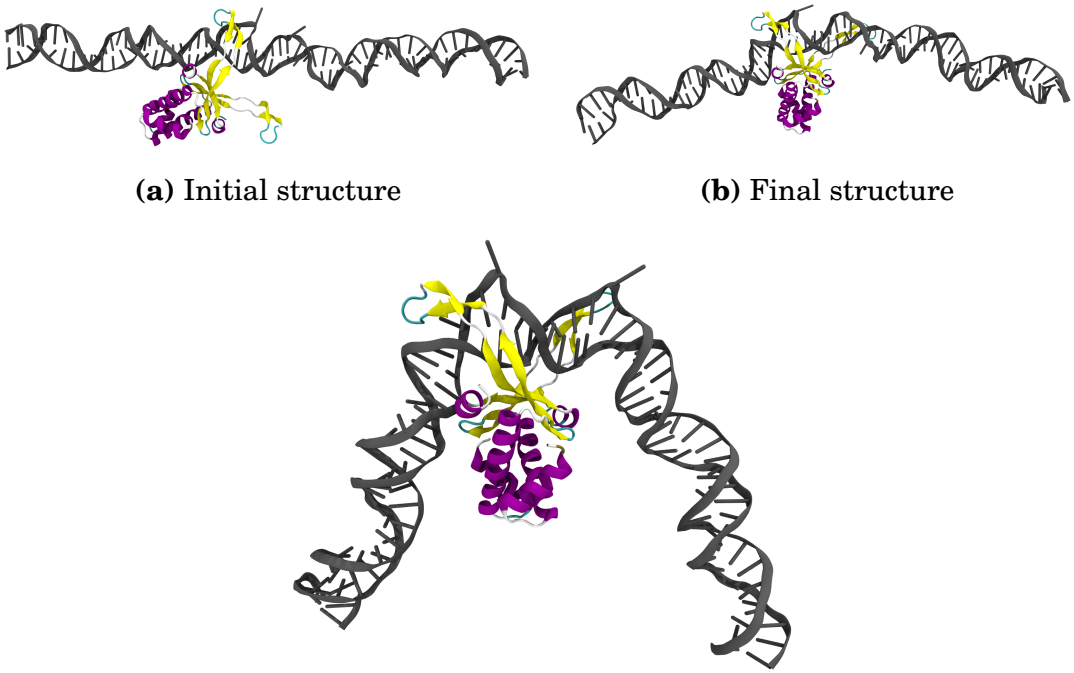
X-ray crystallography revealed a non-specific binding mode for HU, in which the α-helical core of the protein bound parallel to the DNA. The dynamics in this mode are of interest as it is suggested that this is the mechanism through which high concentrations of HU aid in nucleoid compaction. Hence, two structures were created that extended the original structure to have a 61 bp segment of DNA bound.

(a) Initial structure **(b)** Final structure

Figure 3.16: A structure was created by extending the DNA of the 4YFT PDB (a). This was then simulated for 500 ns, with the final result being shown in (b). The protein did bend towards the DNA slightly, but there wasn't a significant change in the DNA beyond slight curvature.

As can be seen in figure 3.16, there was a slight bend in the protein and in the DNA towards one another but there wasn't much change in the overall dynamics of the system.

A second structure was also created with the 63 bp damaged DNA, where the protein would hit the damaged site as it was wrapping the DNA around itself. This was done to view whether the recreation of the canonical binding mode could be reproduced from this initial structure as it would show that it is possible for HU to switch between non-specific and specific binding modes.



(a) Initial structure

(b) Final structure

(c) Almost canonical structure

Figure 3.17: A similar structure was created using a structure containing the damaged DNA (a). Initially, the protein began to more fully wrap the DNA around itself, appearing the resemble the canonical binding mode (b). However, the second β-ribbon arm was unable to fully wrap around and intercalate into the opposing groove, even when the simulation was extended to 2 µs, and the structure unwraps somewhat (c).

Figure 3.17c shows that the HU does initially begin to enter a mode reminiscent of the canonical binding (figure 3.17c). However, the second β-ribbon arm is unable to really wrap around the DNA to intercalate the second proline, which eventually leads the to the DNA unwrapping again (figure 3.5b). This is likely a sampling issue with molecular dynamics, where the intercalation of the second proline has an energy barrier the simulation cannot overcome. However, we can infer from these simulations that such a transition is possible, showing that HU (and likely other architectural proteins) can employ a non-specific binding mode to search DNA and should they encounter a binding site, switching into a specific binding mode is possible.

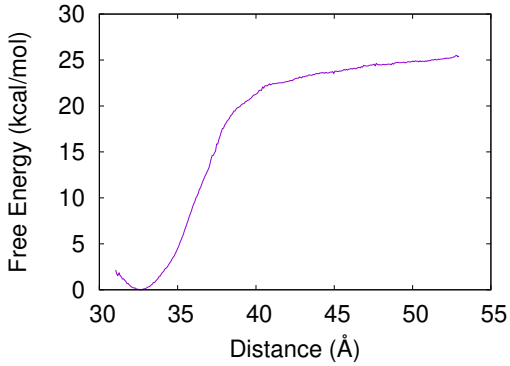
3.8 DNA-HU-DNA bridging

As stated in 1.2.2, previous crystal structures have observed the capability of HU to bind to two DNA strands non-specifically [53]. This is ob-

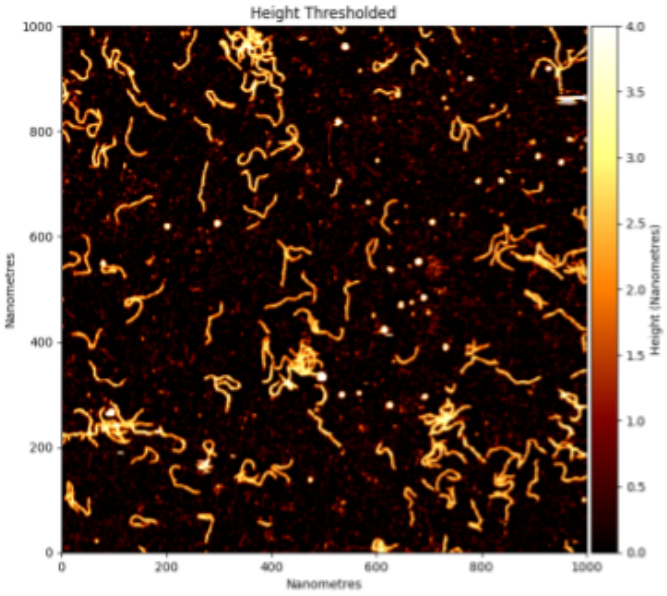
served to occur between the negatively charged DNA backbone and positively charged surfaces on the protein. It has also been found that IHF is capable of bridging two strands of DNA where the main strand containing the binding site is mostly unwrapped whilst a distal strands binds to positively charged amino acids on the other side [35]. These bridging capabilities are viable explanations for the key roles the DNABII proteins play in biofilms, and hence these are of relevance to this study.

In order to investigate, the 4YFT pdb was extended and a second strand was placed on the opposite side of the protein. This system was explicitly solvated and allowed to relax for 500 ns in order to allow a favourable state to be found, the result of which can be seen in figure 3.18a. Umbrella sampling simulations were then performed which slowly pushed the second strand of DNA away from the protein, choosing a backbone atom closest to the centre of mass of the protein and in the DNA backbone. Increasing 2Å increments over a series of 5 ns windows were applied until the PMF was found to plateau.

(a) A bridge formed around the α-helices



(b) Free energy landscape



(c) AFM data shows DNA aggregation by HU

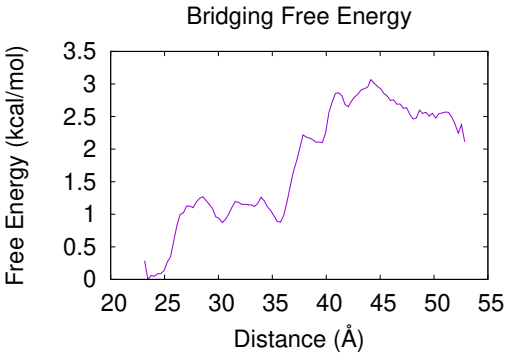
Figure 3.18: (a) shows the DNA-HU-DNA bridge that was formed spontaneously and was then pushed apart using umbrella sampling, (b) showing this state is very energetically favourable with an energy reduction of ~ 25 k_BT. (c) shows clustering of the damaged DNA samples and HU found in AFM imaging (10 ng and 1.5 ng of each respectively).

This mode was found to be very energetically favourable, with the resulting free-energy change to be of $\sim 42 k_{\mathrm{B}} T$ from the bridge to the unbridged state. This is significantly larger than what was found for any of the wrapped states, indicating that given a close enough strand of DNA, HU will bridge these rather than enter the canonical binding mode. A potential avenue of study would be to study the free energy as HU enters the canonical binding for a more direct comparison between the two, but

this was not investigated here. It is also interesting to note that as this non-specific binding mode was suggested to occur at high concentrations of HU, the bridging effect induced should be significantly stronger *in vivo* than was calculated here as there will be multiple HU holding the strands together, not just the single protein simulated here.

As IHF was found to bridge DNA strands with a free energy$\sim 24k_BT$ using a different bridging mechanism, a similar system was create for HU. Here, the initial structure of HU was taken and a second strand was created perpendicular to the original strand and was pulled towards the protein. Then, this system was again explicitly solvated with umbrella sampling used to push the strand away. However, here it was found that the PMF of the system was significantly lower at ~ 3Å. Whilst the minima was during the bridging interaction, the low free energy relative to when they were unbound implies this is unlikely to happen in a real system. This is likely as the DNA here is damaged, making is more flexible. Hence, the canonical strand is more likely to bind to the HU quickly, and electrostatic repulsion between the strands would prevent the second strand from naturally binding to the protein.

(a) A bridge formed around the α-helices



(b) Free energy landscape

Figure 3.19: Initially, a structure was created where HU is bridging DNA in the same mode that IHF does (a), however umbrella sampling revealed this mode is not very energetically favourable, likely as the DNA starts compacting around the protein and conflicts with the second strand.

3.9 Summary

HU's specificity for damaged DNA makes it unique amongst NAPs, adding an additional functionality towards DNA repair beyond the standard genome organisational roles of many NAPs. Here it has been shown that HU ex-

hibits multiple metastable binding modes, and yet has a more general DNA breathing mechanism revealed via umbrella sampling simulation even though there was a questionable choice of reaction coordinate for the left arm. In particular, the preference for damaged DNA can be explained by the results showing that the intercalation of the prolines can only occur in the instances where the DNA is damaged, with the protein becoming unable to intercalate with normal B-DNA.

Perhaps the most novel result would be viewing the mechanism through which the protein travels along DNA, with a direct view of how HU hops along and between DNA strands. This could be extrapolated to other architectural proteins, such as IHF. The bridging capability has also been shown to be highly favourable, explaining the method through which HU stabilises extracellular DNA and compact the genome. Interestingly, the mechanism through which IHF was found to bridge DNA was shown to be unlikely for HU, though no work was performed to validate the binding mode used here for HU on IHF.

Chapter 4

Evolution of DNA:protein specificity

As explained in chapter 1, DNA-protein specificity is of significant interest as it underlines the fundamental basis through which proteins perform their roles in organisms. Of particular note is how this specificity evolved, with many studies considering how changes in DNA occur to allow a new target gene to interact with existing transcription factors [88]. Here, we instead study how both the DNA and protein may have evolved concurrently by studying a pair of evolutionarily related proteins (ParB and Noc). We use molecular dynamics simulations to investigate how they bind to their respective DNA sites (*parS* and *NBS*) (with relevant amino acids and DNA sequence seen in figure 1.9). In particular, X-ray crystallography experiments were unable to resolve the side chains of the key lysine mutations that occur, bringing into question the role that these amino acids were playing, so here the simulations were used to resolve and explain the interactions occurring.

The simulation parameters used were as described in 2.2.5.1 with a 0.1M concentration of NaCl ions. For simplicity in nomenclature, mutations are referred to as the initial wild-type ParB amino acid, followed by the residue number, then the mutated amino acid, such that R173Q represents an arginine at residue 173 mutating to be a glutamine. As the crystal structures contain a dimer, the resulting analysis of interactions was an average of the interactions of both dimers. These simulations were run for 200 ns, with the first 10 ns being discarded as an equilibration step for the system to enter a more favourable state.

Others' Contributions

- Adam Jalal and Tung Le provided the crystal structures and complementary deep mutational scanning data.

4.1 Modelling wild-type ParB-*parS* and Noc-*NBS* interactions

Whilst the key amino acids involved in the recognition of ParB to *parS* and Noc to *NBS* had been identified by our collaborators Dr. Tung Le and Dr. Adam Jalal, a key question that remained was what the role of each mutation was. In particular, crystal structures were unable to resolve the type of interaction the T179K and A184K mutations lead to between the DNA and Noc. Thus, to investigate this, these structures were investigated using molecular dynamics simulations.

The four specificity amino acids were then analysed to study what interactions were occurring in each system. Initially the ParB-*parS* system was simulated, and the four specificity amino-acids were analysed for their interactions with the DNA, as can be seen in figure 4.1. The R173 amino acid maintained a hydrogen bond with *parS* guanine 1 for the whole analysed time. G201 from ParB was found to interact with thymine -6 specifically, but this is only maintained for $\sim 55\%$ of the simulation. Here, it was seen that whilst there is an initial interaction between the base and amino acid, the bond breaks. Meanwhile, both T179 and A184 form no bonds with the *parS* DNA throughout the whole simulation. These can be seen in figure 4.2.

Initially, just hydrogen bonds were searched for, however in the Noc-*NBS* system it was found that the highly conserved lysine mutations did not maintain these interactions. Instead, a system to define salt bridges was programmed and used as described in 2.2.5.1, which revealed solvent-separated ion-pair salt bridging for over 99% of the analysed time between both K179 and K184 and the DNA phosphate backbone, with this being an example of indirect readout as the DNA base itself is not involved. Both Q173 and R201 also showed strong interactions with the DNA base itself, as can be seen in figure 4.3.

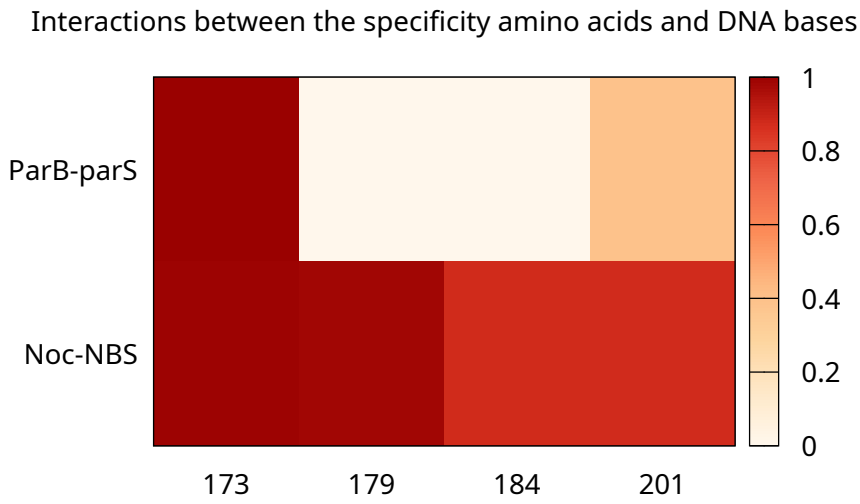
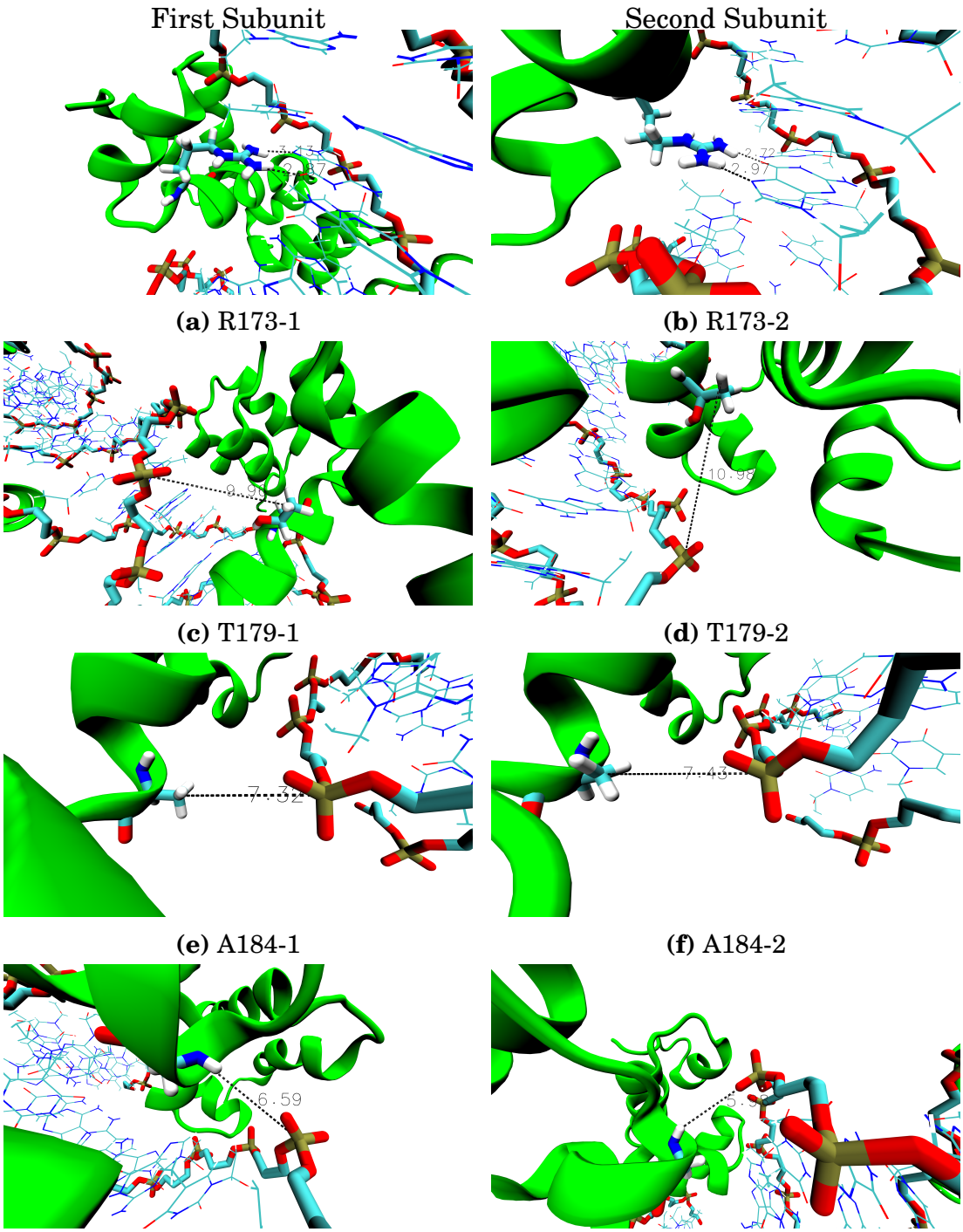


Figure 4.1: How long interactions between the four specificity amino acids and the DNA are maintained throughout the simulations. These are analysed as hydrogen bonds except for the 179 and 184 in the Noc-*NBS* interactions which are as salt bridges.

The long lasting interactions between the *NBS* site and the K179 and K184 (over 99% of the simulations), and the clarification that they occur via salt bridges, was not attainable from the x-ray crystallography data. The initial structure gathered was not of a high enough resolution to capture these interactions, needing the simulations to back up the experimental data. This also leads to another suggestion — that these mutations were permissive mutations to allow other mutations to interact with the DNA. As the interaction was non-specific with the DNA backbone, these mutations provide a general increase in DNA-binding affinity in the protein. This decreases the initial energy barrier towards binding which allows the other base contacts to occur, suggesting such a mutation must occur early in the evolutionary pathway.



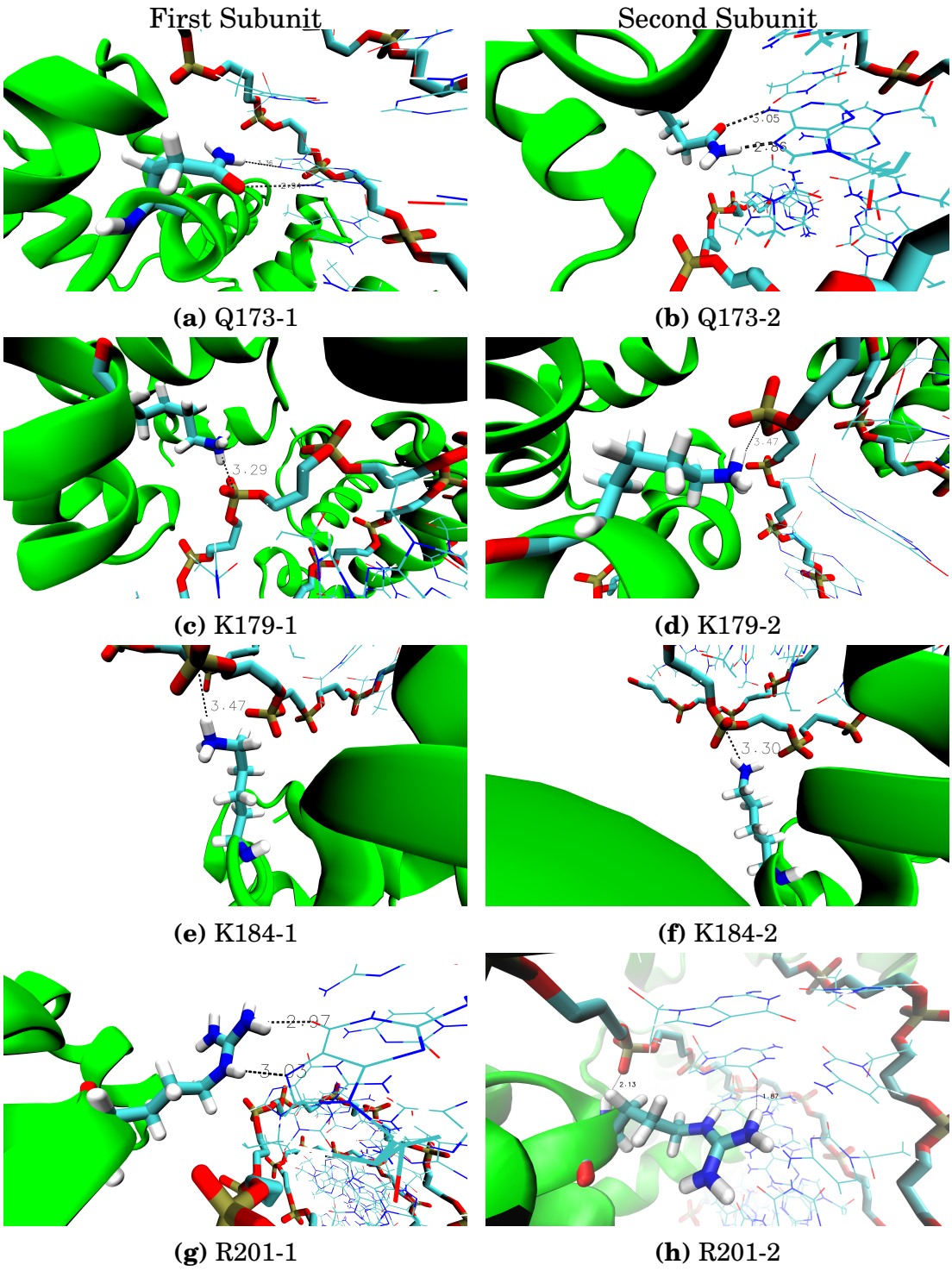
(a) R173-1 **(b)** R173-2

(c) T179-1 **(d)** T179-2

(e) A184-1 **(f)** A184-2

(g) G201-1 **(h)** G201-2

Figure 4.2: The interactions (or lack thereof) between the four specificity amino acids and their relevant DNA bases in the ParB-*parS* complex can be seen here. (a, b) show the hydrogen bonds formed by R173 on each subunit with the G1 base. (c,d,e,f) show that T179 and A184 form no interactions with the DNA on either subunit. (g, h) show the end state of the G201 interaction, which initially does interact with the DNA but the bond gets broken as the simulation progresses.



(a) Q173-1 **(b)** Q173-2

(c) K179-1 **(d)** K179-2

(e) K184-1 **(f)** K184-2

(g) R201-1 **(h)** R201-2

Figure 4.3: The interactions in the Noc-*NBS* complex. (a,b) show the interaction between the Q173 amino acid and the A1 base. (c, d, e, f) show that the K179 and K184 amino acids form interactions with the DNA backbone as opposed to directly to the base, whereas these amino acids did not in the ParB-*parS* system. (g, h) show the interactions R201 forms with G6 on the DNA bases.

4.2 Mapping the evolutionary pathways

In order to study how this change in specificity may have occurred, the different potential evolutionary pathways were mapped to provide the systems to create and simulate. These can be seen in figure 4.4. In order to reduce the number of configurations to simulate, it was considered that each lysine mutates together (as considering each lysine individually increases the number of possible configurations enormously). This is also as it has been shown that only one lysine mutation is necessary to support the switch in binding to NBS.

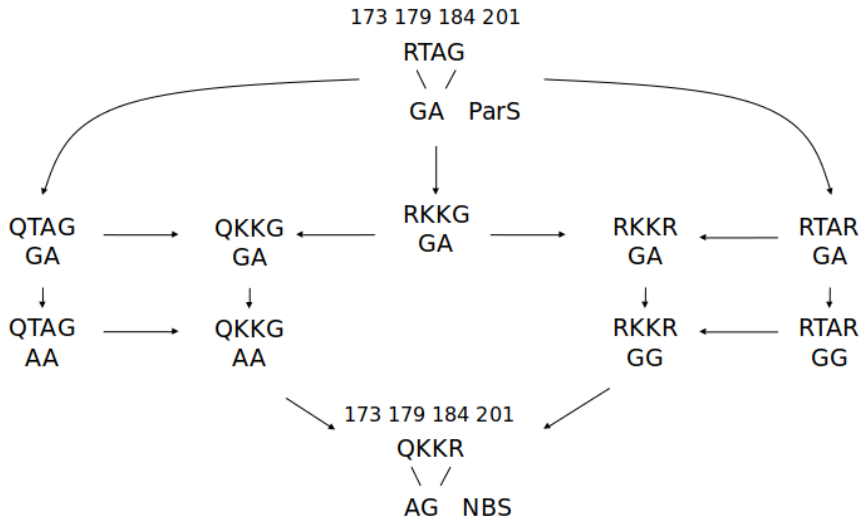


Figure 4.4: The six potential evolutionary pathways of the specificity amino acids. The top four letters refer to the four specificity amino acids, whilst the bottom two letters refer to the 2 DNA bases that mutate from *parS* to *NBS*. RTAG and QKKR for ParB and Noc refer to the key amino acids found in the provided crystal structures.

Again, a contact map was constructed using the key amino acids and base pairs, however this revealed a key issue — the QKKG mutation was interacting with the *parS* site (figure 4.5). As can be seen in figure 4.6, deep mutational scanning experiments show this mutation should only interact with the DNA once the G1 to A1 mutation had occurred — the Q173 amino acid should not be able to interact with the G1 base. As this was a potential sampling issue, as the DNA and protein had started significantly close to one another, and hence extending the simulation would potentially cause the DNA and protein to disassociate with one another. However, even by extending the simulation to 500 ns, the interaction remained indicating that this was likely a potential energy minima but not the global energy

minima. Using enhanced sampling techniques, such as replica exchange molecular dynamics (REMD), may potentially reveal the global minima such that the binding would cease, however these techniques are complicated and require a lot of computing time and so this was not performed.

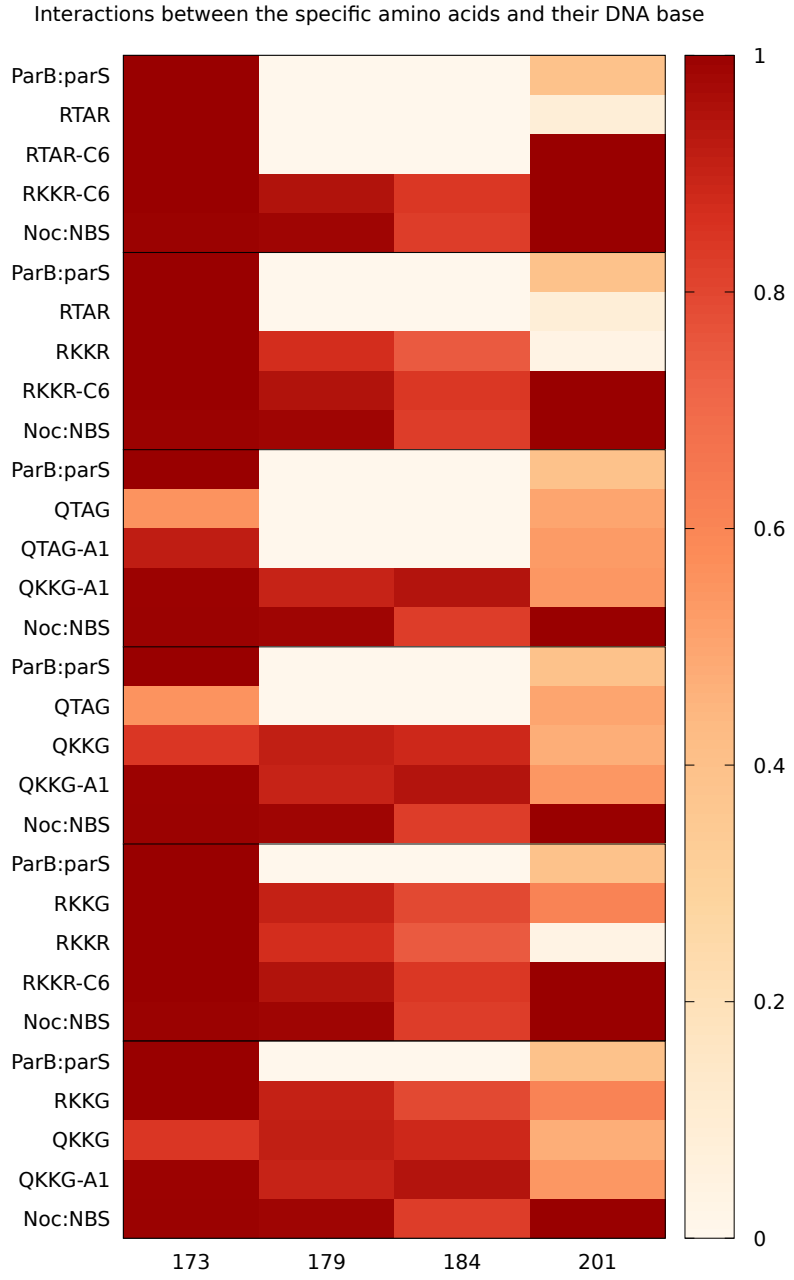


Figure 4.5: A contact map based on the average amount of time interactions occur between the four specificity amino acids and the DNA throughout the simulation of both sub-units of the dimer. Of note, the QKKG mutation was found to bind to parS and yet experimental data suggests this should not occur, with binding only being allowed with the A1 base mutation.

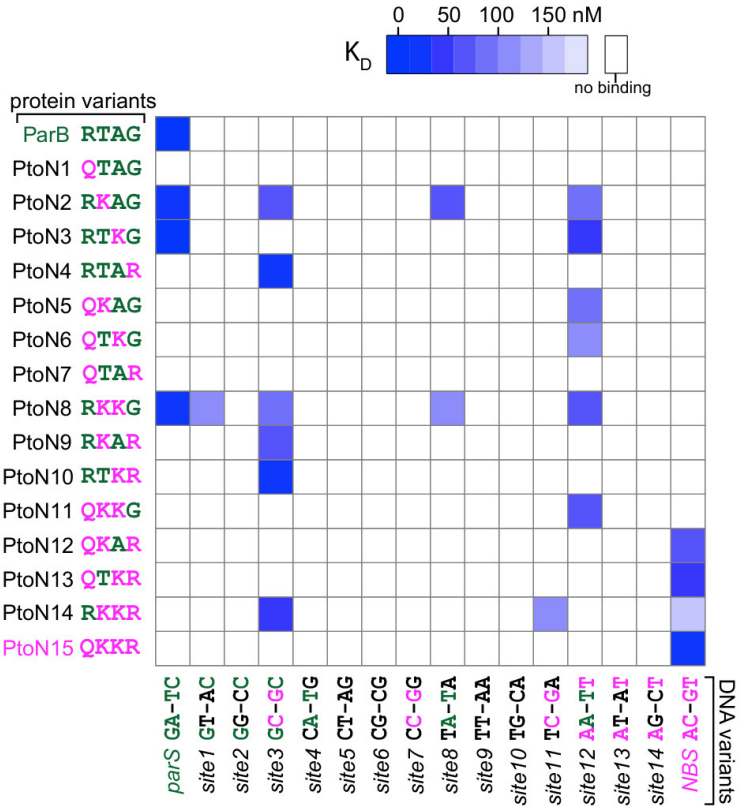


Figure 4.6: Deep mutational scanning shows that the QKKG mutation should only bind to the DNA once the A1 base mutation has occurred [73]. Protein-DNA affinity is represented by K_D, the dissociation constant. In deep mutational scanning, every relevant protein variant is synthesised and the activity of the variant is compared to the wild-type. Here, every variant of the key amino acids were synthesised, and their binding affinity to each DNA sequence is measured.

4.3 Summary

Simulations of ParB-*parS* and Noc-*NBS* complexes have been performed, revealing in atomistic detail exactly how these proteins recognise their binding sites. These complement x-ray crystallography and deep mutational scanning experiments, as the specific interactions occurring could not be captured using these techniques, and prove the usefulness of MD simulations to aid experiments in explaining a full story.

Work also began performing specific simulations to unveil how these 4 amino acids may have evolved over time. Whilst this proved unsuccessful, it does provide an outline of a method through which this could be studied in future. A good case study would be to continue along the lines of work

performed here, taking the structures and performing enhanced sampling techniques such as replica exchange MD, which would take the system to a more global minima in which the interactions would hopefully align with the experimental methods. Further analysis could then be performed on these structures, such as studying the DNA base-pair and base-step parameters for the wild-type systems and the evolving systems to show how the system is changing and validate or disprove certain pathways.

Chapter 5

DNA origami interactions with fluorescent proteins

Recent developments have shown DNA boxes as capable of trapping structures (e.g. the green fluorescent protein) [25]. Here, we create a protocol to enable the simulation of small origami structures atomistically and to answer whether this trapping is mechanical or based on nonbonded interactions between proteins and DNA. A structure of meGFP was created and parametrised to allow simulations to occur to study the mechanism of this encapsulation. This version of GFP has a mutation in residue 206 from an alanine to a lysine, this occurs at its dimerization domain, the additional positive charge stops the aggregation that would be otherwise expected from occurring [81], thus making it a popular choice for experiments. This will boost understanding of experimental data of these DNA nanostructures which cannot probe these length scales and work towards advancing the field of DNA nanotechnology. The main aims of this chapter is to create an efficient methodology for simulating DNA nanostructures atomistically, and show this effectiveness by aligning with experimental data to explain how these structures can capture proteins (in this case, the meGFP).

Others' Contributions

- caDNAno designs, oxDNA simulations and complementary experiments were performed by Matteo Marozzi

5.1 Simulating origami

In order to design DNA origami structures *in silico*, tools such as caDNAno [23] have been invented. These create an output file of the individual DNA strands in a 2D lattice, with staple strands and forces which can then be simulated using the oxDNA [168] coarse-grained simulation program which folds the structure into the final design. After a short production oxDNA phase, tacoxDNA [169] was used which converts these into rudimentary atomistic models. These atomistic models were then minimised with AMBER to allow them to be simulated. A key advantage to all-atom modelling over coarse-grained oxDNA simulation is it allows proteins to be incorporated into the system (though an anisotropic network model representation has been added [170], it does not incorporate electrostatics which are key for many of these interactions)

To test the feasibility of simulating structures to compare with experimental data, a structure of a 25 nm DNA origami box studied experimentally was generated. However, due to the resulting electrostatic energies, AMBER failed to sufficiently minimise the system (with multiple attempts at increasing number of steps using steepest descent before conjugate gradient minimisation) as the system had 9599 bases, likely due to partial overlaps from the conversion to the atomic model but also overstretched bonds contributing to the overall high energy of the system.

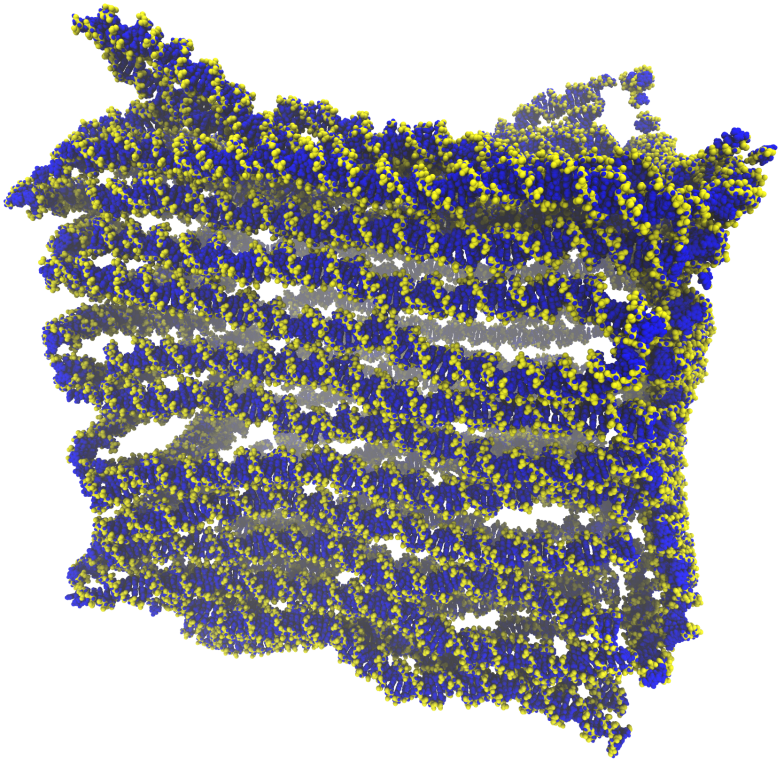


Figure 5.1: A 25 nm by 25 nm DNA origami box, composed of 9599 bases (backbone in yellow, bases in blue). This box is the same as ones designed and used in complementary experiments for the encapsulation of proteins. Attempts were made to simulate this using AMBER but the large electrostatic energies of the system rendered this impossible.

Instead, a new structure of a significantly smaller box (1123 bases) was prepared as can be seen in figure 5.2. This system was then minimised, equilibrated and simulated in explicit solvent twice, once where the system used monovalent cations (200 mM KCl) and a second where divalent cations (14 mM $MgCl_2$) ions were used so that the effect of the ionic conditions on the origami structures could be studied. 200 mM KCl was chosen to approximate the system of a nucleoid as found in nature and was standard in other systems simulated in this thesis and seen in chapter 3, whilst 14 mM $MgCl_2$ was chosen to match experimental conditions used to create the boxes. The Dang ion parameters [160] were chosen for K whilst the Li/Merz ion parameters [171] were used for $MgCl_2$. These simulations were 500 ns long, a significant enough amount of time for the systems to equilibrate into stable structures. This timeframe is sufficient to view the effect that each ion concentration has on the DNA structure, seeing whether the large amounts of DNA are compacted (akin to how divalent cations may compact DNA *in vivo*).

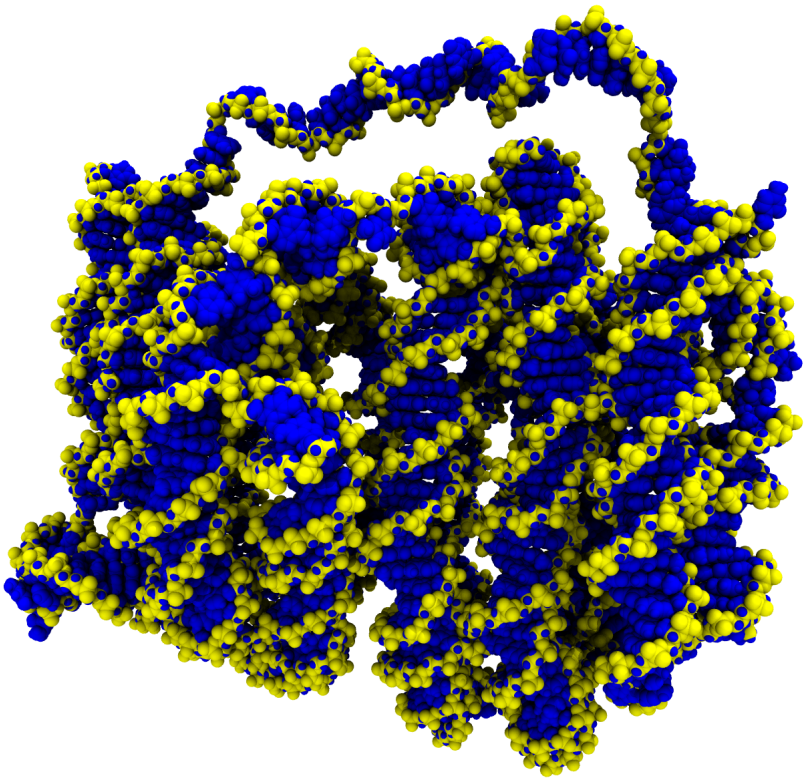
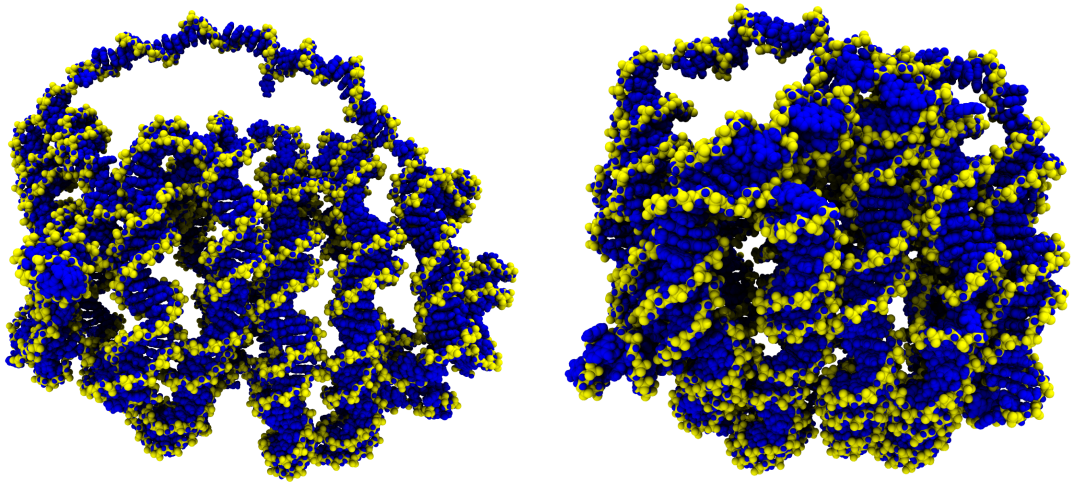


Figure 5.2: In order to be able to simulate DNA origami structures in all-atom, explicit solvent detail, the size of the structure was downscaled. A 3D box structure made up of 1123 bases was created which allowed details about pores between strands to be analysed.

5.2 Ionic effects on DNA origami structures

Superficially, visually inspecting the final state of each simulation, the $MgCl_2$ system saw a compaction in the DNA whilst the KCl did not, as can be seen in figure 5.3. This is expected as multivalent cations promote DNA condensation [172, 173], whilst monovalent cations will simply screen the electrostatic repulsion between strands, rather than create an attractive force between them [174].



(a) The box in KCl. **(b)** The box in $MgCl_2$.

Figure 5.3: The end results after 500 ns simulations show how the monovalent and divalent ions affect the structure of the DNA origami. The DNA structures were prepared twice, once where they were solvated in 200 mM KCl (a) and another where instead 14 mM $MgCl_2$ was used (b). It can be seen that the $MgCl_2$ promoted compaction between the different strands, whilst the KCl did not and so the structures ended up expanding.

To test that the system was behaving as expected, the radial distribution function (a measure of the spatial arrangement of atoms by taking the number of atoms at given distances from a reference point, normalised by the total number of atoms and the density of the system) between the ions and the DNA phosphate backbone was determined as seen in figure 5.4. The radial distribution function indicates the probability of finding an ion at a distance from a DNA phosphorus atom. The first peak indicates that both K and Mg^{2+} ions make a direct contact $\sim 4-5$Å away from the DNA, and the second peak suggests they have water molecule mediated contacts further. The first peak observed in the 200mM KCl is lower then the $MgCl_2$ indicate a lower ion condensation around the DNA with respect to bulk concentration as the DNA is saturated.

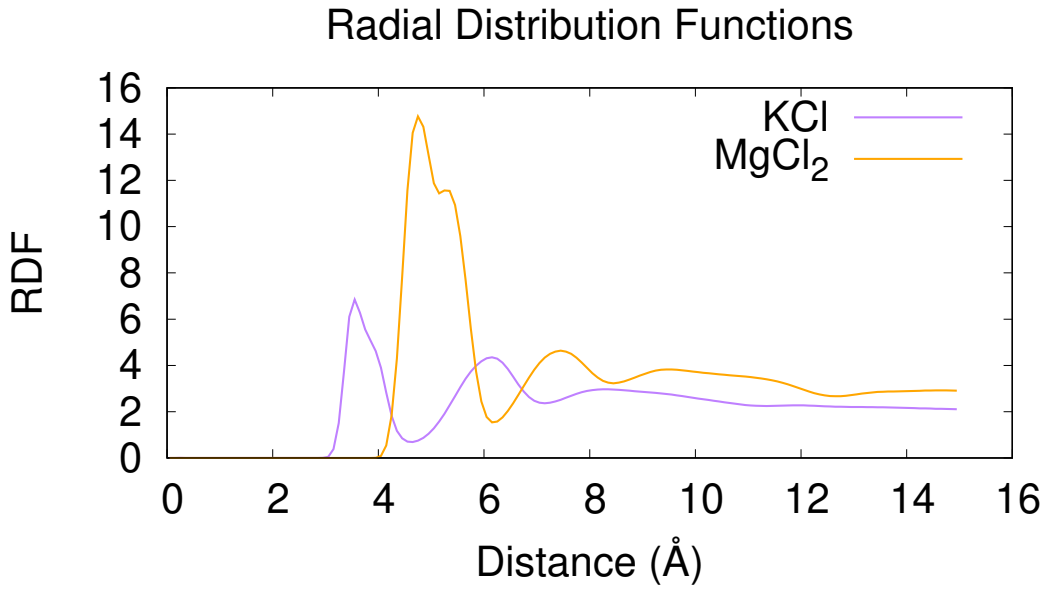
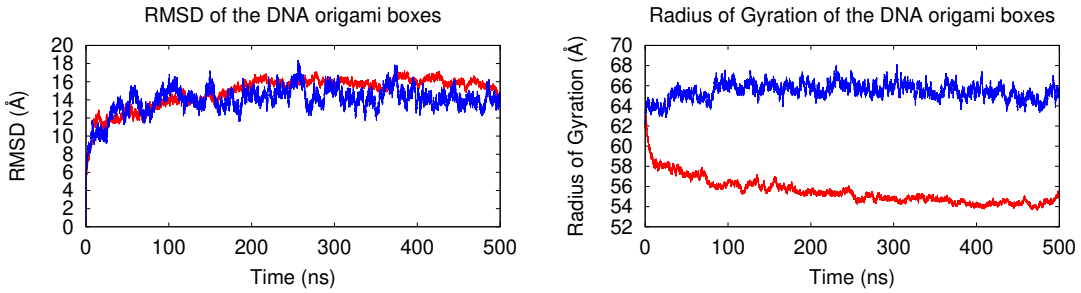


Figure 5.4: The radial distribution function of cations around the DNA phosphate backbone in 200 mM KCl (purple) and 14 mM $MgCl_2$ (orange). It can be seen that there is an initial peak where direct contact between the cations and the phosphate backbone are made, and secondary peaks at water-mediated contacts. Magnesium ions make more direct interactions with the DNA backbone (first peak) and mediated by water molecules (subsequent peaks) than potassium. This also possibly occurs as there is a higher concentration of potassium, leading to a lower ion condensation around the DNA relative to the bulk concentration due to saturation.

In order to quantify the structural changes induced, the root mean square deviation (RMSD) of each system was calculated as can be seen in figure 5.5, comparing the displacement of the atomic positions to the starting frame. It can be seen from the RMSD plots that both systems deviated from the initial structure by similar levels of deviation, but that regardless of ionic condition, there wasn't a notable change in how much the systems changed. However, for the DNA box the $MgCl_2$ system had fewer fluctuations, being less flexible, likely as the box had compacted and remained in this condensed form, whilst the KCl system showed a greater fluctuation. This can be seen by taking the average and standard deviation of each of the last 200 ns (taking the first 300 to be a converging stage of the simulation). The $MgCl_2$ system has an average and standard deviation of $15.882\mathring{A} \pm 0.497$ whilst for KCl $14.25\mathring{A} \pm 0.919$. The standard deviation of the KCl system was approximately double that of the $MgCl_2$, as the KCl is just acting to screen electrostatic charges in the DNA, the general fluctuations

will still occur.

A second method to measure structural changes is to calculate the radius of gyration, which is a measurement for the compactness of a system. These can also be seen in figure 5.5 and show a much clearer picture of the effects the different salts have on the DNA. In the KCl system, the box opens up immediately, whilst the $MgCl_2$ converges on a compacted state with a radius of gyration. By taking the last 200 ns (assuming the first 300 to be a converging stage), the KCl has an average radius of gyration of $65.338\mathring{A} \pm 0.714$ whilst the $MgCl_2$ has $54.476\mathring{A} \pm 0.319$.



(a) RMSD of the boxes.

(b) Radius of gyration of the boxes.

Figure 5.5: The RMSD and radius of gyration of the origami structures in either 14 mM $MgCl_2$ (red) or 200 mM KCl (blue). Whilst the RMSD doesn't show a particular change in either ionic condition (a), it can be seen that there is a decrease in the radius of gyration in $MgCl_2$ when compared to KCl, indicative of the condensation cause by the divalent ions on the DNA (b).

5.3 Preparation of fluorescent proteins

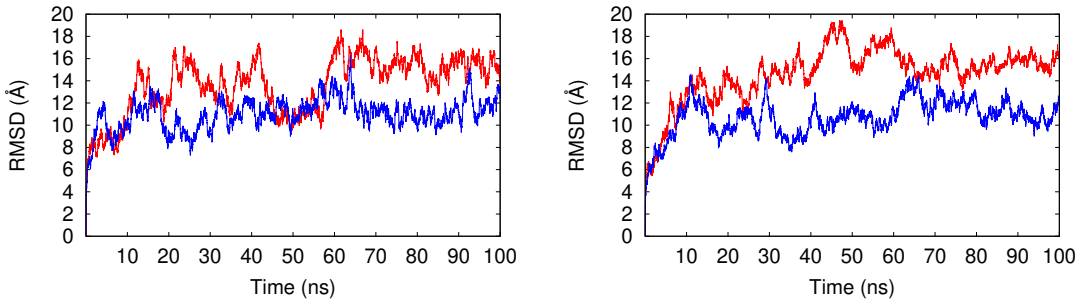
Fluorescent proteins contain a non-standard residue (the chromophore) — that is, in this case, a cyclization of three amino acids that is not represented in force fields. Thus, force field parameters must be generated for the residue before simulation can be performed. Fortunately, the chromophore has been resolved via x-ray crystallography, meaning the generation of atom coordinates is unnecessary, making this task significantly easier. Using the 2Y0G crystal structure of eGFP, containing the chromophore, the Antechamber tool [175] from the AMBER software suite was used to determine charge distributions and atom types within the chromophore residue, notably partial charges were derived using the AM1-BCC charge scheme [176]. Once these parameters were generated, the final meGFP structure was created via *in silico* mutagenesis, turning the 206 residue from alanine to lysine. This structure was then minimised, equilibrated

and a production run of 10 ns was run. A structure was then created using the average atomic positions of the last 1 ns.

Whilst simulation of any mechanical encapsulation of the GFP by a DNA box remains out of reach, the dynamics of a GFP molecule near the wall of the box was still of interest, as it would provide information into whether the protein would interact with a side of the box and whether it could potentially escape through a pore. To this end, a flat sheet of DNA was prepared in the same manner as the box before, and then four structures were created, two in which the sheet was alone in either a 200 mM KCl or 14 mM $MgCl_2$ solution, and two where the protein was placed $\sim$ 15Å away from the DNA in these solutions. These systems were then simulated for 100 ns.

5.4 Simulations of DNA origami alongside fluorescent proteins

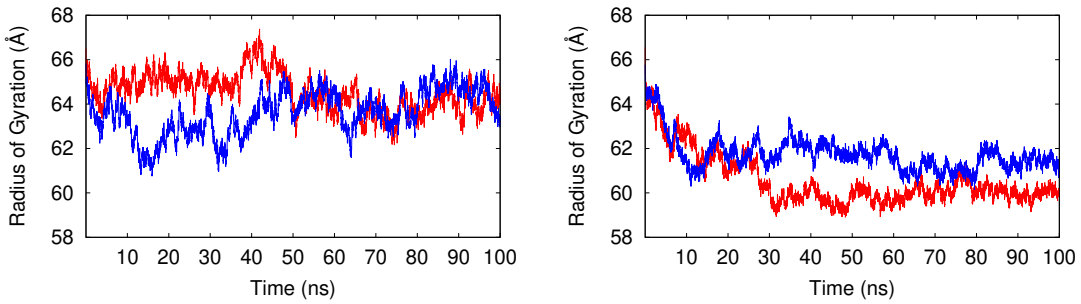
The RMSD of each system was calculated with and without the protein as seen in figure 5.6. Here the difference in the ionic condition alone on the DNA is minimal, this is possibly due to the square having fewer degrees of freedom when compared to the box so the changes possible has a smaller effect on overall structure. The addition of the meGFP does increase the RMSD in both systems, but the effect is more notable in the $MgCl_2$, a result of the protein pulling the DNA into a more compacted structure relative to the system with no protein. These suggest that the systems had converged sufficiently within the 100 ns of production.



(a) GFP-DNA square in KCl. **(b)** GFP-DNA square in $MgCl_2$.

Figure 5.6: Comparison of the RMSD of the DNA origami squares in each ionic condition with and without the meGFP (with the protein is red, without protein is blue). Here the effect of the protein in each system is clear, with it resulting in a higher RMSD in both.

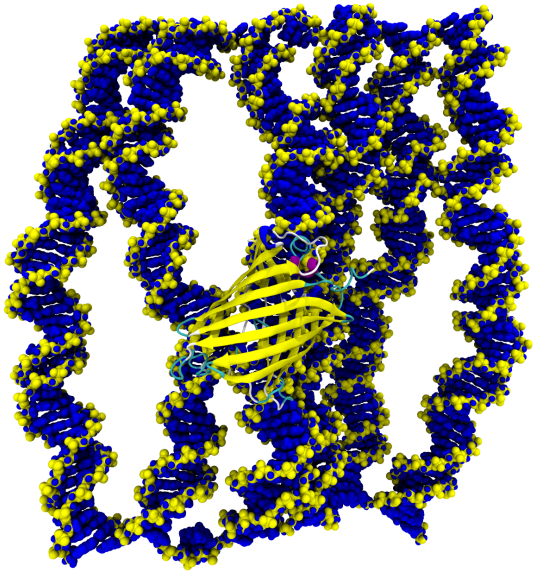
As before, the sheets saw little compaction in the KCl system but compacted in the $MgCl_2$. In order to quantify the condensing of the DNA, the radius of gyration was again found for all systems, as can be seen in figure 5.7. For the systems in the KCl solution, the radius of gyration remained fairly constant throughout each simulation, with the protein not having a noticeable impact in the long term compaction of the DNA. Conversely, as with the DNA box, the $MgCl_2$ buffer promoted the compaction of the square both with and without the protein. Here, the protein did have a noticeable impact, with the DNA having a radius of gyration which was lower by about 2Å when compared to the system without.



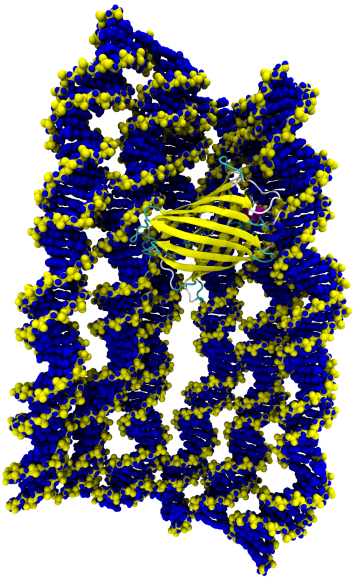
(a) GFP-DNA square in KCl. **(b)** GFP-DNA square in $MgCl_2$.

Figure 5.7: Comparison of the radius of gyration of the DNA origami squares in each ionic condition with and without the meGFP (with the protein is red, without protein is blue). Whilst the difference is minimal in the KCl systems regardless of the meGFP, the $MgCl_2$ system shows the compaction occurring and how the meGFP does further induce the compaction.

To understand the compaction that is found in the $MgCl_2$ systems, the simulations were viewed. It was seen that the condensation of the DNA by the Mg^{2+} caused the sheet to start to hinge in the middle and act to bring opposing sides closer as seen in figure 5.8d. This effect was further promoted by the meGFP, which bound on both ends of its β-barrel structure, thus causing the DNA to bend further inwards. Conversely, the KCl system did not have a similar hinge to compact the DNA, though there an interaction between the GFP and the DNA. There was an initial binding between the protein and the DNA, which quickly broke away. Then the protein jumped away and rebound to a new spot on the DNA where it stayed, with figure 5.9 quantifying this via the number of interactions formed between the meGFP and DNA over time.

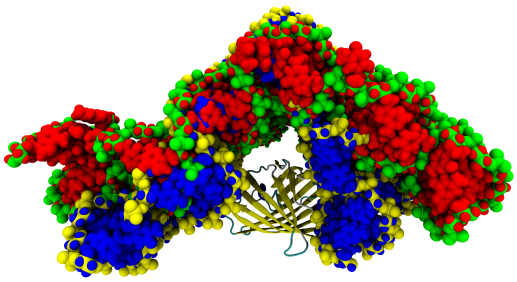


(a) GFP-DNA square in KCl.



(b) GFP-DNA square in $MgCl_2$.

(c) Side on view in KCl.



(d) Side on view in $MgCl_2$.

Figure 5.8: meGFP is seen to interact with DNA in either KCl and $MgCl_2$ buffers. (a) shows how the meGFP does bind to DNA in KCl, though it should be noted that simulations show an initial interaction before moving to a second spot (the second interaction is shown here). (b) shows how the protein strongly binds to the DNA in $MgCl_2$. In this system, the protein binds and stays near the DNA immediately. (c) and (d) show side on views, comparing the simulations with meGFP (DNA in blue and yellow) vs without meGFP (DNA in red and green). The meGFP can be seen to actively promote compaction of the DNA origami in the $MgCl_2$ buffer.

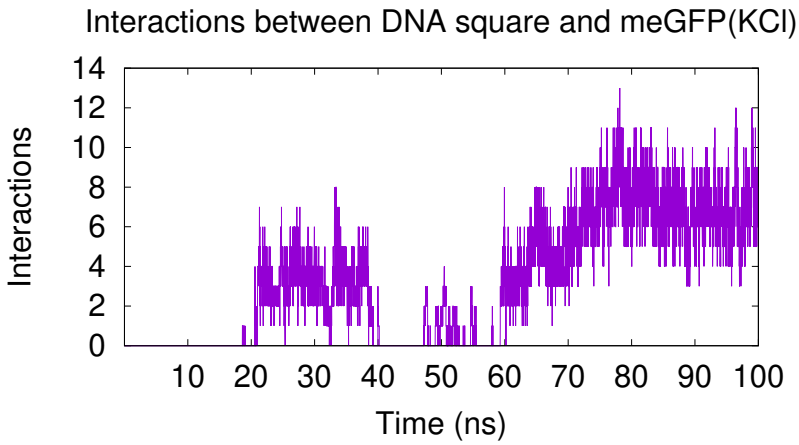


Figure 5.9: A plot showing the number of interactions between the DNA and protein in KCl over time. The period of $\sim 40-50$ ns shows no interactions, which is where the protein quickly moves from one section of the DNA to another, which allows more interactions to occur for longer.

A key interaction of interest was the 206 residue of the protein, as this contains a mutation from an alanine to a lysine to prevent the aggregation of these proteins. Hence, there is the introduction of an extra positive charge which may enhance the DNA-binding capabilities of the meGFP variant compared to regular eGFP. Figure 5.10 shows the lysine in the initial 20 ns of the simulation in the $MgCl_2$ buffer. It was seen that the lysine is attracted to the DNA at the start of the simulation, aiding in pulling the meGFP towards the DNA almost immediately.

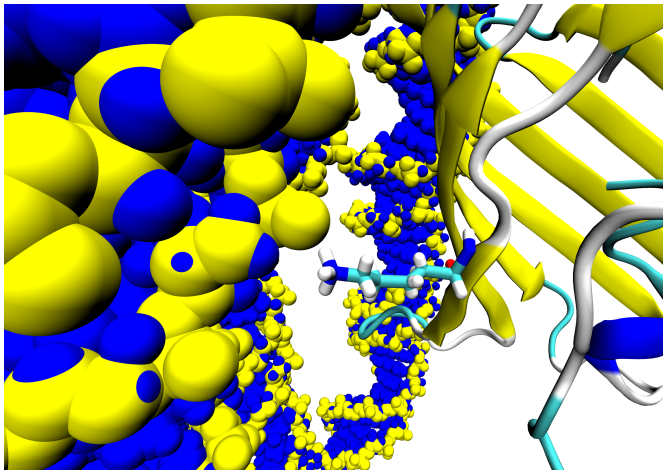


Figure 5.10: Initial interaction of 206 lysine to the DNA. Within the first 20 ns, the meGFP drifts towards the DNA and the lysine is seen attracting towards the DNA quite strongly in the $MgCl_2$ buffer.

From this point, the DNA naturally begins to curve inwards, towards the ends of the meGFP. A repositioning occurs, whereby the ends of the β-barrels interact with the DNA. This occurs as each side of the β-barrel has multiple arginine (arg73, arg168) and lysine (lys43, lys79, lys101i, lys140) which form salt bridges with the DNA backbone. Due to this, the middle of the meGFP stops forming direct interactions with the DNA as can be seen in figure 5.11, though the lysine remains attracted and extending towards the square.

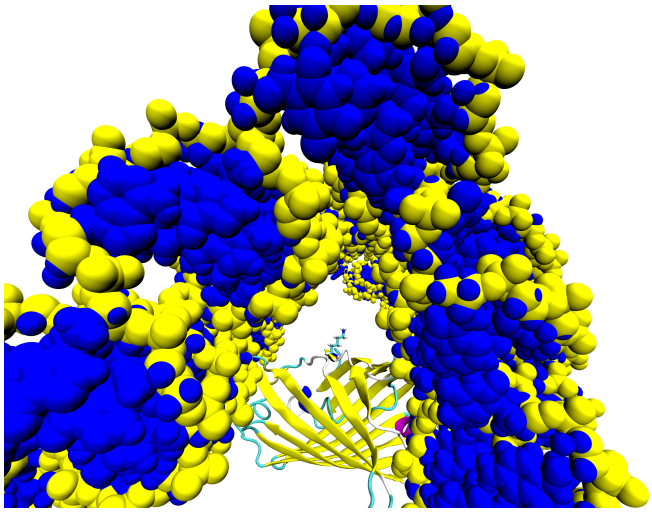


Figure 5.11: At the 100 ns mark, the DNA has closed in around the meGFP. Due to the manner in which this occurs, there is no direct interaction between the lysine and the DNA, though the lysine can be seen to still be attracted towards the DNA.

In the KCl solution, the interactions between meGFP and DNA were weaker but still occur. Initially there was an interaction that started at the 35 ns mark and ended after ~ 15 ns, as can be seen in figure 5.12. This interaction did not involve the lysine, and the meGFP drifted away from the DNA.

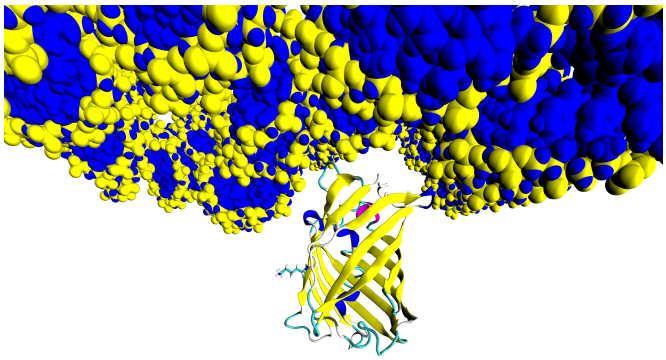


Figure 5.12: The initial interaction between the meGFP and the DNA square in KCl. Here one side of the β-barrel does interact with the DNA (via the arg73 as seen in the $MgCl_2$ buffer), and the lysine can be seen to not be in any interaction.

The protein did then land again on the DNA, this time in an interaction which involved the lysine, as seen in figure 5.13. This was a much more stable interaction, with the protein remaining in place for the remaining 40 ns.

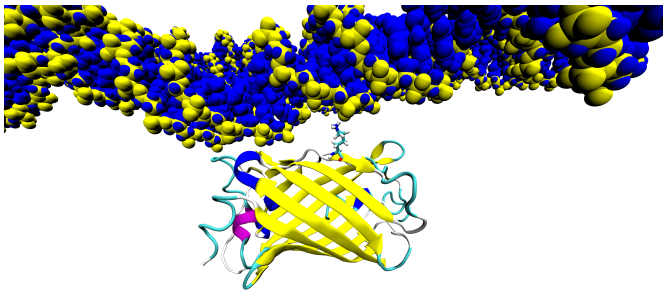
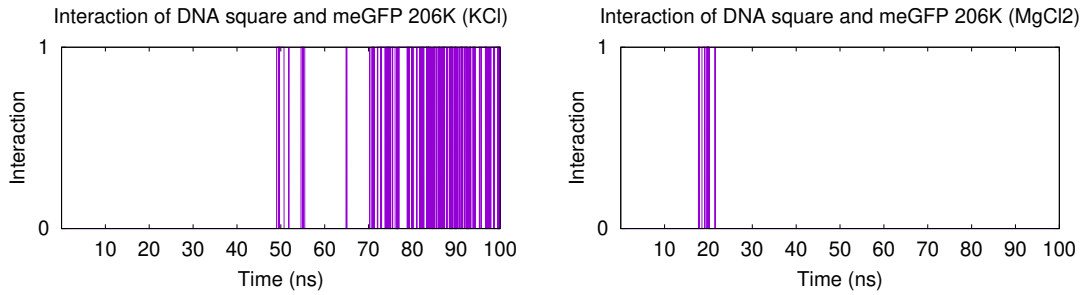


Figure 5.13: The end interaction of the meGFP and DNA square in KCl. Here, the lysine can be seen to be in a stable interaction with the phosphate backbone.

In order to quantify the differing interactions occurring in the KCl and the $MgCl_2$ solutions, the interactions between the lysine and DNA were measured. Figure 5.14a shows that for the second binding of the protein and DNA in KCl, the lysine is in constant interaction with the DNA, and has none during the initial contact (from ~ 35 to ~ 50 ns). Meanwhile figure 5.14b shows the initial search of the DNA by the lysine at ~ 20 ns whilst showing the lysine has no direct contact when the DNA wraps around the β-barrel.



(a) KCl. **(b)** $MgCl_2$.

Figure 5.14: Comparison of the analysis for interactions between the 206K residue and the DNA in KCl (a) and $MgCl_2$ (b). The second binding of the meGFP can be seen with the 206K helping stabilise it in the KCl. In the $MgCl_2$, the initial attraction between the DNA and lysine can be seen, before the bond breaks due to the wrapping of the square around the β-barrel itself.

5.5 Summary

Whilst the simulation of larger DNA origami structures proved unattainable with current software, atomistic molecular dynamics has been shown to be capable of simulating smaller structures. This is a powerful tool for understanding how these structures can interact with proteins, which coarse-grained models such as oxDNA cannot study.

The modelling performed has shown the role the ionic conditions has on the DNA origami conformation, with divalent cations inducing a compaction in the DNA. In the $MgCl_2$ solution, the DNA origami has less flexibility than in the KCl solution which can be seen in the RMSD calculations.

It has been shown that meGFP is capable of interacting with DNA, with the A206K mutation from eGFP helping facilitate this. This is particularly interesting as the protein is often used in *in vivo* experiments of the nucleoid, so noting the interaction between the DNA (which is highly compacted in the nucleoid) and the protein here suggests the data from these experiments should be treated with care (for example, tracking diffusion *in vivo* using this protein, the diffusion will be slower than if the protein being tracked was untagged, as the meGFP will non-specifically interact with the DNA). However, it should be noted that the β-barrel has a number of lysine and arginine surrounding it. Thus, proper validation of the role of the mutation would require simulation of non-mutated eGFP alongside to compare. Preliminary experimental work by Matteo Marozzi using TIRF

microscopy suggests that meGFP colocalises with both 2D and 3D DNA origami structures and has a slower diffuses rate than when alone, further suggesting that there is a non-specific interaction occurring.

Chapter 6

Discussion

This work builds on recent developments in the field of atomistic modelling of DNA-protein dynamics. Of note, all-atom simulations showing the diffusion mechanism of a protein along and between DNA strands is relatively novel (although coarse-grained and lattice model approaches have been used previously [177, 178]). The combination of simulation and experiment presented in this thesis further builds upon previous work in the field of DNA-protein interactions, with in-liquid AFM being used in place of in-air to better represent relevant biological conditions.

This work also provides early work into studying how specific interactions between DNA and proteins may have evolved. Making single amino acid or nucleotide changes in a complex to view how these interactions can be made or broken to view likely evolutionary pathways. Whilst the work presented here is incomplete, it lays the groundwork for a methodology that would be more fruitful, driving the field forward towards understanding the evolution to develop new regulatory functions in proteins.

Work developed here has also shown the development of a protocol to study DNA origami structures in all-atom simulation. Whilst this isn't strictly new territory in the field of DNA nanotechnology, previous work has relied on either having a flat sheet to be folded in the all-atom MD [179], or reliant on using elastic network restraints to fold into place [180]. Instead, the oxDNA tool was used here which allows for much more efficient initial folding of the DNA origami structure before conversion into an all-atom model for simulation in PDB format. The created model has clearly shown the effect different ionic conditions can have on DNA, showing the divalent cations such as Mg^{2+} caused a noticeable compaction in the DNA relative to the monovalent K. In particular, these systems have been used to show that meGFP is capable of interacting with DNA. This

has potentially far reaching consequences as it is often used in *in vivo* fluorescence experiments studying the nucleoid, under the assumption it has a minimal effect beyond allowing the targeted protein to be tracked.

6.1 Future Work

The work on HU focused on the specific site of damage found in the 1P78 PDB, other forms of damage such as nicks in the DNA would be of immediate interest due to their relevance *in vivo*. At current, DNA force fields tend to overemphasize the base stacking effect, such that nicks in the DNA tend to have less of an effect than would be found in real life. Studying the DNA with the damage in different structural positions would also be of interest, changing the groove positions relative to the HU would likely change the energetic landscape calculated in this work. Studying the effect supercoiling has on the DNA-HU complex would be of interest, as DNA *in vivo* tends to be negatively supercoiled and this work has been conducted for IHF. A less involved study would be whether IHF exhibits a similar bridging mode to that found for HU here, as the similar surface electrostatic profiles suggest this may be possible. Lastly, a similar study in the proline intercalation of IHF, and how the sequence affects this, would be interesting to note the change in DNA-binding capability of the protein.

As stated before, an immediate line of study to further understanding of the evolution of ParB to Noc would be to perform replica exchange MD, as this would allow the ground state of each mutation to be found which may show whether the interactions were possible or not. A similar study of other related proteins (such as HU and IHF) would also be of interest to study why such similar proteins exist and to further elucidate their differences.

As computing power continues to increase, the size of systems that can be studied will also increase. Thus, studying more complex DNA origami will continue to become possible. At present, simply trying the large DNA box in other molecular dynamics suites would be of interest, as both GROMACS and NAMD have been used to study much larger systems than used in this work. This would potentially allow a more in-depth study in the capability of a DNA box in encapsulating GFP than was shown here. An immediate note is the pores in the large box were much larger than those in the square, the possibility of the GFP simply escaping through these may exist but cannot be sampled at current. As simulation and experiment continue to converge, relevant single-molecule studies may be developed to

allow the GFP-DNA interaction suggested here to be visualised.

The work presented here has shown that experiment and simulation go can hand in hand to explain dynamics in biological systems. In principle, any interacting system that is sufficiently small can take advantage of the methodology used, immediately obvious suggestions would be other nucleoid-associated proteins such as Fis or H-NS, though the potential systems that can be studied are endless. Computational power is increasing and experiments to study smaller systems are being developed, so perhaps soon a true convergence between simulation and experiment will be fully possible.

Appendix A

DNA Sequences

A.1 DNA sequences in the 305 bp damaged DNA

The DNA sequence containing the damaged section. The binding site containing the damage is in bold.

GAAGTGTCGCTACGGTCTCAGACATCACAGTCTACTACTGGCATG
AGTTTGTAGCCCAACTCATAATAGAATGACAAAGAAATGTATTTGTAA
CGACTATGGCAAATCGACTTTGCTGTATGTAACGTTCCTCAAATATTT
ACTC**CATATCAATTTGTTG**CTCATTTATAAACTCCTTGCAATGTATGT
CGTTTCAGCTAAACGGTATCAGCAATGTTTATGTAAAGAAACAGTAA
GATAATACTCAACCCGATGTTTGAGTACGGTCATCATCTGACACTAC
AGACTCTGGCATCGCTGTGAAG

A.2 DNA sequences in the 303 bp B-DNA

The DNA sequence containing the corrected section. The corrected site is in bold.

GAAGTGTCGCTACGGTCTCAGACATCACAGTCTACTACTGGCATG
AGTTTGTAGCCCAACTCATAATAGAATGACAAAGAAATGTATTTGTAA
CGACTATGGCAAATCGACTTTGCTGTATGTAACGTTCCTCAAATATTT
ACTC**CAACAAATTGTTG**CTCATTTATAAACTCCTTGCAATGTATGTC
GTTTCAGCTAAACGGTATCAGCAATGTTTATGTAAAGAAACAGTAAG
ATAATACTCAACCCGATGTTTGAGTACGGTCATCATCTGACACTACA
GACTCTGGCATCGCTGTGAAG

A.3 DNA oligonucleotides

The DNA oligonucleotides used to create the DNA constructs used in AFM imaging

	Sequence (all 5'-3')
Sense 1	GAAGTGTCGCTACGGTCTCAGACATCACAGTC TACTACTGGCATGAGTTTGTAGCCCAACTCATA ATAGAATGACAAAGAAATGTATTTGTAAC
BDNA Sense 2	GACTATGGCAAATCGACTTTGCTGTATGTAAC GTTCCTCAAATATTTACTCCAACAAATTGTTGC TCATTTATAAACTCCTTGCAATGTATGTC
DDNA Sense 2	GACTATGGCAAATCGACTTTGCTGTATGTAAC GTTCCTCAAATATTTACTCCATATCAATTTGTT GCTCATTTATAAACTCCTTGCAATGTATGTC
Sense 3	GTTTCAGCTAAACGGTATCAGCAATGTTTATGT AAAGAAACAGTAAGATAATACTCAACCCGATG TTTGAGTAC
Sense 4	GGTCATCATCTGACACTACAGACTCTGGCATC GCTGTGAAG
Antisense 4	CTTCACAGCGATGCCAGAGTCTGTAGTGTCAG ATGATGACCGTACTCAAACATCGGGTTGAGTA TTATCTTACTGTTTCTTTACATAAACATTG
DDNA Antisense 3	CTGATACCGTTTAGCTGAAACGACATACATTGC AAGGAGTTTATAAATGAGCATATCAATTTGTTG GAGTAAATATTTGAG
BDNA Antisense 3	CTGATACCGTTTAGCTGAAACGACATACATTGC AAGGAGTTTATAAATGAGCAACAATTTGTTGG AGTAAATATTTGAG
Antisense 2	GAACGTTACATACAGCAAAGTCGATTTGCCAT AGTCGTTACAAATACATTTCTTTGTCATTCTAT TATGAGTTGG
Antisense 1	GCTACAAACTCATGCCAGTAGTAGACTGTGAT GTCTGAGACCGTAGCGACACTTC

Table A.1: DNA oligonucleotides used to create the DNA constructs used in AFM imaging.

Glossary

A adenine

AFM atomic force microscpy

AMBER Assisted Model Building with Energy Refinement

bp base pair

C cytosine

CIP contact ion pair

CFA Coloumb Field Approximation

CTD C-terminal domain

DNA deoxyribonucleic acid

eYFP enhance Yellow Fluorescent Protein

G guanine

GB-HCT Generalised Born (Hawkins, Cramer, Truhlar)

GB-OBC Generalised Born (Onufriev, Bashford, Case)

GFP Green Fluorescent PRotein

HTH helix-turn-helix

HU histone-like protein from *E. coli* strain U93 *or* heat-unstable protein

IHF integration host factor

NAB nucleic acid builder

NAP nucleoid-associated protein

NBS Noc binding site

NMR nuclear magnetic resonance

NTD N-terminal domain

MD molecular dynamics

meGFP monomeric enhanced Green Fluorescent Protein

PDB Protein Data Bank

PLL poly-l-lysine

PMF potential of mean force

QPD quadrant-photodiode

RMSD root-mean-square deviation

RNA ribonucleic acid

SMC Structural Maintence of Chromosomes

T thymine

TIP3P transferable intermolecular potential with 3 points

TIRF total internal reflection fluorescence

U uracil

WHAM weighted histogram analysis method

Bibliography

[1] Crick FHC, 1958. "On Protein Synthesis". *Symp. Soc. Exp. Biol* **12** 138–163

[2] Raiber EA, Hardisty R, van Delft P, and Balasubramanian S, 2017. "Mapping and elucidating the function of modified bases in DNA". *Nature Reviews Chemistry* **1** s41570–017

[3] Yakovchuk P, Protozanova E, and Frank-Kamenetskii MD, 2006. "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix". *Nucleic Acids Research* **34** 564–74

[4] Dickerson R, 1989. "Definitions and nomenclature of nucleic acid structure components". *Nucleic Acids Research* **17**(5) 1797–1803. ISSN 0305-1048

[5] Lu X and Olson WK, 2003. "3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures". *Nucleic Acids Research* **31**(17) 5108–5121

[6] Dickerson RE, 1992. "DNA structure from A to Z" **211** 67–111. ISSN 0076-6879

[7] Rich A and Zhang S, 2003. "Z-DNA: the long road to biological function". *Nature Reviews Genetics* **4** 566–572

[8] Wing RA, Drew HR, Takano T, Broka CA, Tanaka S, Itakura K, and Dickerson RE, 1980. "Crystal structure analysis of a complete turn of B-DNA". *Nature* **287** 755–758

[9] Smith NC and Matthews JM, 2016. "Mechanisms of DNA-binding specificity and functional gene regulation by transcription factors". *Current Opinion in Structural Biology* **38** 68–74

[10] Turner PR and Denny WA, 1996. "The mutagenic properties of DNA minor-groove binding ligands". *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **355**(1) 141–169

[11] Timoféeff-Ressovsky NW, Zimmer KG, and Delbrück M, 1935. "Über die Natur der Genmutation und der Genstruktur". *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Fachgruppe VI* **23** 189–245

[12] Walker GC, Siede W, Wood RD, Schultz RA, and Ellenberger T, 2005. *DNA Repair and Mutagenesis*. ISBN 9781555813192

[13] Balasubramanian B, Pogozelski WK, and Tullius TD, 1998. "DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone". *Proceedings of the National Academy of Sciences* **95**(17) 9738–9743

[14] Lindahl T, 1993. "Instability and decay of the primary structure of DNA". *Nature* **362** 709–715

[15] Reisz JA, Bansal N, Qian J, Zhao W, and Furdui CM, 2014. "Effects of ionizing radiation on biological molecules–mechanisms of damage and emerging methods of detection." *Antioxidants & redox signaling* **21 2** 260–92

[16] Sutherland BM, Bennett PV, Sidorkina O, and Laval J, 2000. "Clustered DNA damages induced in isolated DNA and in human cells by low doses of ionizing radiation". *Proceedings of the National Academy of Sciences* **97**(1) 103–108

[17] Seeman N and Sleiman H, 2018. "DNA nanotechnology". *Nature Reviews Materials* **3** 17068

[18] Xu F, Xia Q, and Wang P, 2020. "Rationally Designed DNA Nanostructures for Drug Delivery". *Frontiers in Chemistry* **8**

[19] Seeman NC, 1982. "Nucleic acid junctions and lattices". *Journal of Theoretical Biology* **99**(2) 237–247

[20] Kallenbach NR, Ma RI, and Seeman NC, 1983. "An immobile nucleic acid junction constructed from oligonucleotides". *Nature* **305** 829–831

[21] Chen J and Seeman NC, 1991. "Synthesis from DNA of a molecule with the connectivity of a cube". *Nature* **350** 631–633

[22] Rothemund PWK, 2006. "Folding DNA to create nanoscale shapes and patterns". *Nature* **440** 297–302

[23] Douglas SM, Marblestone AH, Teerapittayanon S, Vazquez A, Church GM, and Shih WM, 2009. "Rapid prototyping of 3D DNA-origami shapes with caDNAno". *Nucleic Acids Research* **37** 5001 – 5006

[24] Tikhomirov G, Petersen P, and Qian L, 2017. "Fractal assembly of micrometre-scale DNA origami arrays with arbitrary patterns". *Nature* **552** 67–71

[25] Burns JR, Lamarre B, Pyne ALB, Noble JE, and Ryadnov MG, 2018. "DNA Origami Inside-Out Viruses". *ACS Synthetic Biology* **7**(3) 767–773

[26] Pumm AK *et al.*, 2022. "A DNA origami rotary ratchet motor". *Nature* **607** 492 – 498

[27] Blattner FR *et al.*, 1997. "The Complete Genome Sequence of Escherichia coli K-12". *Science* **277** 1453–62

[28] Thanbichler M, Wang SC, and Shapiro L, 2005. "The bacterial nucleoid: A highly organized and dynamic structure". *Journal of Cellular Biochemistry* **96**(3) 506–521

[29] Lebowitz J, 1990. "Through the looking glass: the discovery of supercoiled DNA". *Trends in Biochemical Sciences* **15**(5) 202–207. ISSN 0968-0004

[30] Verma SC, Qian Z, and Adhya SL, 2019. "Architecture of the Escherichia coli nucleoid". *PLoS Genetics* **15**

[31] Dame RT, Rashid FZM, and Grainger DC, 2019. "Chromosome organization in bacteria: mechanistic insights into genome structure and function". *Nature Reviews Genetics* **21** 227 – 242

[32] Dame RT, Kalmykowa OJ, and Grainger DC, 2011. "Chromosomal Macrodomains and Associated Proteins: Implications for DNA Organization and Replication in Gram Negative Bacteria". *PLoS genetics* **7** e1002123

[33] Yang CC and Nash HA, 1989. "Targeting a bacterial DNABII protein with a chimeric peptide immunogen or humanised monoclonal antibody to prevent or treat recalcitrant biofilm-mediated infections". *Cell* **57** 869–880

[34] Swinger KK and Rice PA, 2004. "IHF and HU: flexible architects of bent DNA". *Current opinion in structural biology* **14** 28–35

[35] Yoshua SB, Watson GD, Howard JAL, Velasco-Berrelleza V, Leake MC, and Noy A, 2021. "Integration host factor bends and bridges DNA in a multiplicity of binding modes with varying specificity". *Nucleic Acids Research* pages 8684–8698

[36] Rice PA, wei Yang S, Mizuuchi K, and Nash HA, 1996. "Crystal Structure of an IHF-DNA Complex: A Protein-Induced DNA U-Turn". *Cell* **87** 1295–1306

[37] Hwang DS and Kornberg A, 1992. "Opening of the replication origin of Escherichia coli by DnaA protein with protein HU or IHF". *Journal of Biological Chemistry* **267** 23083–23086

[38] Kobryn K, Lavoie BD, and Chaconas G, 1999. "Supercoiling-dependent site-specific binding of HU to naked Mu DNA." *Journal of molecular biology* **289 4** 777–84

[39] Arfin SM, Long AD, Ito ET, Tolleri L, Riehle MM, Paegle ES, and Hatfield GW, 2000. "Global gene expression profiling in Escherichia coli K12. The effects of integration host factor". *The Journal of biological chemistry* **275** 29672–84

[40] Jamal M, Ahmad W, Andleeb S, Jalil F, Imran M, Nawaz MA, Hussain T, Ali M, Rafiq M, and Kamil MA, 2018. "Bacterial biofilm and associated infections." *Journal of the Chinese Medical Association : JCMA* **81 1** 7–11

[41] Gustave JE, Jurcisek JA, Goodman SD, and O BL, 2013. "Targeting bacterial integration host factor to disrupt biofilms associated with cystic fibrosis". *Journal of Cystic Fibrosis* **12** 384–389

[42] Devaraj A, Justice SS, Bakaletz LO, and Goodman SD, 2015. "DNABII proteins play a central role in UPEC biofilm structure". *Molecular Microbiology* **96** 1119–35

[43] Devaraj A, Buzzo J, Rocco CJ, Bakaletz LO, and Goodman SD, 2018. "The DNABII family of proteins is comprised of the only nucleoid associated proteins required for nontypeable Haemophilus influenzae biofilm structure". *MicrobiologyOpen* **7**(3) e00563

[44] Novotny LA, Jurcisek JA, Goodman SD, and Bakaletz LO, 2016. "Monoclonal antibodies against DNA-binding tips of DNABII proteins disrupt biofilms in vitro and induce bacterial clearance in vivo". *EBioMedicine* **10** 33–44

[45] Novotny LA, Goodman SD, and Bakaletz LO, 2020. "Targeting a bacterial DNABII protein with a chimeric peptide immunogen or humanised monoclonal antibody to prevent or treat recalcitrant biofilm-mediated infections". *EBioMedicine* **59** 102867

[46] Velmurugu Y, Vivas P, Connolly M, Kuznetsov SV, Rice PA, and Ansari A, 2017. "Two-step interrogation then recognition of DNA binding site by Integration Host Factor: an architectural DNA-bending protein". *Nucleic Acids Research* **46**(4) 1741–1755

[47] Khrapunov S, Brenowitz M, Rice PA, and Catalano CE, 2006. "Binding then bending: A mechanism for wrapping DNA". *Proceedings of the National Academy of Sciences* **103**(51) 19217–19218

[48] Lin J, Chen H, Dröge P, and Yan J, 2012. "Physical Organization of DNA by Multiple Non-Specific DNA-Binding Modes of Integration Host Factor (IHF)". *PLoS ONE* **7**

[49] Fosado YAG, Howard JAL, Weir S, Noy A, Leake MC, and Michieletto D, 2022. "Fluidification of Entanglements by a DNA Bending Protein." *Physical review letters* **130 5** 058203

[50] Thakur B, Arora K, Gupta A, and Guptasarma P, 2021. "The DNA-binding protein HU is a molecular glue that attaches bacteria to extracellular DNA in biofilms". *Journal of Biological Chemistry* **296** 100532

[51] Wojtuszewski K and Mukerji I, 2003. "HU Binding to Bent DNA: A Fluorescence Resonance Energy Transfer and Anisotropy Study †". *Biochemistry* **42** 3096–104

[52] Swinger KK, Lemberg KM, Zhang Y, and Rice PA, 2003. "Flexible DNA bending in HU-DNA cocrystal structures". *The EMBO journal* **22** 3749–60

[53] Hammel M, Amlanjyoti D, Reyes FE, Chen JH, Parpana R, Tang HYH, Larabell CA, Tainer JA, and Adhya S, 2016. "HU multimerization shift controls nucleoid compaction:". *Science Advances* **2** e1600650–e1600650

[54] Claret L and Rouviere-Yaniv J, 1997. "Variation in HU composition during growth of Escherichia coli: the heterodimer is required for long term survival1". *Journal of Molecular Biology* **273** 93–104

[55] Aki T, Choy HE, and Adhya SL, 1996. "Histone-like protein HU as a specific transcriptional regulator: co-factor role in repression of gal transcription by GAL repressor". *Genes to Cells* **1**

[56] Lia G, Bensimon D, Croquette V, Allemand JF, Dunlap D, Lewis DEA, Adhya SL, and Finzi L, 2003. "Supercoiling and denaturation in Gal repressor/heat unstable nucleoid protein (HU)-mediated DNA looping". *Proceedings of the National Academy of Sciences of the United States of America* **100** 11373 – 11377

[57] Schumacher MA and Funnell BE, 2005. "Structures of ParB bound to DNA reveal mechanism of partition complex formation". *Nature* **438** 516–9

[58] Livny J, Yamaichi Y, and Waldor MK, 2007. "Distribution of Centromere-Like parS Sites in Bacteria: Insights from Comparative Genomics". *Journal of bacteriology* **189** 8693–8703

[59] Funnell BE, 2016. "ParB Partition Proteins: Complex Formation and Spreading at Bacterial and Plasmid Centromeres". *Frontiers in Molecular Biosciences* **3**

[60] Hwang LC, Vecchiarelli AG, Han YW, Mizuuchi M, Harada Y, Funnell BE, and Mizuuchi K, 2013. "ParA-mediated plasmid partition driven by protein pattern self-organization". *The EMBO journal*

[61] Zhang H and Schumacher MA, 2017. "Structures of partition protein ParA with nonspecific DNA and ParB effector reveal molecular insights into principles governing Walker-box DNA segregation". *Genes and Development* **31**

[62] Böhm K, Giacomelli G, Schmidt A, Imhof A, Koszul R, Marbouty M, and Bramkamp M, 2020. "Chromosome organization by a conserved condensin-ParB system in the actinobacterium Corynebacterium glutamicum". *Nature Communications* **11**

[63] Fisher GLM *et al.*, 2017. "The structural basis for dynamic DNA binding and bridging interactions which condense the bacterial centromere". *eLife* **6**

[64] Chen BW, Lin MH, Chu CH, Hsu CE, and Sun YJ, 2015. "Insights into ParB spreading from the complex structure of Spo0J and parS". *Proceedings of the National Academy of Sciences* **112**(21) 6613–6618

[65] Jalal ASB and Le TBK, 2020. "Bacterial chromosome segregation by the ParABS system". *Open Biology* **10** 200097

[66] Bramkamp M and van Baarle S, 2009. "Division site selection in rod-shaped bacteria". *Current opinion in microbiology* **12** 683–8

[67] Wu LJ and Errington J, 2004. "Coordination of Cell Division and Chromosome Segregation by a Nucleoid Occlusion Protein in Bacillus subtilis". *Cell* **117** 915–25

[68] Pang T, Wang X, Lim H, Bernhardt TG, and Rudner DZ, 2017. "The nucleoid occlusion factor Noc controls DNA replication initiation in Staphylococcus aureus". *PLoS genetics* **13** e1006908

[69] Wu LJ, Ishikawa S, Kawai Y, Oshima T, Ogasawara N, and Errington J, 2009. "Noc protein binds to specific DNA sequences to coordinate cell division with chromosome segregation". *The EMBO journal* **28** 1940–52

[70] Wu LJ and Errington J, 2011. "Nucleoid occlusion and bacterial cell division". *Nature reviews. Microbiology* **10** 8–12

[71] Sievers J, Raether B, Perego M, and Errington J, 2002. "Characterization of the parB-Like yyaA Gene of Bacillus subtilis". *Journal of bacteriology* **184** 1102–11

[72] Edgar RC, 2004. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic acids research* **32 5** 1792–7

[73] Jalal AS, Tran NT, Stevenson CE, Chan EW, Lo R, Tan X, Noy A, Lawson DM, and Le TB, 2020. "Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family". *Cell Reports* **32**(3) 107928

[74] Shimomura O, Johnson FH, and Saiga Y, 1962. "Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan, Aequorea". *Journal of Cellular and Comparative Physiology* **59**(3) 223–239

[75] Ormö M, Cubitt AB, Kallio K, Gross LA, Tsien RY, and Remington SJ, 1996. "Crystal Structure of the Aequorea victoria Green Fluorescent Protein". *Science* **273**(5280) 1392–1395

[76] Ward WW, Cody CW, Hart RC, and Cormier MJ, 1980. "Spectrophotometric identity of the energy transfer chromophores in renilla and Aequorea green-fluorescent proteins". *Photochemistry and Photobiology* **31**(6) 611–615

[77] Morise H, Shimomura O, Johnson FH, and Winant J, 1974. "Intermolecular energy transfer in the bioluminescent system of Aequorea". *Biochemistry* **13**(12) 2656–2662

[78] Chalfie M, Tu Y, Euskirchen G, Ward WW, and Prasher DC, 1994. "Green Fluorescent Protein as a Marker for Gene Expression". *Science* **263**(5148) 802–805

[79] Cormack BP, Valdivia RH, and Falkow S, 1996. "FACS-optimized mutants of the green fluorescent protein (GFP)". *Gene* **173**(1) 33–38

[80] Burnette DT, Sengupta P, Dai Y, Lippincott-Schwartz J, and Kachar B, 2011. "Bleaching/blinking assisted localization microscopy for superresolution imaging using standard fluorescent molecules". *Proceedings of the National Academy of Sciences* **108**(52) 21081–21086

[81] Zacharias DA, Violin JD, Newton AC, and Tsien RY, 2002. "Partitioning of Lipid-Modified Monomeric GFPs into Membrane Microdomains of Live Cells". *Science* **296**(5569) 913–916

[82] Luscombe NM and Thornton JM, 2002. "Protein-DNA Interactions: Amino Acid Conservation and the Effects of Mutations on Binding Specificity". *Journal of molecular biology* **320** 991–1009

[83] Pabo CO and Nekludova L, 2000. "Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?" *Journal of molecular biology* **301** 597–624

[84] Choo Y and Klug A, 1997. "Physical basis of a protein-DNA recognition code". *Current opinion in structural biology* **7** 117–25

[85] Seeman NC, Rosenberg JM, and Rich A, 1976. "Sequence-specific recognition of double helical nucleic acids by proteins". *Proceedings of the National Academy of Sciences of the United States of America* **73** 804–8

[86] Rohs R, Jin X, West SM, Joshi R, Honig B, and Mann RS, 2010. "Origins of Specificity in Protein-DNA Recognition". *Annual review of biochemistry* **79** 233–69

[87] Abe N, Slattery M, Dror I, Rohs R, Honig B, and Mann RS, 2015. "Deconvolving the Recognition of DNA Shape from Sequence". *Journal of Biomolecular Structure and Dynamics* **31** 43

[88] McKeown A, Bridgham J, Anderson DW, Murphy MN, Ortlund E, and Thornton JW, 2014. "Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module". *Cell* **159** 58–68

[89] Anderson DW, McKeown AN, and Thornton JW, 2015. "Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites". *eLife* **4**

[90] Tracewell CA and Arnold FH, 2009. "Directed enzyme evolution: climbing fitness peaks one amino acid at a time." *Current opinion in chemical biology* **13 1** 3–9

[91] Riggs AD, Bourgeois S, and Cohn M, 1970. "The lac represser-operator interaction: III. Kinetic studies". *Journal of Molecular Biology* **53**(3) 401–417

[92] Berg OG, Winter RB, and Von Hippel PH, 1981. "Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory". *Biochemistry* **20**(24) 6929–6948

[93] Halford SE and Marko JF, 2004. "How do site-specific DNA-binding proteins find their targets?" *Nucleic Acids Research* **32**(10) 3040–3052

[94] Park S, Lee OC, Durang X, and Jeon JH, 2021. "A mini-review of the diffusion dynamics of DNA-binding proteins: experiments and models". *Journal of the Korean Physical Society* **78** 408–426

[95] Dahirel V, Paillusson F, Jardat M, Barbi M, and Victor JM, 2009. "Nonspecific DNA-protein interaction: why proteins can diffuse along DNA." *Physical review letters* **102 22** 228101

[96] Kabata H, Kurosawa O, Arai I, Washizu M, Margarson S, Glass RE, and Shimamoto N, 1993. "Visualization of single molecules of RNA polymerase sliding along DNA." *Science* **262 5139** 1561–3

[97] Terakawa T, Kenzaki H, and Takada S, 2012. "p53 searches on DNA by rotation-uncoupled sliding at C-terminal tails and restricted hopping of core domains." *Journal of the American Chemical Society* **134 35** 14555–62

[98] Ober RJ, Tahmasbi A, Ram S, Lin Z, and Ward ES, 2015. "Quantitative Aspects of Single Molecule Microscopy." *IEEE signal processing magazine* **32 1** 58–69

[99] Hedglin M and O'Brien PJ, 2010. "Hopping enables a DNA repair glycosylase to search both strands and bypass a bound protein." *ACS chemical biology* **5 4** 427–36

[100] Lomholt MA, van den Broek B, Kalisch SM, Wuite GJL, and Metzler R, 2009. "Facilitated diffusion with DNA coiling". *Proceedings of the National Academy of Sciences* **106** 8204 – 8208

[101] Roe DR and Brooks BR, 2020. "A protocol for preparing explicitly solvated systems for stable molecular dynamics simulations." *The Journal of chemical physics* **153 5** 054123

[102] Macke TJ and Case DA, 1998. "Modeling unusual nucleic acid structures". ACS Publications

[103] Dill KA, Ozkan SB, Shell MS, and Weikl TR, 1993. "The protein folding problem." *Annual review of biophysics* **37** 289–316

[104] wwPDBconsortium, 2018. "Protein Data Bank: the single global archive for 3D macromolecular structure data". *Nucleic Acids Research* **47** D520 – D528

[105] Jumper JM *et al.*, 2021. "Highly accurate protein structure prediction with AlphaFold". *Nature* **596** 583 – 589

[106] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, and Kollman PA, 1996. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules". *Journal of the American Chemical Society* **118**(9) 2309–2309

[107] Case DA, 2018. "AMBER 18". URL `https://ambermd.org/`

[108] Hooke R, 1678. *De Potentia Restitutiva, or of Spring. Explaining the Power of Springing Bodies*. London: Royal Society

[109] Eisenschitz R and London F, 1930. "Über das Verhältnis der van der Waalsschen Kräfte zu den homöopolaren Bindungskräften". *Zeitschrift für Physik* **60** 491–527

[110] Price DJ and Brooks CL, 2004. "A modified TIP3P water potential for simulation with Ewald summation". *The Journal of chemical physics* **121** 10096–103

[111] Poisson SD, 1823. *Mémoire sur la Théorie du Magnétisme*

[112] Holst MJ and Saied F, 1992. "Multigrid solution of the Poisson—Boltzmann equation". *Journal of Computational Chemistry* **14**

[113] Kirkwood JG, 1934. "Theory of Solutions of Molecules Containing Widely Separated Charges with Special Application to Zwitterions". *Journal of Chemical Physics* **2** 351–361

[114] Tanford C and Kirkwood JG, 1957. "Theory of Protein Titration Curves. I. General Equations for Impenetrable Spheres". *Journal of the American Chemical Society* **79** 5333–5339

[115] Onufriev AV and Case DA, 2019. "Generalized Born Implicit Solvent Models for Biomolecules." *Annual review of biophysics* **48** 275–296

[116] Still WC, Tempczyk A, Hawley RC, and Hendrickson TF, 1990. "Semianalytical treatment of solvation for molecular mechanics and dynamics". *Journal of the American Chemical Society* **112** 6127–6129

[117] Onufriev AV, Bashford D, and Case DA, 2000. "Modification of the Generalized Born Model Suitable for Macromolecules". *Journal of Physical Chemistry B* **104** 3712–3720

[118] Hawkins GD, Cramer CJ, and Truhlar DG, 1996. "Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium". *The Journal of Physical Chemistry* **100** 19824–19839

[119] Onufriev AV, Bashford D, and Case DA, 2004. "Exploring protein native states and large-scale conformational changes with a modified generalized born model". *Proteins: Structure* **55**

[120] Mongan JT, Simmerling C, Mccammon JA, Case DA, and Onufriev AV, 2007. "Generalized Born model with a simple, robust molecular volume correction." *Journal of chemical theory and computation* **3** 156–169

[121] Debye P and Hückel E, 1923. "Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen". *Physikalische Zeitschrift* **24** 185–206

[122] Nguyen H, Pérez A, Bermeo S, and Simmerling C, 2015. "Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins". *Journal of Chemical Theory and Computation* **11**(8) 3714–3728

[123] Geney R, Layten M, Gomperts R, Hornak V, and Simmerling C, 2006. "Investigation of Salt Bridge Stability in a Generalized Born Solvent Model." *Journal of chemical theory and computation* **2 1** 115–27

[124] Lang EJM, Baker EG, Woolfson DN, and Mulholland AJ, 2022. "Generalized Born Implicit Solvent Models Do Not Reproduce Secondary Structures of De Novo Designed Glu/Lys Peptides". *Journal of Chemical Theory and Computation* **18** 4070 – 4076

[125] Euler L, 1768. "Institutionum calculi integralis"

[126] Hockey RW, 1970. "The potential calculation and some applications". *Methods in Computational Physics* **9** 136 – 211

[127] Verlet L, 1967. "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". *Physical Review* **159** 98–103

[128] Birdsall CK and Langdon AB, 1985. *Plasma Physics via Computer Simulations*. McGraw-Hill Book Company

[129] Berendsen H, Postma J, van Gunsteren W, DiNola A, and Haak J, 1984. "Molecular-Dynamics with Coupling to An External Bath". *The Journal of Chemical Physics* **81** 3684

[130] Langevin P, 1908. "Sur la theorie du mouvement brownien". *Comptes rendus de l'Académie des Sciences* **146** 530 – 533

[131] Ryckaert JP, Ciccotti G, and Berendsen HJC, 1977. "Numerical integration of cartesian equations of motion of a system with constrants: Molecular dynamics of n-alkanes". *Journal of Chemical Physics* **23** 327 – 341

[132] Torrie GM and Valleau JP, 1977. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". *Journal of Computational Physics* **23** 187–199

[133] von Helmholtz H, 1891. *On the thermodynamics of chemical processes*. Physical Memoirs Selected and Translated from Foreign Sources

[134] Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, and Kollman PA, 1992. "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method". *Journal of Computational Chemistry* **13**(8) 1011–1021

[135] Binnig G, Quate C, and Gerber C, 1986. "Atomic Force Microscope". *Physical Review Letters* **56** 930–933

[136] Pyne A, Thompson R, Leung C, Roy D, and Hoogenboom B, 2014. "Single-Molecule Reconstruction of Oligonucleotide Secondary Structure by Atomic Force Microscopy". *Small* **10** 3257–3261

[137] Asakawa H, Ikegami K, Setou M, Watanabe N, Tsukada M, and Fukuma T, 2011. "Submolecular-scale imaging of α-helices and C-terminal domains of tubulins by frequency modulation atomic force microscopy in liquid." *Biophysical journal* **101 5** 1270–6

[138] Dufrêne YF, 2014. "Atomic Force Microscopy in Microbiology: New Structural and Functional Insights into the Microbial Cell Surface". *mBio* **5**

[139] Ammar HB, 2017. "Investigation of ternary AlInN and quaternary AlGaInN alloys for High Electron Mobility Transistors by Transmission Electron Microscopy"

[140] Hansma PK *et al.*, 1994. "Tapping mode atomic force microscopy in liquids". *Applied Physics Letters* **64** 1738–1740

[141] Haynes PJ, Main KHS, and Pyne AL, 2020. "Atomic Force Microscopy of DNA and DNA-Protein Interactions v1". *protocols.io*

[142] Müller DJ, Fotiadis DI, Scheuring S, Müller SA, and Engel A, 1999. "Electrostatically balanced subnanometer imaging of biological specimens by atomic force microscope." *Biophysical journal* **76 2** 1101–11

[143] Moreno-Herrero F and Gómez-Herrero J, 2012. "AFM: basic concepts"

[144] Casuso I, Kodera N, le Grimellec C, Ando T, and Scheuring S, 2009. "Contact-mode high-resolution high-speed atomic force microscopy movies of the purple membrane." *Biophysical journal* **97 5** 1354–61

[145] Hansma PK *et al.*, 1994. "Tapping mode atomic force microscopy in liquids". *Applied Physics Letters* **64** 1738–1740

[146] Kwaśniewska A, Świetlicki M, Prószyński A, and Gładyszewski G, 2021. "The Quantitative Nanomechanical Mapping of Starch/Kaolin Film Surfaces by Peak Force AFM". *Polymers* **13**(2)

[147] Su C, Hu S, Hu Y, Erina N, and Slade A, 2010. "Quantitative Mechanical Mapping of Biomolecules in Fluid". *MRS Proceedings* **1261**

[148] Alsteens D, Dupres V, Yunus S, Latgé JP, Heinisch JJ, and Dufrêne YF, 2012. "High-resolution imaging of chemical and biological sites on living cells using peak force tapping atomic force microscopy." *Langmuir : the ACS journal of surfaces and colloids* **28 49** 16738–44

[149] Rico F, Su C, and Scheuring S, 2011. "Mechanical mapping of single membrane proteins at submolecular resolution." *Nano letters* **11 9** 3983–6

[150] Picas L, Rico F, and Scheuring S, 2012. "Direct measurement of the mechanical properties of lipid phases in supported bilayers." *Biophysical journal* **102 1** L01–3

[151] Khan Z, Leung C, Tahir BA, and Hoogenboom BW, 2010. "Digitally tunable, wide-band amplitude, phase, and frequency detection for atomic-resolution scanning force microscopy". *Review of Scientific Instruments* **81**(7) 073704

[152] Akrami SMR, Nakayachi H, Watanabe-Nakayama T, Asakawa H, and Fukuma T, 2014. "Significant improvements in stability and reproducibility of atomic-scale atomic force microscopy in liquid". *Nanotechnology* **25**

[153] Sumino A, Sumikama T, Uchihashi T, and Oiki S, 2019. "High-speed AFM reveals accelerated binding of agitoxin-2 to a K+ channel by induced fit". *Science Advances* **5**

[154] Hansen O and Boisen A, 1999. "Noise in piezoresistive atomic force microscopy". *Nanotechnology* **10** 51–60

[155] Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, and Woods RJ, 2005. "The Amber biomolecular simulation programs". *Journal of Computational Chemistry* **26** 1668–1688

[156] Salomon-Ferrer R, Case DA, and Walker RC, 2013. "An overview of the AMBER biomolecular simulation package". *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**

[157] Tsui V and Case DA, 2000. "Theory and application of the Generalized Born solvation model in macromolecules simulations". *Biopolymers* **56** 275 – 291

[158] Mongan J, Simmerling C, McCammon JA, Case DA, and Onufriev A, 2007. "Generalized Born Model with a Simple, Robust Molecular Volume Correction". *Journal of Chemical Theory and Computation* **3**(1) 156–169

[159] Onufriev A, Bashford D, and Case DA, 2004. "Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model". *Proteins* **55** 383–94

[160] Dang LX, 1995. "Mechanism and Thermodynamics of Ion Selectivity in Aqueous Solutions of 18-Crown-6 Ether: A Molecular Dynamics Study". *Journal of the American Chemical Society* **117**(26) 6954–6960

[161] Ivani I *et al.*, 2015. "Parmbsc1: A refined force field for DNA simulations". *Nature Methods* **13** 55–58

[162] Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, and Simmerling C, 2015. "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB". *Journal of Chemical Theory and Computation* **11**(8) 3696–3713

[163] Grossfield A, 2017. "WHAM: The weighted histogram analysis method". URL `http://membrane.urmc.rochester.edu/?page+id=126`

[164] Chen C, Esadze A, Zandarashvili L, Nguyen D, Pettitt BM, and Iwahara J, 2015. "Dynamic Equilibria of Short-Range Electrostatic Interactions at Molecular Interfaces of Protein—DNA Complexes". *J. Phys. Chem. Lett.* **6**(14) 2733–2737

[165] Roe DR and Cheatham TE, 2013. "PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data". *Journal of Chemical Theory and Computation* **9** 3084–3095

[166] Beton JG, Moorehead R, Helfmann L, Gray R, Hoogenboom BW, Joseph AP, Topf M, and Pyne ALB, 2020. "TopoStats – A program

for automated tracing of biomolecules from AFM images". *Methods (San Diego, Calif.)* **193** 68 – 79

[167] Dame RT, van Mameren J, Luijsterburg MS, Mysiak ME, Janicijevic AC, Paździor G, van der Vliet PC, Wyman C, and Wuite GJL, 2005. "Analysis of scanning force microscopy images of protein-induced DNA bending using simulations". *Nucleic Acids Research* **33** e68 – e68

[168] Ouldridge TE, Louis AA, and Doye JPK, 2010. "DNA Nanotweezers Studied with a Coarse-Grained Model of DNA". *Physical review letters* **104** 178101

[169] Suma A, Poppleton E, Matthies M, Šulc P, Romano F, Louis AA, Doye JPK, Micheletti C, and Rovigatti L, 2019. "TacoxDNA: A user-friendly web server for simulations of complex DNA structures, from single strands to origami". *Journal of Computational Chemistry* **40**(29) 2586–2595

[170] Procyk J, Poppleton E, and Šulc P, 2021. "Coarse-grained nucleic acid-protein model for hybrid nanotechnology." *Soft matter* **17** 3586–3593

[171] Li Z, Song LF, Li P, and Merz KMJ, 2020. "Systematic Parametrization of Divalent Metal Ions for the OPC3, OPC, TIP3P-FB, and TIP4P-FB Water Models". *Journal of Chemical Theory and Computation* **16** 4429–4442

[172] Pelta J, Livolant F, and Sikorav JL, 1996. "DNA Aggregation Induced by Polyamines and Cobalthexamine (*)". *The Journal of Biological Chemistry* **271** 5656 – 5662

[173] Bloomfield VA, 1997. "DNA condensation by multivalent cations." *Biopolymers* **44 3** 269–82

[174] Bednar J, Furrer P, Stasiak A, Dubochet J, Egelman EH, and Bates AD, 1994. "The twist, writhe and overall shape of supercoiled DNA change during counterion-induced transition from a loosely to a tightly interwound superhelix. Possible implications for DNA structure in vivo." *Journal of molecular biology* **235 3** 825–47

[175] Wang J, Wang W, Kollman PA, and Case DA, 2006. "Automatic atom type and bond type perception in molecular mechanical calculations." *Journal of molecular graphics & modelling* **25 2** 247–60

[176] Jakalian A, Jack DB, and Bayly CI, 2002. "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation". *Journal of Computational Chemistry* **23**

[177] Tan C, Terakawa T, and Takada S, 2016. "Dynamic Coupling among Protein Binding, Sliding, and DNA Bending Revealed by Molecular Dynamics." *Journal of the American Chemical Society* **138 27** 8512–22

[178] Teif VB and Rippe K, 2010. "Statistical–mechanical lattice models for protein–DNA binding in chromatin". *Journal of Physics: Condensed Matter* **22** 414105

[179] Yoo J and Aksimentiev A, 2013. "In situ structure and dynamics of DNA origami determined through molecular dynamics simulations". *Proceedings of the National Academy of Sciences* **110** 20099 – 20104

[180] Maffeo C, Yoo J, and Aksimentiev A, 2016. "De novo reconstruction of DNA origami structures through atomistic molecular dynamics simulation". *Nucleic Acids Research* **44** 3013 – 3019