Action! Modelling DNA nanomachines for deciphering their molecular mechanisms

Antoinette Alevropoulos-Borrill

Doctor of Philosophy

University of York
Physics

September 2021

Abstract

One of the most fundamental nanomachines for life is the helicase, as it is key in the process of replicating DNA. It functions as a DNA zip, separating the two strands of DNA, and allowing other nanomachines to create new DNA strands. Experimental research has increased the understanding of these helicases. However, there is still a lot to know about how these nanomachines function. Molecular dynamics can fill this gap, by providing predictions and supporting the experimental work, in order to gain a better understanding of helicase action.

One such helicase, the E1, separates the DNA strands in the papillomavirus. Within the structure of the helicase, lie many channels and chambers, which play a role in the function of the protein. The work for this thesis devised and utilised a new method, which increased the number of accessible conformations. The results found that inside one of the chambers, the DNA is pulled apart by an inner component of the helicase. Revealing that the helicase acts on the DNA, grabbing it and separating the strands in the process.

An entirely different helicase, Rep, clears the DNA of other proteins that can act as "road-blocks", as well as separating the DNA strands of E. coli. It is formed of four segments, with previous work discovering that these four segments fold into two distinct conformations, called open and closed. The work presented here, along with the complementing experiments, indicated that there were multiple other conformations. The work also demonstrated that the distribution of these conformations is affected by the salt concentration of the environment.

Contents

List of Figures

Preface

We have come a long way since the start of this field in the 1950s. Though I feel that we can sometimes forget both how far we have come, and how much further we must go. Since the dawn of modern computers our understanding of the world has increased at such a rate, it feels like we are constantly being swept off our feet by the wave of time, hastily landing in a far-flung future. We live in an age where the technology that propelled men into space, can now fit, quite literally, in the palm of our hand.

And yet, at the same time, some progress feels slower. While modern computers were being developed, and the first MD simulations produced, pioneering x-ray crystallographers developed an image of what would later reveal the structure of DNA. Photo 51, the accomplishment of Rosalind Franklin and her PhD student, Raymond Gosling, allowed the structure of DNA to be deduced. Since then both the fields of molecular dynamics and x-ray crystallography have developed, and intertwined, creating the road for me to conduct the research for this thesis. However, it is useful to note that Rosalind Franklin's role in this path is not often remembered as a pinnacle of scientific endeavour, but part of a different narrative altogether.

Our view of the world, whether it is through science or history, has a tendency of valuing certain viewpoints, and lacking a holistic perspective. While we might want to focus on some advancements more than others, we should not lose sight of what can be achieved together. Science, like all human endeavours, relies upon collaboration, both in the present and with the past; our work after all, is based upon the achievements of our predecessors.

This PhD thesis alone would not have been what it was, without the communication between two different groups, two different universities and three disciplines; biology, chemistry and physics. There is so much to be learnt by utilising a wide scope of lenses. Whether these are through different scientific views, like microscopes or computers, or social ones, like gender, class, race or culture. Perhaps the direction science could take in the future will not be dependent on one perspective but many. As the physics joke goes*, we do not exist in a vacuum.

* A farmer asks some scientists how to cure his cow. The biologist looks at the cow and says "I know the species of the cow, but I can’t cure it". The chemist examines the cow and says "I know the chemical composition of the cow, but I can’t cure it". The physicist says nothing, but starts writing pages of notes. After many days, without looking up, the physicist says, "I can cure your cow, but it only works for spherical cows in a vacuum".

Acknowledgements

Firstly, I would like to acknowledge my supervisors and collaborators, without whom I would not have learned so much. The weekly supervision of Dr Agnes Noy was invaluable, as well as the meetings with Prof Fred Antson, Dr Cyril Sanders, Dr Vladimir Levdikov, Prof Mark Leake, Dr Steve Quinn, Dr Jamieson Howard, Dr Ben Ambrose and Dr Tim Craggs. This work would also not have been possible without the funding from the EPSRC as well as the computational resources of Cambridge CSD3, JADE and the University of York (YARCC and Viking). A final thanks to Dr George Watson whose expertise on Linux and AMBER was inspirational, and the introduction of bash scripts one of the best gifts a computational PhD candidate could ask for.

I would also like to thank the amazing Department of Physics IT Support Staff. In particular, Matt Pickin, Dr Ed Higgins and David, who enabled my research to be accomplished significantly faster, by compiling pmemd with cuda support. Thank you also to Dr Agnes Noy for providing an extremely fast GPU. I have appreciated it very much.

On a more personal note, I have to thank my biggest supporter, best friend and incredible partner, Luke Tetley. I do not think I could describe in words how important you have been in helping me both through my PhD, as well as the time before. You have been my absolute rock, weathering my storms with me with absolute grace. I could not have started this PhD, let alone finish it, without you.

I would also like to thank my wonderful friends. I do not think I could find a better bunch of lovely people to have spent my free time with, especially during Covid. Thank you especially to Sarah, Joe and Ruth for your constant reassurances. Thanks also to my lovely discord MMA friends; Neel, Nathan, Ella, Tom J, Kian, Sofia, Sam, Ose, Tudor, Daisy and Ella. You are all brilliant! Thank you for the discord chats and games of Among Us, Civ, Minecraft and Werewolf, and for dealing with me when I am very bad at all of them. In essence, thank you for keeping me sane.

It seems strange to talk about friends and not mention my longest friendship of all. Thank you Harry and Becca, although we rarely see each other I could not do without our pages-long messages and hours-long talks on life.

A final thank you goes to the Open Door Team. Like so many, I have had times of poor mental health, and having the kind support of the admin staff and counsellors has made so much difference.

In Memory of Melitta

DECLARATION

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 an award at this, or any other, University. All sources are acknowledged as References.

Work that was conducted by others, for which this thesis depends, is declared throughout the text. The contributions for the E1 helicase being Professor Fred Antson and Dr Vladimir Levdikov who determined the crystal structure of the E1 helicase used throughout this work. The contributions for the FRET analysis of Rep was conducted by Dr Jamieson Howard, Dr Steve Quinn, Dr Ben Ambrose and Dr Tim Craggs.

Publications

The following publications arose from this project:
Vladimir Levdikov, Antoinette Alevropoulos-Borrill, Cyril Sanders, Agnes Noy, Alfred Antson *Mechanism of DNA strand separation by the papillomavirus E1 helicase* In preparation.

Jamieson A. L. Howard, Steven D. Quinn, Ben Ambrose, Maria Dienerowitz, Antoinette Alevropoulos-Borrill, Agnes Noy, Michael Borsch, Mahmoud Abdelhamid, Timothy D. Craggs and Mark C. Leake. *Molecular-level Insights Reveal the Rapid Structural Remodelling of Single Rep Helicases* In preparation

Chapter 1

Introduction

1.1 Proteins and their Structure

1.1.1 Primary Structure

Zooming into a living organism, past any tissues or organs, right into the cells, you find each one brimming with a buzz of movement from molecular nanomachines packed inside. It is much like a hive where each molecular machine is like a bee carrying out its own role in the collective life of the system. These molecular machines are not made from steel or tyres, as they are in the world we know, but are from what was believed to be the smallest unit of matter in existence. The atom. First theorised in ancient Greece [7], it is difficult to comprehend how far we have come in our knowledge of this unbelievably small component of life.

The simplest atom of all, hydrogen, made of only a single proton and electron, is the most prevalent in the human body [8]. After hydrogen, there are three other elements which together constitute over 95% of the human body; nitrogen, oxygen and carbon [9, 8, 10]. It seems incredible that from only a small selection of elements, so much biological material can be made. These four atoms make up the bulk of amino acids, which build up to form chains called polypeptides, the primary structure. The primary structure is the sequence of amino acids which fold into complex nanomachines called proteins, key in the workings of the organism. Again, the simplicity is beautiful in that each amino acid has these same four components, each one bound to a central carbon atom. Firstly, the amino group, composed of a nitrogen and two bound hydrogens, which gives the amino acid its name. Next is the carboxyl group, a carbon atom with a single-bound and double-bound oxygen. Finally there is a single-bound hydrogen atom and a side chain, consisting of one or more bonded atoms. Since the carboxyl group is an acid it, together with the amino group, gives the molecule its name, amino acid [11].

When amino acids join together, they form a polypeptide chain, as in figure 1.1. In these chains, one side ends with the nitrogenous amino group, called the N-terminus. Meanwhile the other side ending with the carboxyl group, is called the C-terminus. There are only twenty commonly found amino acids [12], each with a different side chain, that build up to form the vast array of proteins. The amino acid sequence determines the shape of the protein, and thus its function[13]. As a result, proteins with similar functions often have sequentially and structurally similar components, called motifs. The overall shape and function of these proteins will also be similar. When proteins have amino acid sequences where the amino acids are similar, they are said to be conserved [14].

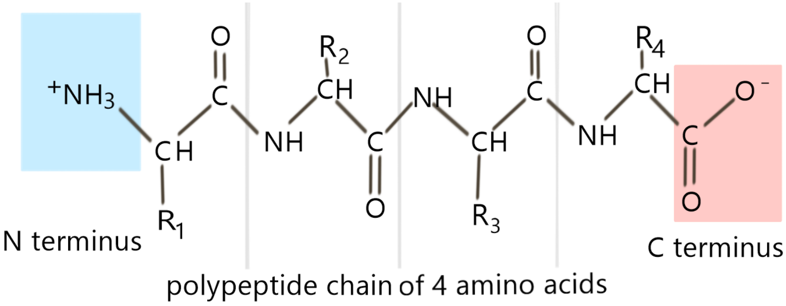

Figure 1.1: Polypeptide chain of 4 amino acids with R denoting the side chains. *N-terminus - blue, C - terminus - red.*

It is a central dogma in biochemistry that the amino acid sequence affects how the polypeptide folds and thus creates the shape or structure of the protein [11]. In turn, this shape affects the how the protein will move and therefore how it functions. The question as to how the sequence affects the shape of the protein is called the "protein folding problem" [15] and has been of great significance for over 50 years [16]. The shape of the protein is as a result of the chemical properties of the amino acids found in their side chains, including their charge and polarity [12].

There are four properties of the amino acids; polar, apolar, positively and negatively charged [12]. Polar amino acids are hydrophilic, meaning they interact well with water, and are often found on the surface of the protein, where they can interact with the water around them. Conversely, apolar or nonpolar groups are hydrophobic, Greek for water haters, who are repulsed by water. They are often buried inside the protein, hidden away from the water molecules. The properties of the amino acids not only affect their position in the protein once it has folded, but also how the protein behaves in different environments. For example, the concentration of salt, or charged ions, around the protein can change the shape of the protein.

Moving onto the side chains that are elecrostatically charged, these amino acids can have acidic and basic properties. Acids are gift-givers, donating a bound hydrogen atom as a hydrogen ion or proton. These protons are then accepted by basic molecules, binding to one of the atoms, such as nitrogen. Positively charged amino acids are acidic, while those that are negatively charged are basic. Whether the amino acid residues in a protein are protonated depends on the pH of the environment, the level of protons in the solution, as well as its pK value, the propensity of a proton to dissociate from the amino acid [17]. So amino acids with a high pK will be protonated at a neutral pH, while those with low pK will be unprotonated.

Lysine is one of the positively charged amino acids, where its side chain consists of a linear carbon chain ending with an amino group. This nitrogen is fully

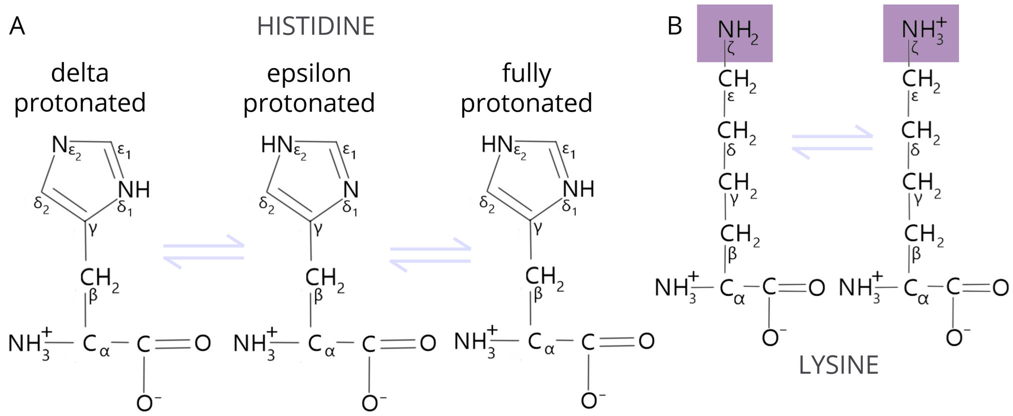

Figure 1.2: Protonation states of histidine, showing the delta, epsilon and fully protonated states (A) and lysine, showing the two protonation states of the end chain amino group (B). *Lysine zeta amino group shown in purple.*

protonated at the cellular pH of 7, because lysine has a high pK of around 10. The protonation state is less clear for other amino acids like histidine. Histine has a pentameric ring with two nitrogens, one at the delta and one at the epsilon position. These nitrogens can be doubly protonated, or singly protonated with a hydrogen at either the delta or epsilon site. Both the epsilon and delta protonated states can be present at a pH of 7, cellular pH. [18]. The different protonated states of lysine and histidine when found in polypeptides is displayed in figure 1.2.

1.1.2 3D Structure

The large chains of amino acids fold up into a diverse array of 3D structures. On the smallest level, called the secondary structure, are small local structures, which commonly fold into only two distinct structures; alpha helices and beta sheets [19]. These secondary structures can be joined together by different types of loops. The beta sheets in particular often have many sheets joined together in a plane by beta loops. If the beta sheets are aligned, and joined together by a sharp-turning loop of 3-5 amino acids, they are referred to as beta hairpins, or simply hairpins [20, 21, 22, 23]. The tertiary structure of a protein comes when it has folded into multiple alpha helices, beta sheets and loops, demonstrated in figure 1.3. The

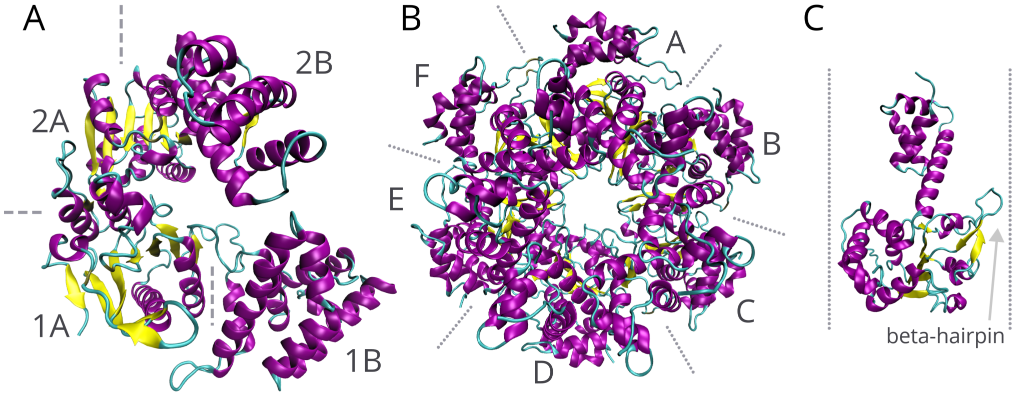

Figure 1.3: Three dimensional structures of proteins including a monomeric protein (Rep) made of four subdomains (A), and a hexameric protein (E1) made of six subunits (B), with each subunit made of two subdomains, connected by an alpha helix (C). *Alpha helices - purple, beta sheets - yellow, loops - cyan.*

structure may contain subdomains containing many alpha helices, beta sheets and loops, which join to the other subdomains via other secondary structures. Such is the case in the Rep helicase, where one strand folds into four subdomains; 1A, 1B, 2A and 2B [1].

The largest proteins have a final structural layer, its quartenary structure, when multiple amino acid chains, or subunits, come together [24]. The smallest quarternary structure would be a protein with two subunits, a dimer, however proteins can consist of many more. Proteins associated with DNA replication can often have six subunits, called hexamers, such as the E1 helicase, with its six subunits arranged in a ring [1], labelled A to F, as showing in figure 1.3B.

Proteins are not rigid, inflexible structures like a sculpture set in stone, but can move fluidly and continuously, almost as if they were a living thing [25]. Each different shape that the protein occupies is called a conformation, with the protein passing through different conformations as it carries out its function. The dihedral angles [26] result in different energies for each conformation, so that as the protein moves, it explores an energetic and conformational landscape.

1.2 DNA Structure

DNA is a fundamental component of life. It holds the genetic information of the organism, acting as an instruction manual for the life-form. Incredibly, DNA is written in an alphabet that contains only four letters. The arrangement of these letters into longer sequences of "words" distinguishes many of the differences between species and individuals. The four letters are A (adenine), T (thymine), G (guanine) and C (cytosine) and they represent different types of a molecule called nucleobases, often referred to simply as bases [11] (see figure 1.4).

The bases have two categories, pyrimidines and purines. The pyrimidines (thymine and cytosine), are smaller and contain one hexameric ring. The purines on the other hand, (guanine and adenine), are much larger and comprise of the same hexameric ring as the pyrimidines, but also a smaller pentameric ring alongside. Each base has a complementary pair, formed of a purine and pyrimidine. Adenine forms a pair with thymine, as does guanine and cytosine. In RNA, a similar molecule to DNA, thymine is replaced by a different pyrimidine, uracil.

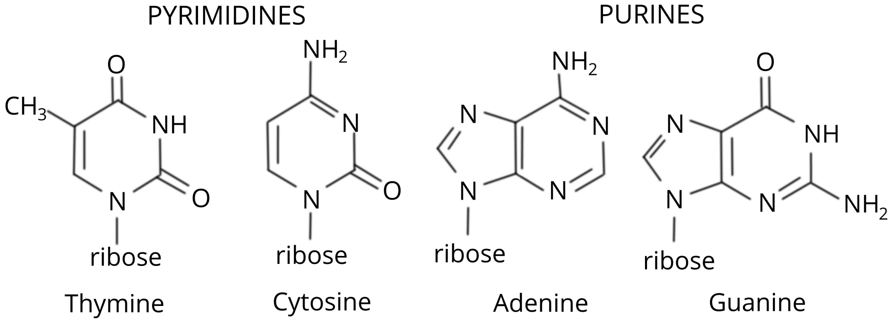

Figure 1.4: The four nucleobases of DNA: thymine, cytosine, adenine and guanine. The two purines, adenine and guanine have two aromatic rings, while the pyrimidines have one.

DNA is made not only from nucleobases but another two components; a deoxyribose sugar and a phosphate. The deoxyribose sugar is a pentose ring with four carbons and one oxygen, while the phosphate has one phosphorous atom with four oxygens. Together these three components, the nucleobase, sugar and phosphate form a nucleotide. These nucleotides then form large chains, creating a single strand of DNA. The nucleic acid chains have directionality, where the DNA strands can go from the 3′ carbon to the 5′ carbon and vice versa. The directionality comes into play when proteins bind and travel along, or translocate, DNA, since the proteins have a particular direction they travel along the DNA track.

The sugar-phosphate backbone is of particular importance in the properties of DNA, as it gives DNA a negative electrostatic charge. This allows other charged molecules, such as certain amino acids, to interact with the DNA via the Coulomb potential. Interactions can be via long-range electrostatic interactions as well as shorter range interactions like hydrogen bonds. Hydrogen bonds are weak interactions caused by the partial charge from an electronegative atom, an atom that has a tendency to attract a bonded pair of electrons. The result is a partial positive charge on the hydrogen, and a partial negative charge on the nitrogen or oxygen. A hydrogen bond is then formed between one of these hydrogen atoms and another electronegative atom.

DNA is most commonly found as double stranded DNA (dsDNA), where two strands are aligned anti-parallel. Here, one strand will be 3′-5′, while the other is aligned opposite to it, 5′-3′. The two strands are held together via hydrogen bonds between the complementary base pairs called WC hydrogen bonds (figure 1.5), as well as between the stacked bases above and below [27]. In guanine and cytosine there are three hydrogen bonds between the bases. The eletronegative atoms of 06, N1 and N2 on guanine bind to the N4, N3 and 02 atoms of cytosine, respectively.

The distinct 3D structure of dsDNA is iconic, and has been widely circulated in films and TV. The image is clear, with the bases on the inside, and the sugar-phosphate backbone of each strand, often referred to as the phosphate backbone, encircling each other [28, 29, 30, 31]. However, since the use of the image is often

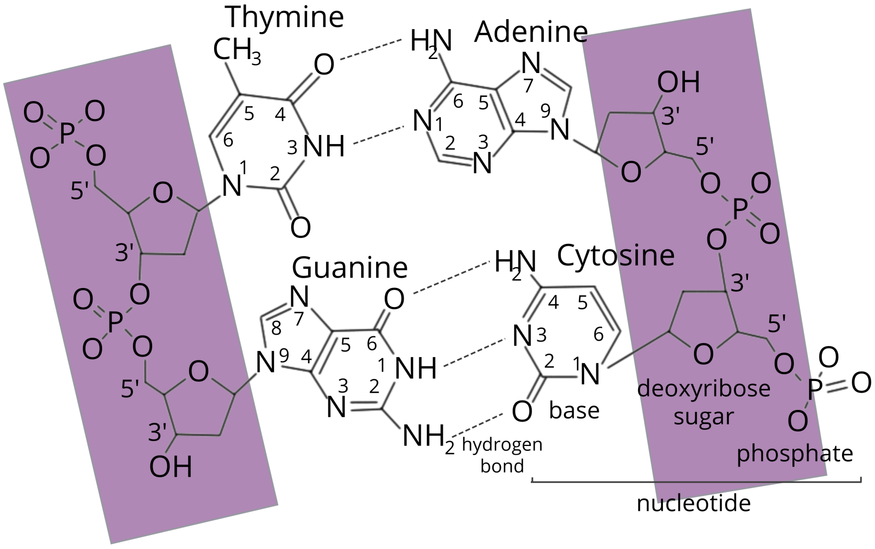

Figure 1.5: Two base pairs of DNA. The sugar and phosphate form the backbone of DNA, while all three together (base, sugar and phosphate) are called a nucleotide. The complementary base pairs are joined together by three WC hydrogen bonds in the case of guanine and cytosine, and two in the case of thymine and adenine. *Sugar phosphate backbone shown in purple.*

for entertainment, not education, the details in these media are often unclear. For example, the strands are often seen wrapping around each other evenly, with the same uniform gaps between every turn. However this is not the case, with the backbone of the two strands closer together (or further apart) than is often seen, creating a minor and major groove between the two strands [32]. The groove where the two backbones are closer being the minor groove, and the one where they are further apart, the major groove (as demonstrated in figure 1.6A). Another key characteristic of DNA is the twist. The twist, the degree to which one strand wraps around the other, can be defined as the angle between one base pair and the next [32], shown in figure 1.6B. The most common form of DNA found in nature has 10 base pairs a turn [33] so the twist between one base pair and the next is 36°.

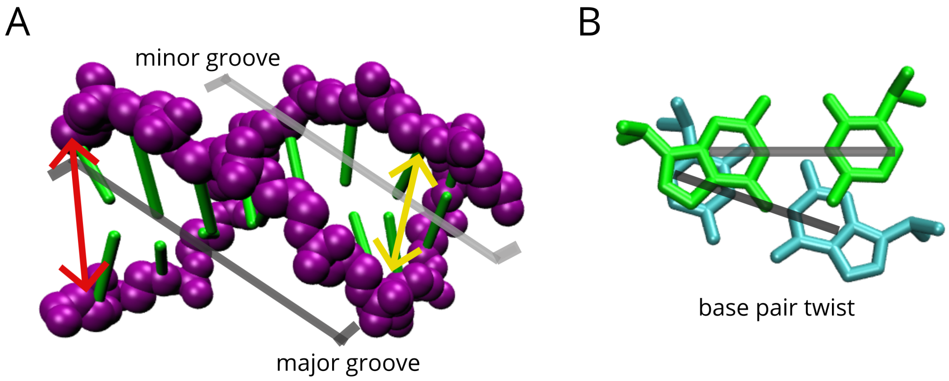

Figure 1.6: The major and minor grooves of double stranded DNA (dsDNA) (A) and the base pair twist (B). *A: DNA backbone - purple, base pairs - green, minor groove - light grey, major groove - dark grey, minor groove width - yellow, major groove width - red. B: a single base pair - green, and the one below - cyan.*

1.3 Helicases

1.3.1 Helicases and DNA Replication

In order for life to continue to exist, an organism needs to replicate itself along with its DNA. The process of DNA replication starts with a nanomachine called the helicase, a protein which separates the two strands of double-stranded DNA (dsDNA). The separated, now single-stranded DNA (ssDNA) is used as a template from which a different nanomachine, polymerase, creates a new DNA strand. The helicase translocates along the strand, while behind it, the polymerase creates the new strand on top of the template strand [34], and it is here that the complementary bases come into play. The polymerase identifies the base on the template strand and then assembles the complementary base onto the new strand, ensuring the new DNA strand is identical to the other existing DNA strand. Other proteins bind to the DNA in order to help with replication, however these may need to be removed for replication to occur [35, 36].

The replication process often begins on a specific location on the DNA called the origin of replication (ori). The helicase binds to the ori, or in some cases is assembled there by another nanomachine, and then unzips the DNA. The point on the DNA where the dsDNA is separated to form ssDNA is called the replication fork junction or simply the fork. Helicases can not only separate the two strands of DNA for replication but for other DNA processes such as DNA repair, or transcription, when the DNA is read and copied into RNA.

1.3.2 Helicase Superfamilies

Helicases have amino acid sequences that are conserved across different helicases, resulting in similar structures and mechanisms. These similarities allow the helicases to be grouped into six superfamilies (SF) [1]. The helicases in SF1 and SF2 are structurally similar and are most often monomers, while those in SF3-6 are hexameric, having six subunits, which form a ring around the DNA. There are different methods of DNA translocation across the superfamilies. Helicases in SF1 only translocate along ssDNA while helicases in other superfamilies have been found to translocate along either ssDNA or dsDNA. The directionality also differs with SF3 helicases being the only superfamily to translocate with 3′-5′ directionality, from the 3′ to 5′ carbon atoms. Helicases within other superfamilies move either 3′-5′ or 5′-3′.

The most studied helicases in SF1 are Rep, UvrD and PcrA, which have a 3′-5′ directionality [4, 37, 38]. Within the SF3 group are viral DNA and RNA helicases, including the E1 helicase, the most studied helicase of this group [39]. The hexameric helicases in the other superfamilies, like the T7gp4 in SF4 and MCM in SF6, have a similar structure to that of E1. Each helicase forms a hexameric ring around the DNA with beta hairpins that protrude into a central channel where the DNA is located. These loops translocate the DNA, like wheels on a track.

1.3.3 ATP hydrolysis

Nanomachines have surprising similarity with their macroscopic counterparts, despite being nine orders of magnitude smaller. They take in a fuel, a molecule called ATP, which provides energy to the moving parts of the helicase, allowing it to move along a track of DNA. The energy comes from breaking a bond in adenosine triphosphate (ATP) which releases energy. As the name suggests, ATP is formed of adenosine (adenine bound to a ribose sugar ring), which in turn is bound to three phosphate groups. When the bond between the last phosphate group is broken, and the phosphate is cleaved off, energy is released creating adenosine diphosphate (ADP) with the remaining two phosphates, as well as the separated phosphate. This process is called ATP hydrolysis and is coordinated by particular amino acid motifs called the AAA+ and Rec-A folds. The subunits in proteins have different states depending on the state of ATP hydrolysis. In hexameric helicases the states are T when ATP is bound, D when ATP has been hydrolysed and converted to ADP, and E when the state is empty and ADP has dissociated [1].

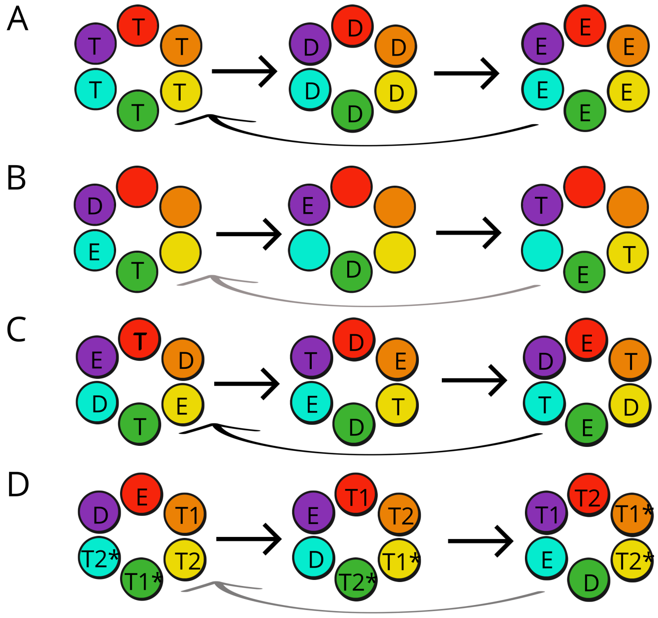

Figure 1.7: Four different mechanism of ATP hydrolysis by hexameric helicases. (A) The concerted method; (B) the stochastic method; (C) The sequential method and (D) The rotary method. *Adapted from figure 8 Singeleton et al. 2007 [1].*

Since hexameric helicases have more than one subunit, there are different mechanisms by which each subunit hydrolyses ATP, including the concerted, stochastic or sequential mechanism. In the concerted mechanism, ATP hydrolysis is coordinated with all subunits cycling through the same states at the same time, as shown in figure 1.7A. The stochastic method, shown in 1.7B, is more erratic with states binding ATP and hydrolysing independently of the others. Finally, in the sequential mechanism a rotary wave takes place throughout the subunits, as in 1.7C. These conformations cycle around the protein much like a "Mexican wave", with each subunit changing conformation with the hydrolysis of ATP.

In some helicases, such as the Rho and E1 helicase, the sequential method is extended so that each subunit occupies a different state. Instead of cycling through the three states, the subunits cycle through six, E-T1-T2-T1*-T2*-D [40, 41], as shown in 1.7D. When the helicase bind evermore tightly to the ATP through its T states, with hydrolysis occurring during the T1* or T2* states. This mechanism is often referred to as a rotary wave.

1.3.4 Passive vs Active Helicases

All helicases can be considered active as they convert energy in order to translocate DNA. However the degree to which a helicase separates strands of DNA can be described as being "active" or "passive". Despite converting energy, passive helicases wait for thermal fluctuations to separate the strands of DNA and cause it to fray. They then capture the ssDNA nucleotides, and use the energy from ATP hydrolysis to translocate along the strand. On the other hand, active helicases destabilise the DNA, with the energy from ATP being used to both translocate along the DNA and cause separation of the two strands. The distinction between active and passive helicases is not always clear, with the mechanism of the helicase having both active and passive components.

1.4 The E1 Papillomavirus Helicase

1.4.1 The Papillomavirus

The human papillomavirus (HPV) has become of particular interest due to its correlation with cervical cancer [42, 43]. There are 100 different types of HPV with most infections of HPV causing no problems for the host, though in some cases it can develop into skin warts or papillomas [44]. In the worst cases, HPV can cause cancer of the cervix, throat or neck. There is no current cure to the disease, but a vaccine has been developed and has been administered in the UK to school age girls since 2008 as part of the national immunisation programme [45]. Further confirming the severity, the World Health Organisation (WHO) recommends all countries vaccinate against HPV as part of preventative measures [46, 47].

Even in countries, such as the UK, that have administered the vaccine, it only has the best efficacy in people who are not yet sexually active. As a result anyone older than the first generation to be vaccinated, or anyone in a country without access to the vaccine, is still vulnerable to the disease [48]. As there is currently no cure for the papillomavirus anyone infected will be at risk of developing cancer. Drugs tackling HPV can therefore be used as a vital weapon in the "war on cancer", with the helicase being targeted due to its role as a key component in replication of the virus [49]. If it is understood how the helicase works, it may be easier to devise a way to stop its function, and thus prevent the papillomavirus from replicating.

The human papillomavirus is not the only papilloma variant of concern or importance. The papillomavirus is found in 20 different mammalian species where it also develops into papillomas [44]. The bovine papillomavirus, found in key livestock like cows, can effect the lives of the cows, as well as the farmers in developing countries upon whom their livelihood depends [50, 51]. Elements of the bovine papillomavirus (BPV), such as the E1 helicase is much easier to study than the human papillomavirus (HPV), since the whole helicase structure has been resolved in BPV but not HPV. The E1 helicase of BPV is widely studied and can be used as a model to compare to others [52]. As a result, it was the BPV E1 helicase that was used in the simulations, found within this thesis.

1.4.2 The E1 Helicase

Once the papillomavirus has infected a host cell, its helicase, named E1, begins to separate the DNA strands [53]. The E1 helicase binds to the origin of replication (ori) on the viral dsDNA as two hexamers, separating the two strands, with each hexamer encompassing a different DNA strand (figure 1.8) [2]. The helicases then translocate in opposite directions along the DNA, separating the strands as they go [54].

Each helicase consists of four major parts; the N-terminal domain, the ori DNA binding domain, the collar domain (CD) and the AAA+ motor domain [39]. Together, the CD and AAA+ motor domain make up the helicase domain which separates the DNA strands. In bovine papillomavirus, the first 308 amino acid residues in each subunit make up the N-terminal domain and DNA-binding domain. The helicase domain then goes from residue 308 to 605. The breakdown of the different components and residue numbers can be seen in figure 1.9. Inside the AAA+ domain are the translocating beta hairpins, with residue numbers 504 to 508. These five amino acids are aspartic acid (D504), arginine (R505), lysine (L506), histidine (H507) and a final lysine (K508) [39]. In this case, all but the aspartic acid is positively charged, which is negatively charged.

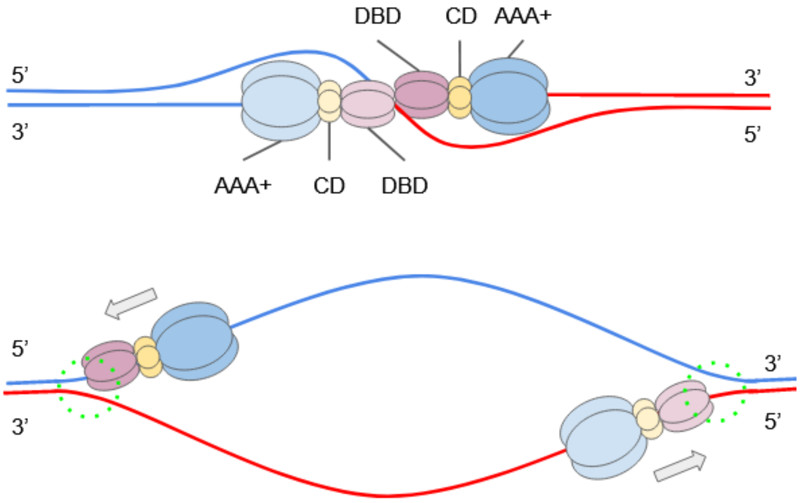

Figure 1.8: The E1 helicase attaching at the ori as a double hexamer, formed of the DNA-binding domain (DBD), collar domain (CD) and AAA+ motor domain on dsDNA. The DBD separates the dsDNA, and translocates along the dsDNA in the 3′-5′ direction of the active strand. The green circle shows the area of the replication fork junction (RFJ), where the literature presumed the dsDNA is separated into ssDNA. *Adapted from figure 7 Lee et al 2014 [2]*

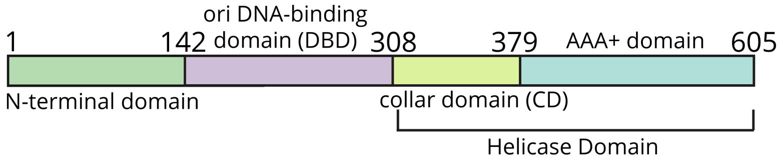

Figure 1.9: Amino acid residue numbers and domains of the E1 helicase *Adapted from figure 1 Chaban et al 2015 [3]*

1.4.3 Passive vs Active Mechanisms

Although the passive vs active mechanism has been investigated in other helicases [55, 56], it has not yet been conducted for E1. As a result, the question is still open as to whether E1 passively unwinds and separates dsDNA, or whether it does so actively. It was previously believed that E1 utilised a steric exclusion mechanism, whereby one strand passed through the helicase, with the other excluded. The excluded strand was believed to play a passive role, and was consequently named the passive strand [52]. If there was an active role of the helicase it would then be on the translocating strand, duly named the active strand. Recent developments found that the dsDNA fork was found inside the helicase [3]. As a result the helicase may be more likely to have an active mechanism since it would be able to interact with the DNA fork.

1.4.4 Rotary Wave Mechanism

The E1 helicase has a sequential mechanism of ATP hydrolysis, also called the rotary wave mechanism [40]. When the rotary wave (the wave of changing conformations around the AAA+ domain) occurs, the subunits do not change position, but change their conformation, in particular, the height of the DNA-binding loops. When a conformational change has occurred, these translocating loops inside the AAA+ motor domain, move the DNA along by one step. The different ATP states of the AAA+ motor domain remain the same whether or not the helicase is in the presence of DNA or ATP [57].

1.4.5 Escort Mechanism

Inside the AAA+ motor domain are the translocating loops, beta hairpins, which track ssDNA backbone in the central channel. The hairpins are arranged in a right-handed spiral staircase around the central channel [39], as shown in figure 1.10. Each of the six hairpins is part of one of the six subunits that form the AAA+ hexameric ring. When each subunit cycles through six stages of ATP binding, the hairpin moves in conjunction with the state of their corresponding subunit. A subunit in its ATP state, has the the hairpin at the top, when it is in its apo state (without ATP or ADP) the hairpin will be in the bottom position, and will be at intermediary positions when in its intermediate states. The hairpins translocate the ssDNA via an "escort mechanism", whereby each hairpin grabs a free ssDNA nucleotide at the top of the staircase and "escorts" it down the channel until it reaches the bottom and disengages. It then moves back to the top of the channel and repeats the journey with a new nucleotide [39] as shown in figure 1.11.

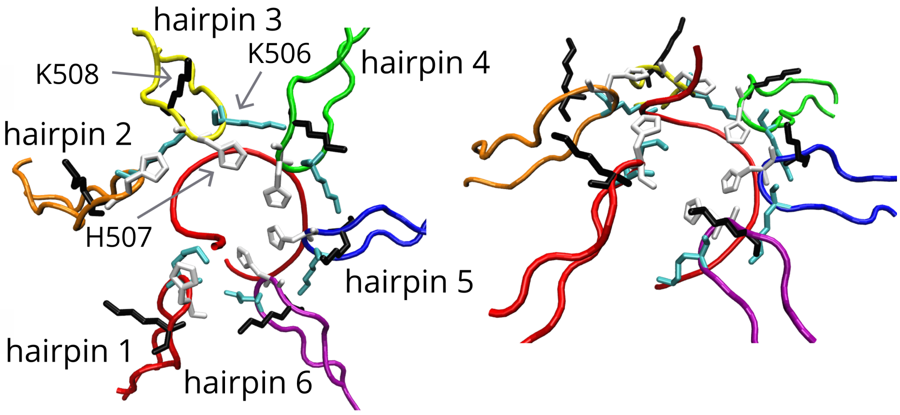

Figure 1.10: The six translocating hairpins, and amino acids lysine K508, histidine H507 and lysine K506. Left. View from above. Right. View from the side. *DNA backbone - red, K508 - black, H507 - white, K506 - cyan.*

As can be seen in figure 1.10 the hairpins have a complex structure of amino acids, whose interaction bind the DNA and maintain the staircase structure. Going round the hairpin starting from the highest residue, there is lysine K508, histidine

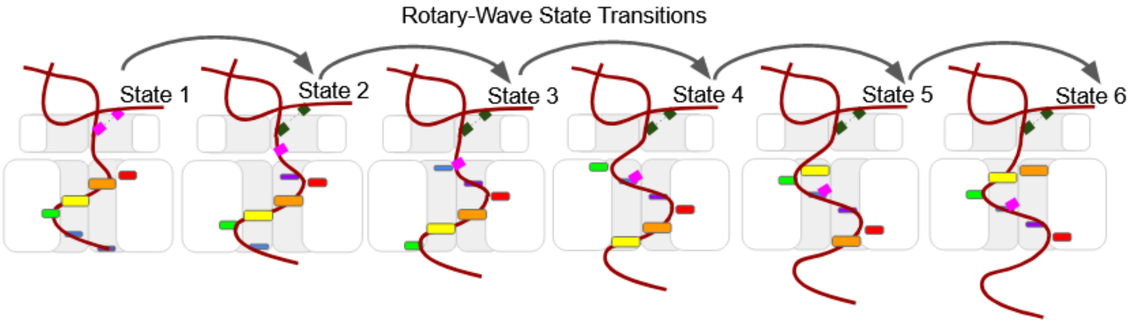

Figure 1.11: Schematic of E1 helicase with staircasing hairpins in the AAA+ motor channel. The subunits A to F cycle through the different ATP states, corresponding with different hairpin heights. *Hairpin of each subunit: A - red, B - orange, C - yellow, D - green, E - blue, F - purple. DNA backbone - dark red, last dsDNA base pair in each state - green, one translocating nucleotide - pink. Same colour code as figure 1.7.*

H507, lysine K506, arginine R505 and aspartic acid D504. The staircase structure is maintained by hydrogen bonds between these amino acids in different hairpins [39]. The lysine K506, tracks the DNA backbone and binds to the DNA via a hydrogen bond between its zeta amino group and the phosphate backbone. It also has a role in maintaining the hairpin staircase by hydrogen bonding between the zeta group and the main chain amino group of K508 and R505, as well as the side chain of D404. In the crystallographic structure, the only other hairpin amino acid to bind to the DNA is the histidine, H507, which forms hydrogen bonds with the backbone of an adjacent nucleotide between the phosphate and its main chain amino group.

1.4.6 Recent Developments

Recently, electron microscopy found that the helicase had many channels and chambers where the DNA could pass in or out [3], as can be seen in figure 1.12A and B. The DNA could not be observed explicity, but further work, labelling the DNA with protein tags, determined the location of the DNA in the helicase [3].

As can be seen in figure 1.12C, the dsDNA was found to enter via a side channel between the N-terminal and DNA-binding domain. It passes through the helicase as dsDNA, until it reaches the collar domain (CD). Here, the two strands are separated and unwound, with the active strand passing through the CD, and onwards through the AAA+ domain. It is inside the AAA+ domain where the hairpins bind and translocate the ssDNA strand. The passive strand passes over the top of the CD and exits through a narrow channel between the DNA-binding domain and the AAA+ domain. The location of the dsDNA and point of unwinding was found to be fixed, so that the dsDNA rotated one nucleotide with each translocating step. The positions of the exiting passive and active strands were also fixed.

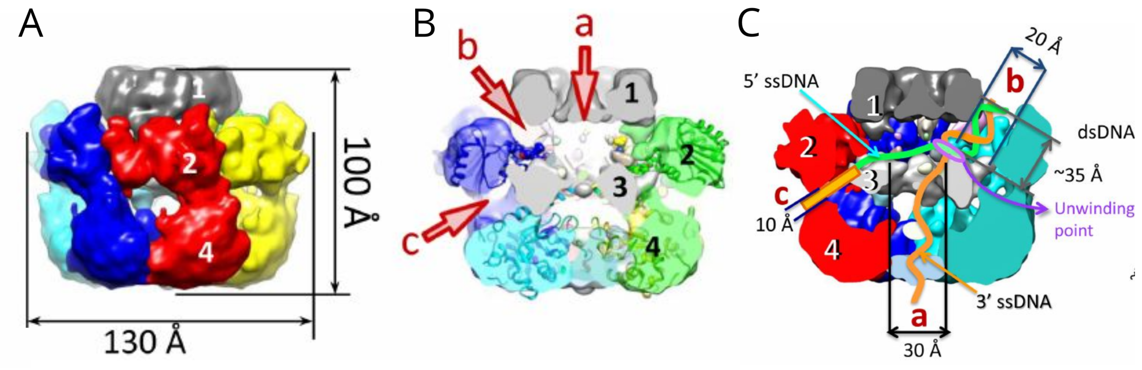

Figure 1.12: The full structure of the BPV E1 helicase (A) with cross-section showing three side channels a, b and c (B) where the DNA enters and exits (C). *1 - The N-terminal domain, 2 - DNA binding domain, 3 - Collar Domain 4 - AAA+ domain. From figure 3 Chaban et al 2015 [3]*

An unpublished crystallographic structure from the Antson group, confirmed the results from the electron microscopy and DNA labelling. It found that ten base pairs of dsDNA enter the helicase via the side channel, with each nucleotide having a twist of 36°. The dsDNA separates in the collar domain, with the first ssDNA nucleotides on the active strand entering into a large bell-shaped chamber. The rest of the ssDNA passes through the AAA+ domain, with six ssDNA nucleotides inside, a twist of 60° with the rest exiting the channel.

The atomic resolution allowed most of the DNA conformations to be observed. However, the two first two ssDNA nucleotides in the chamber were a mystery. The resolution of these nucleotides was very low, signifying dynamic movement and conformational flexibility. Additionally, the chamber appeared large enough that a variety of conformations could be allowed. However, it was unknown what these were. Equally as unknown, was the role of the chamber. There was also still the question as to how the DNA was separated, and how the hairpins grab the DNA.

Professor Fred Antson believed simulations of the helicase could answer these questions. As a result, a collaboration was set up between the Antson group in the York Structural Biology Laboratory and the Noy group in the University of York Physics Department. Preliminary molecular dynamics simulations showed that the two nucleotides in the chamber were highly flexible, moving through different conformations. As a result, it was hypothesised that the chamber could act as an adapter. It would provide the space for the DNA to adjust its conformation, allowing it to transition from ten nucleotides a turn in the collar domain, to six in the AAA+ domain. The adapter could also act as a "croupier", or card dealer, passing the next nucleotide like cards into the "hands" of the upcoming hairpin. Further simulations would determine if this hypothesis was true, and how it might occur, which will be discussed in this thesis.

1.5 The E. Coli Rep Helicase

1.5.1 E. coli replication

Escherichia coli or E.coli is a bacteria that lives in the lower intestines of warm-blooded animals such as humans. Unlike papillomavirus, most E.coli are harmless and do not cause adverse effects in their host, though there are a few that can cause urinary tract infections and diarrhea, which can be fatal for infants [58]. E. coli is widely studied, with research including the cells swimming behaviour [59], as well as molecular research in DNA replication [58].

E.coli has a replicative helicase called Rep. Rep is part of the SF1 family and, like many other SF1 helicases, is a monomer. There are two conformations of Rep, open and closed, that are also observed in two other structurally similar SF1 helicases, UvrD [60, 61, 62] and PcrA [63, 38]. These helicases not only separate strands of DNA, but can clear the DNA of proteins that act as "road-blocks" [64, 65], which need to be removed for the dsDNA to be separated [35, 36].

Rep has different amounts of helicase and protein-removal activity, depending on its conformation. Rep is a monomer in the absence of DNA [66], and when DNA is introduced it can travel along ssDNA as a monomer [67]. However, it cannot separate dsDNA as a single monomer, and instead works together with another Rep protein, forming a dimer [68]. Rep can separate dsDNA as a monomer if it is forced into the closed state [69], or if one of its subdomains, 2B, is removed [70, 71, 72], though this reduces its ability to remove road-blocks [72]. It seems that the closed conformation is preferable for helicase activity on dsDNA, while the open conformation is more likely to associated with translocation along ssDNA [73].

1.5.2 Rep structure

Rep is formed of four subdomains; 1A, 1B, 2A and 2B, each joined to the next via a loop that can function as a hinge between the subdomains [4]. Subdomain 1A has amino acid residues 1-84 and 196-275, with 1B formed of the residues in between, from 85-195. The same goes for 2A and 2B, with 2A carrying on from 276-374 along with 544-641. The remaining residues 375-543 then form 2B. The hinge between 1A and 1B is located between residues 71-84, while the hinge between 2A and 2B has two loops; 367-381 and 536-552 [4]. The whole sequence and allocation of domains can be seen in figure 1.13.

Rep has so far been found in two conformations, open and closed (see figure 1.14),

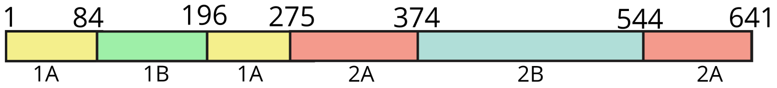

Figure 1.13: Amino acid residue numbers and subdomains of the Rep helicase.

where the subdomains have different orientations with respect to each other [4]. The open conformation was also observed in the structurally similar helicase PcrA, and was described by the researchers as looking like a "crab claw". The 1A and 2A domains at the bottom, with the 2B and 1B like two pincers, with 2B much larger.

The closed conformation differs to the open conformation by the 2B subdomain. When the 1A, 1B and 2A subdomains of the open and closed conformations are superimposed, the 2B subdomain has rotated by 130°, around the 2A-2B hinge [4]. The 2B subdomain is positioned next to 1B in the closed conformation, with the rotation in the open conformation locating it further away, resembling the large pincer.

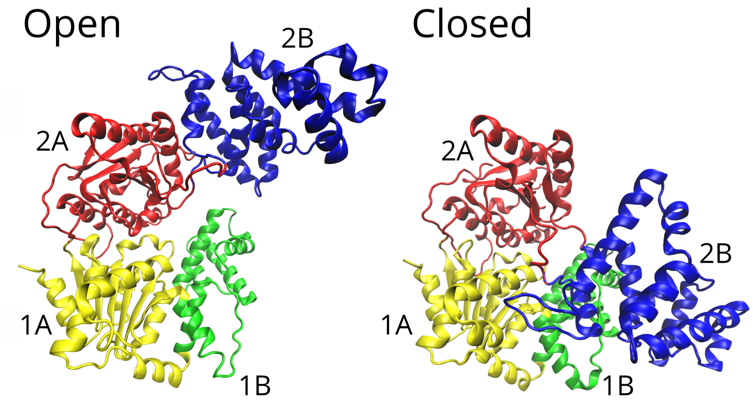

Figure 1.14: The open and closed conformations of the *e. coli* Rep helicase as from the 1UAA structure obtained by Korolev *et al* [4, 5]

1.5.3 Conformational changes

The question as to how Rep transitions between the open and closed states, is a question this thesis aims to answer. Answers have been found for the structurally similar helicase, UvrD. In solution, and in the absence of DNA, UvrD transitions between open and closed states. In low salt concentrations, the closed conformation is more common, while in high salt the open is more common [73]. The helicase was also found to pass through other conformations, totalling four different conformations [60].

Since Rep and UvrD are structurally similar, the Leake group hypothesised that Rep could also have four different states, as well as be affected by the salt concentration. Two concurrent methods were devised to see if this were true. The first used Fluorescence Resonance Energy Transfer spectroscopy (FRET) [74], which allows the distance between two dyes to be measured [75]. The dyes were attached to residues 97 (1B) and 473 (2B), to measure the distance between subdomain 1B and 2B. When the dyes, and thus the subdomains, were close together, there would be a high FRET efficiency (with the greatest being 1). Likewise, when the dyes were further apart there would be a low FRET efficiency (closer to 0).

The second method was to simulate Rep, in the same conditions as the experiments, to determine structural information. The crystallographic structures of the open and closed conformations could be used as starting structures for molecular dynamics. Then the changes in both the open and closed structures could be observed. Physical properties such as radius of gyration, and the rotation of the subdomains could also be used to categorise the different conformations.

Chapter 2

Methods

2.1 Molecular Dynamics

Molecular dynamic simulations (MD) are a method of modelling the time-evolution of particular systems, from the movement of atoms that make up a biomolecule to the swimming motion of E .coli cells [76]. MD simulations were first devised in the 1950s, by Alder and Wainwright [77], when looking at the interactions between hard spheres, and have since developed to look firstly at liquids [78, 79, 80, 81], and then at proteins [82], as well as other biomolecules such as nucleic acids and lipids. The principle underlying molecular dynamics is the same, however there can be a significant amount of variety on how it is implemented. The underlying principles are:

1. determine the initial positions and velocities of the particles
2. determine the forces acting on each of the particles
3. determine the velocity of each particle with respect to the forces
4. move the particles as determined by their velocities
5. repeat steps 2 to 4 until the number of steps has been reached

2.2 Starting structures

Before the simulation can begin there needs to be a good starting structure from which the dynamics can evolve. For biomolecular systems, the most common method for obtaining a starting structure is from experimental x-ray crystallography. Experimental structures are uploaded to the protein data bank, with codes corresponding to each structure, such as 1UAA for Rep [4, 5].

At the time of writing, the bovine papillomavirus (BPV) E1 helicase structure used for this thesis had not yet been published, and so has not yet been uploaded to the protein data bank. However, the open and closed structures for Rep can be found on the protein data bank website. As the structures are obtained as part of structural research, and not purely as the start for molecular modelling, parts of the system are missing or need to be changed. Firstly, the parts of a biomolecule that are most interesting to simulate are the ones that have the most dynamic motion. Unfortunately, this motion means it is difficult to obtain a clear structure in these areas using x-ray crystallography. In the case of Rep, two sections of the protein were missing in the starting structure. Obviously, these needed to be filled with the same sequence of amino acids, and in a reasonable structure. Fortunately, Rep has structurally similar proteins such as UvrD and PcrA [4], which can be used to fill in the gap.

The next difference between the x-ray crystallography and the requirements for MD concerns the bonded hydrogen atoms. The x-ray crystallography does not provide the coordinates of the bonded hydrogen atoms, and as a result, lacks the information about the protonated states of the amino acids. Depending on the pH (a measure of proton/hydrogen ion concentration in solution), certain amino acids can have different protonated states. The most significant of these is histidines, which are one of the amino acids making up the hairpin loops in the BPV E1 helicase. Since the protonation state of histidines is sensitive, it is important to ensure it is in the correct state. These states can be determined computationally by using a web based application such as the H++ server created by Virginia Tech [83]. The server takes an uploaded pdb and calculates the pK of the relevant amino acids in the structure. Depending on the pK of the amino acids, and the pH of the molecules environment (which is inputted by the user), the H++ server then adds any missing hydrogens. The protonated structure is then returned, as well as the calculated pK of the amino acids, all of which can be downloaded to the users local computer.

2.3 Force-fields

Once the initial positions have been determined, the next step is to calculate the forces that act upon the particles, in this case the atoms. For atomistic simulations the forces are determined by a number of energy potentials. The first one to be used was the Lennard-Jones potential [84, 85, 86], shown in figure 2.1, which describes the non-bonded potential between a pair of neutral atoms:

$$V_{LJ}(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{2.1}$$

Here, V_{LJ} is the potential, epsilon, ϵ the depth of the well (shown in figure 2.1), sigma, σ the distance where the particle interaction is zero and r is the distance between the two atoms. The potential is governed by a repulsive and attractive force, each from a different physical principle. The interaction that governs over longer ranges is the Van der Waal's force, which is generated when the electron clouds of two neutral atoms are effected by the presence of the other, causing a shift in the distribution of electrons. This shift creates opposing partial charges in both atoms, which causes an attractive force. In the Lennard-Jones potential this is conveyed by the power of six term, which occurs over distances greater than sigma. The power twelve term is then the highly repulsive term, occurring at distances smaller than sigma, where the position of the atoms would overlap. Here, Pauli's exclusion principle dominates, which states that no two fermions can occupy the same quantum state. Since the quantum state of the electrons depends, amongst other things, on its position, then obviously no two atoms can occupy the same position in space. As a result, the atoms would experience a strong repulsive force should their positions begin to overlap.

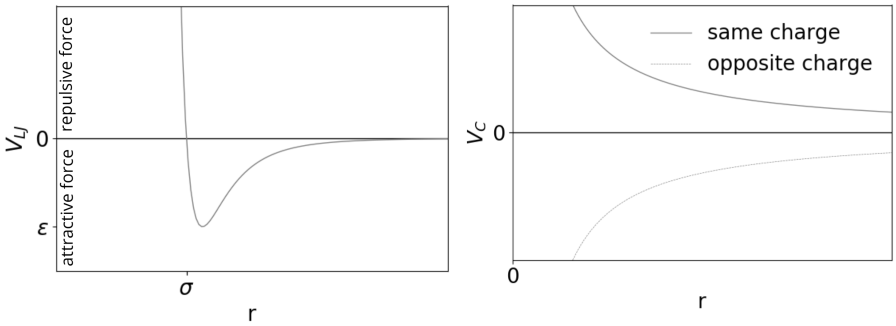

Figure 2.1: Left. The Lennard-Jones potential. When atoms are within one sigma distance, there is a strong repulsive potential. This turns into a moderate attractive potential with a depth of epsilon, which decays over longer distances. Right. The Coulomb potential, where the potential decays continuously.

The Lennard-Jones potential can be used solely to model the behaviour of non-bonded neutral charges, such as argon atoms [84, 85]. However, it does not model the effects of charged particles. For this, the Coulomb potential [87, 88] is needed:

$$V_C(r) = \frac{q_1 q_2}{4\pi\epsilon_0 r} \tag{2.2}$$

Where V_C is the Coulomb potential, q_1 the charge of atom 1, q_2 the charge of atom 2, and ϵ_0 the permittivity of free space, the capacity of the electric field to permeate a vaccuum. The Coulomb potential, shown in figure 2.1 can describe the forces on charged atoms, but extra terms are required to model any bonded potentials.

The covalent bonds between atoms are as a result of the behaviour of electrons, which is not modelled in classical MD. However the bonds can be sufficiently modelled using classical mechanics. There are three terms that can model the

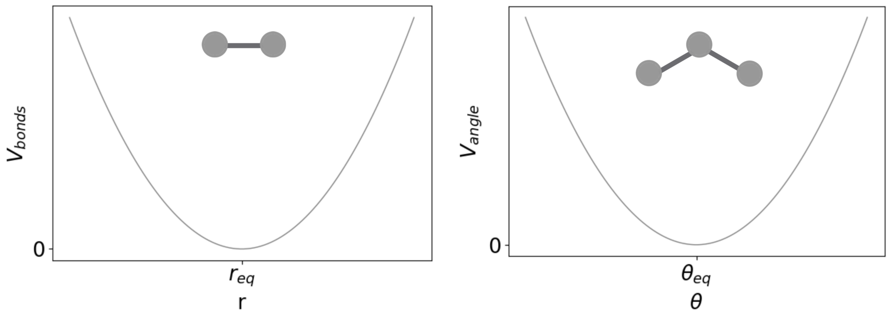

Figure 2.2: Left. The potential for a pair of covalently bonded atoms. Right. The potential for the angle between two atoms bonded to a third

dynamics of bonded interactions. The first two model the bond length and angle as Hookean springs, shown in figure 2.2. the equation for the potential for the bonds is then:

$$V_{bonds}(r) = b(r - r_{eq})^2 \tag{2.3}$$

Where, again, r is the distance between two atoms, r_{eq} the equilibrium position between the pair, and b a constant. The same Hookean spring potential can be used to model the angle between two atoms, where each atom is bonded to the same third atom.

$$V_{angle}(\theta) = a(\theta - \theta_{eq})^2 \tag{2.4}$$

Where theta, θ is angle between the atoms, θ_{eq} the equilibrium angle and a, a constant. The final bonded term comes from the dihedral angle, the angle of the two planes between four bonded atoms. When there are four sequentially bonded atoms, say A, B, C and D, then the dihedral angle is the angle between the plane formed by A, B and C, and that formed by B, C and D. The potential is described as:

$$V_{dihed} = \sum[1 + cos(n\phi - \gamma_n)] \tag{2.5}$$

This is a periodic function whereby the ϕ denotes the dihedral angle, and γ is the energy of the first peak where n=1. An example potential is described in figure 2.3.

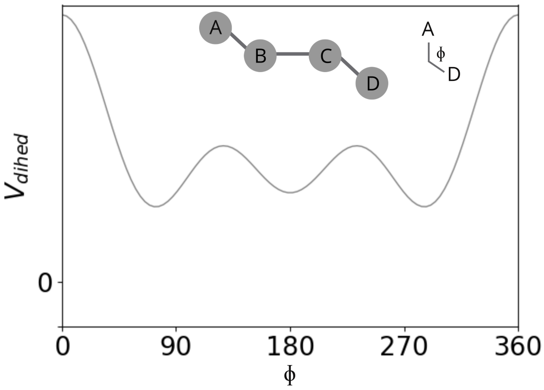

Figure 2.3: Example of the dihedral angle potential

The full potential is then an addition of all these terms, with each term summed over every atom i, and its pair j. This becomes the force-field which in the MD package AMBER is described as:

$$\begin{aligned} V_{AMBER} = \sum_{bonds} b_i(r_i - r_{i,eq})^2 + \sum_{angles} a_i(\theta_i - \theta_{i,eq})^2 \\ + \sum_{dihedrals} \sum_n (V_{i,n}/2)[1 + cos(n\phi_i - \gamma_{i,n})] \\ + \sum \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum \frac{q_i q_j}{4\epsilon_0 r_{ij}} \end{aligned} \tag{2.6}$$

Force-fields have been developed to more accurately reflect certain molecules. The ff14SB force field is the most commonly used for proteins [89], having been adapted from the previous ff94 force field [90], with both force-fields improving and stabilising the alpha helices. As for DNA, it was found that certain dihedral angles in the backbone moved away from expected values over long simulations, around 50ns [91, 92]. This was amended in the bsc0 force-field developed in Barcelona [93], with further improvements made in 2016 with the parmbsc1 force-field [94].

2.4 Integrators of Motion

Once the force-field has been determined, it is time for the method of motion. The motion of the atoms cannot be solved analytically as it is a many-body problem. So it must be solved numerically, by evolving the system step-by-step as a function of time. There are a number of techniques for this, each done in discrete time steps, where the time step is chosen based upon the time of the highest frequency movement. One such way is the Euler method:

$$v_{n+1} = v_n + a_n \Delta t \tag{2.7}$$

$$x_{n+1} = x_n + v_n \Delta t \tag{2.8}$$

Where the position, x, of each atom at time is moved $v_n \Delta t$ from its previous position with each step. v_n is the velocity of the atom at time n and Δt is the time step. At the same time, the velocity is also updated from its previous value by $a_n \Delta t$, where a_n is the acceleration of the atom at the previous step. The initial positions are set at the start of the simulation from the starting structure. The velocities are also set at the beginning of the simulation, and through stages of gradual heating, will take on the Maxwell-Boltzmann distribution. Meanwhile, the acceleration is calculated at each step, based upon the force-fields chosen at the beginning of the simulation. The acceleration is calculated using Newton's second law:

$$a = \frac{d^2x}{dt^2} = \frac{1}{m}F \tag{2.9}$$

where

$$F = -\nabla V(x) \tag{2.10}$$

The value F is the force felt by each atom, as determined from the gradient of the potential, described by the force fields.

The accuracy of the Euler method is limited and can lead to instabilities, so other methods are often used. One such method is the leap-frog method. Here, the position and velocity are not updated simultaneously, but "leap-frog" over one another, with each being updated a half-step after the other.

$$v_{n+\frac{1}{2}} = v_{n-\frac{1}{2}} + a_n\Delta t \tag{2.11}$$

$$x_{n+1} = x_n + v_n\Delta t \tag{2.12}$$

This technique is more stable than the Euler method and is used in molecular dynamics engines such as AMBER. The Euler method provides stable simulations provided Δt is sufficiently small. The value for Δt is the time-scale of the smallest time-scale dynamic, or the highest frequency oscillation. In atom MD simulations Δt is conventionally set to 2fs [95].

2.5 Solvent Models

Now that the basic algorithms are set, the question remains as to how to make the simulations realistic. Just as the molecules need to be in a solute in real life, the same physics applies in simulations. The purpose of the solvent then, is to accurately model the molecule's environment, if observed in the laboratory. This includes modelling the bulk liquid the biomolecule would be in.

There are two different models for doing this; explicit and implicit solvent. As the names suggests, in explicit solvents the water and ions are explicitly expressed with each atom or molecule as its own particle. The reverse then is true for the implicit solvent, where instead of being explicitly defined, the solvent is incorporated within the calculations of the simulations. There are advantages and disadvantages to both solvents, with the type of solvent being chosen depending on the aim of the simulation. With explicit solvent, the solute-solvent interactions are modelled more accurately, however it is at the expense of time and dynamics. Conversely, by incorporating the solvent implicitly within the simulation calculations, the viscosity of the solvent can be changed. This means that for larger dynamics, or for faster sampling, implicit solvents are preferred.

2.5.1 Explicit Solvent

Biomolecules need to be modelled in a solvent in order to match real environments. The most commonly used model for water molecules is the TIP3P model [96], which models the oxygen and two bound hydrogen atoms as three points, each with partial charges on the points. These introduce a more accurate environment, however the molecules themselves are often of little interest. As a result, there are more atoms and calculations, creating an additional computational cost, while the number of atoms of interest stays the same. The solution then is to reduce the number of calculations while keeping the solvent effects.

The first method is to use an efficient volume or "solvent box", which reduces the number of superfluous molecules. In an ideal box, the boundaries would all be the minimum distance from the solute, forming a sphere. However, as will become apparent later, the box needs to be able to tessellate, which a sphere does not.

Although a cube tessellates, it does not have an efficient use of volume, with the diagonals of the cube being further away from the solute than the distance required for the sides. The more optimum shape is a truncated octahedron, a cube with the edges cut off, which is often used in AMBER [97]. It satisfies both the criterion of tessellation while providing a more efficient use of space. Thereby reducing the number of molecules needed, and thus the number of calculations.

If the solvent was modelled only using the water molecules in the initial truncated octahedral box then, firstly, the solute could quite quickly move outside of the box and interact with the vacuum. Secondly, should it stay inside the box, there is still a large proportion of water molecules at the boundaries. These molecules would then be interacting only with a fraction of other molecules, as on one side there would still be the vacuum [98]. This is not physical, and does not accurately reflect a bulk solvent.

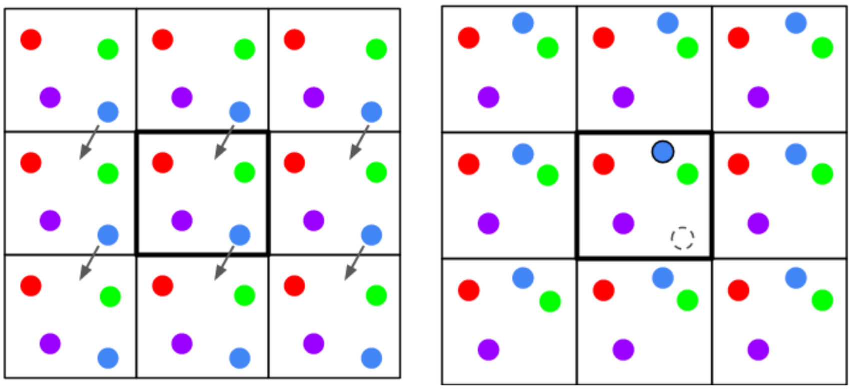

Figure 2.4: First layer of repeating periodic cells in 2D. The blue atom in the primary cell moves out of the bottom boundary, but with the periodic cells it now appears at the top. *Primary cell in thick black lines, replica cells in thin black lines, atoms are shown by coloured circles*

To get over the problem of the small systems having such a large proportion interacting with a boundary, periodic boundary conditions are introduced. Here, the unit cell is repeated in three dimensions as shown in figure 2.4. The calculations only being solved for the primary cell, making the dynamics in all the cells identical. If a particle moves in the primary cell, it does so in all the other periodic cells [99, 98]. The result is a solution to the problems of boundary effects. Firstly, there are now atoms outside the boundaries of the cell, so that the atoms here are no longer interacting with a vacuum. Secondly, the boundaries themselves are now removed, as a particle reaching a boundary can pass into adjacent cell. To maintain the number of atoms in the system, and because all the cells are identical, an atom moving out of the bottom boundary, would now appear at the top of each cell. In the same way, if a particle reached the top of the cell, it would pass into the bottom of the one below, and thus the bottom of every cell. Doing this preserves the number of particles in the system, as even when a particle moves outside of the cell boundaries, it is still contained within the system.

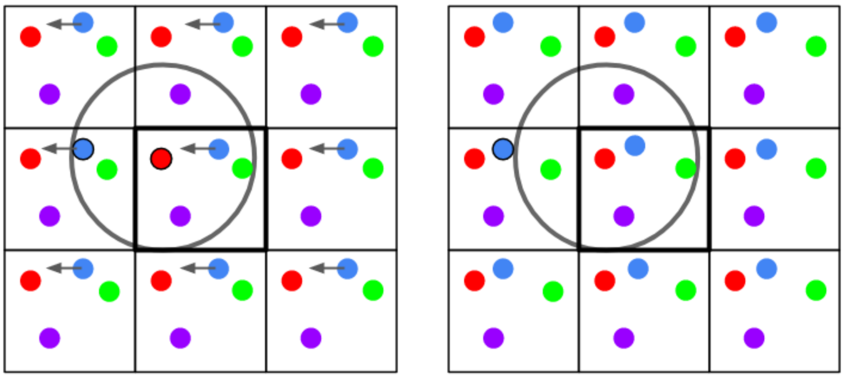

Figure 2.5: The problems with using a cut-off. The blue atom on the left would be included in the calculations for the red atom in one step, but would then be absent in the next. *Grey circle represents the cut-off area*

Another way of reducing the computational cost is using a cut-off. When calculating the forces on an atom, only those within the cut-off would be taken into account. This reduces the number of calculations and can speed up simulations. The standard cut-off distance is 12.0Å [97] and it works well for certain long-range potentials like the Lennard-Jones. As can be seen in figure 2.1, the Lennard-Jones potential decays to zero much faster than the Coulomb potential. At distances past the cut-off the lack of inclusion of Lennard-Jones does not significantly affect the calculations. It does however make a difference with the Coulomb potential [98]. The cut-off can result in the Coulomb potential not being appropriately included, as atoms can have an effect in one-step but be absent the next, as shown in figure 2.5. The solution to this is to use particle Ewald mesh calculations, which smooth out the Coulomb potential so that the contribution of atoms outside the cut-off are effectively included.

These Ewald calculations add a screening cloud, a Gaussian of opposite charge to each point charge. However, this now means that the electrostatic contribution would be of the screened charges, not of the point charges. To amend this, compensating charges are added, these are Gaussians of the same charge as the particle. The result cancels out the screened charges, while creating a periodic function that can be represented by a Fourier series. The Fourier series smooths out the Coulomb potential and speeds up the calculations [99].

2.5.2 Implicit Solvent

In implicit solvent, the explicit solvent molecules are removed, and are replaced by the dielectric constant of the bulk solvent, including the water and ions. Due to the removal of the collisions between the molecules, the implicit solvent has a lower viscosity, and as a result can sample conformational space much faster [100]. This increases the computational efficiency at the cost of the solvent's structural information.

The implicit solvent is incorporated into MD simulations by finding the free energy of solvation, which is then be integrated into the force-fields [101]. The free energy of solvation is the amount of energy required to introduce the solute in a vacuum to a solvent. It can be calculated using the Generalised Born equation [102], which begins:

$$G = -\frac{1}{2}\int \rho(r)\phi(r)dr \tag{2.13}$$

Where ρ is the charge density, ϕ is the potential, and r is the distance from the centre of a charged object, such as an atom.

With this equation, one can calculate the free energy of a charged sphere, such as a monatomic ion, with a radius α and charge q. To begin with, the entirety of the charge will be distributed across the surface of the sphere as it repels from itself [103]. The charge density is then the charge divided by the surface area of the sphere:

$$\rho(s) = \frac{q}{4\pi\alpha^2} \tag{2.14}$$

The potential ϕ can then be calculated using the Coulomb potential as described in equation 2.15 [87, 88].

$$\phi(r) = -\frac{q}{\epsilon|r|} \tag{2.15}$$

Where ϕ is the potential, q is the charge, ϵ the dielectric constant of the environment and r the distance from the centre of the sphere.

Thus, equations 2.14 and 2.15 can be fed back into equation 2.13, becoming equation 2.16 in order to calculate the free energy of the charged sphere.

$$G = -\frac{1}{2}\int_s \frac{q}{4\pi\alpha^2}\frac{q}{\epsilon\alpha}ds \tag{2.16}$$

After integrating over the surface of the sphere, equation 2.16 then becomes equation 2.17

$$G = -\frac{1}{2}\frac{q^2}{\epsilon\alpha} \tag{2.17}$$

The energy of solvation can then be determined as the difference between the free energy of the molecule in its gas phase where ϵ=1 and the free energy in a solvent with a dielectric constant of ϵ. The resulting equation is described by equation 2.18, and is called the Born equation.

$$G_{ele} = G(\epsilon = 1) - G(\epsilon) = \frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\frac{q^2}{\alpha} \tag{2.18}$$

The Born equation describes the solvation energy for a monotomic ion, however most biomolecules are made of much more than a single ion. The Born equation can thus be adapted for a molecule made up of many atoms, as described by the Generalised Born Model, equation 2.19 [104]. The equation sums up the free energy of each atom in the molecule, as well as incorporating the effects that the atoms have on each other, with a screening function f_{GB}.

$$G_{ele} = -\frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{f_{GB}} \tag{2.19}$$

where

$$f_{GB} = \sqrt{r_{ij}^2 + \alpha_i \alpha_j e^{-D_{ij}}} \tag{2.20}$$

and

$$D_{ij} = \frac{r_{ij}^2}{d\alpha_i\alpha_j} \tag{2.21}$$

Where N is the total number of atoms, and r_{ij} is the distance between atom i and atom j, with charges q_i and q_j and Born radius α_i and α_j, respectively, as well as a constant d.

The Born radius of each atom is not a property of the atom itself but is calculated based upon the molecular geometry. Calculating the Born radii can be timeconsuming, and so methods have been developed to determine approximations for the Born radius, in order to reduced the computational cost. One such model is the GBneck2 model, which is currently the most accurate [105].

2.6 Thermostats and Barostats

2.6.1 Langevin Thermostat

In MD simulations, the boundary conditions create a microcanonical (NVE) ensemble, where the number of particles, volume and energy are conserved. The number of particles cannot change, since every time a particle leaves the "box", it emerges from the other side. The box size is set when the molecules of interest are solvated, and it remains unchanged, therefore the volume also remains unchanged as well. Finally, the energy remains constant since it does not exchange energy outside of the system. The total energy remains constant with the kinetic energy and potential energy fluctuating, converting from one to the other, while the total remains constant.

In physiological conditions, the temperature is constant, with the kinetic energy of the system based upon the temperature as in equation 2.22.

$$E_k = \frac{3}{2} N k_B T \tag{2.22}$$

Where k_B is the Boltzman constant.

In MD simulations, the kinetic energy is proportional to the atomic velocities squared. Thus equation 2.22 is reversed with the temperature at each time step based upon the total kinetic energy of each atom, as in equation 2.23.

$$T = \frac{1}{3} \sum_{i=1}^{N} \frac{m_i v_i^2}{N k_B} \tag{2.23}$$

The temperature can then be maintained by coupling the system to a heat bath at T_0. The coupling is achieved using a Langevin equation which adapts the equations of motion by adding in a frictional and stochastic term, as in equation 2.24 [106].

$$m_i a_i = F_i - m_i \gamma_i v_i + R_i(t) \tag{2.24}$$

While the first term remains as before, the force felt by the atom, the second term is a damping term that corresponds to collisions with particles. The last term, $R(t)$, provides an additional random force, being a stochastic variable with a Gaussian distribution with a mean of zero. The intensity of $R(t)$ is defined by equation 2.25

$$< R_i(t) R_j(t + \Delta t) >= 2 m_i \gamma_i k_B T_0 \delta(\Delta t) \delta_{ij} \tag{2.25}$$

The Langevin equation acts as a thermostat by maintaining the kinetic energy for a set temperature. It also introduces viscosity into implicit solvent simulations with the γ providing the frequency for collisions. When $\gamma < 2$ the frequency of collisions is small and the system can explore the conformational space further.

2.6.2 Berendsen Thermostat and Barostat

In the Berendsen thermostat, the temperature at each step is maintained close to the set temperature, T_0 of the thermal bath, by scaling the velocities [106]. The velocity of each atom, v_i, is scaled by a factor λ, defined in equation 2.26. Scaling the velocities allows the sum of the total velocities to be minimised, which also constrains the kinetic energy.

$$\lambda = 1 + \frac{\Delta T}{2\tau_T} \left(\frac{T_0}{T} - 1 \right) \tag{2.26}$$

where ΔT is the difference between the temperature at each step, T, and the set temperature of the thermal bath T_0. The value τ_T is the time constant of the coupling.

In terms of the algorithm, the velocities are determined first using equation 2.11, and then scaled using equation 2.27.

$$v_{n+\frac{1}{2}} \to \lambda v_{n+\frac{1}{2}} \tag{2.27}$$

The value for λ was can be used from the previous step, as the value does not change significantly over time [106].

The pressure also needs to be maintained in order to match conditions in a cell or in the laboratory. It is assumed that the pressure is 1 atm[97], and can be maintained using a similar algorithm, the Berendsen barostat. Here, it is the distances between particles that are scaled, instead of the velocities, by a factor μ, as described in equation 2.28 [106].

$$\mu = 1 - \frac{\beta \Delta T}{3\tau_P}(P_0 - P) \tag{2.28}$$

The distances are scaled by adjusting the particles' positions after the velocities have been scaled, using equation 2.12. The positions can then be scaled to maintain the pressure using equation 2.29.

$$x_{n+1} \to \mu x_{n+1} \tag{2.29}$$

2.7 Constraints and Restraints

The fastest frequency oscillation in atomic MD simulations is that of the covalent bonds between hydrogen and another atom [95]. As the bond lengths are not of great concern, the bond lengths can be constrained to a fixed length using the

SHAKE algorithm [95, 107]. The absence of the fastest oscillation allows the time step to be increased, so that the simulation evolves faster. The SHAKE algorithm is applied in a similar way to the Berendsen thermostats and barostats, by adjusting the positions as part of the Leapfrog algorithm similar to equations 2.26 and 2.28 [106].

It may also be beneficial to fix the positions of atoms, such as for replica exchange molecular dynamics. One way of achieving this is via positional restraints, which restrain the atoms' position according to a reference structure. A harmonic potential is used to maintain the position of the atoms, so that they conform to the reference structure. The potential is applied to each of the restrained atoms, so that when the atoms move away from their position, an opposing force is applied which pushes the atom back towards its desired position. The strength of the harmonic potential can be increased or decreased whether or not the atoms need to be highly restrained or have larger deviations [108].

2.8 Simulation Parameters

The MD simulations in this thesis was carried out using AMBER versions 16 [109] and 18 [97], using both explicit and implicit solvent. The implicit solvent simulations used a Generalised Born model with GBneck2 corrections [105]. A Langevin thermostat was used with a collision frequency of 0.01ps^{-1} or 1ps^{-1} in order to reduce the viscosity and sample more of the conformational space. The higher value of the two, 1ps^{-1}, was used to reduce the speed of denaturing. All of the simulations ran with a salt concentration of 0.15M unless they required a low or high concentration, which was set at 0.01M and 0.5M respectively. The explicit solvent simulations were solvated with TIP3P water [96], and Na and Cl ions [110], with the temperature maintained using a Berendsen thermostat. Both the explicit and implicit simulations used the ff14SB [89, 90] and BSC1 force fields [94, 111] for the protein and DNA respectively.

2.9 Enhanced sampling techniques

For complex systems, such as proteins and DNA, MD simulations can feasibly sample behaviour in the order of nanoseconds. For dynamics on a time-scale longer than nanoseconds, there would be a high computational cost, requiring considerable resources such as larger computer clusters, and for a greater time. In order to get around this problem, a number of different enhanced sampling techniques were developed, which extend the effective simulation time at a much lower computational cost [112].

The reason why some conformations are rarer, and thus take more time to sample, is because they are separated by large energetic barriers. The MD conformations can get stuck in local minima, unable to access the conformations [113, 6], as in figure 2.6. The enhanced techniques were to developed to allow more of the conformational space to be sampled, within the same simulation time, by overcoming the energetic barriers.

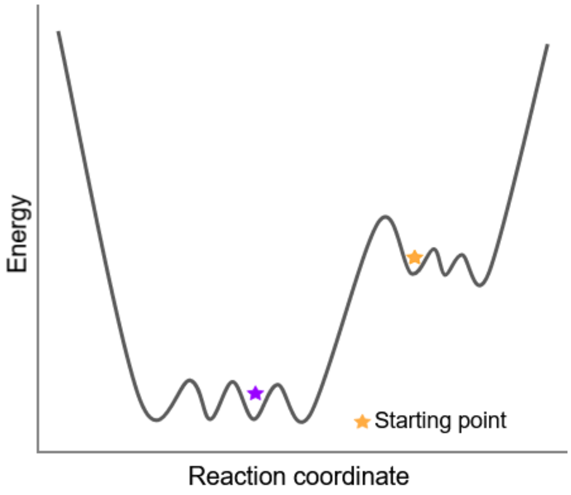

Figure 2.6: Example of an energy landscape. If a state was at the point of the gold star, then it may not reach the conformations at the point of the purple star, due to the large energy barrier between them. *Figure based upon figure 1 Corbett et al 2021 [6]*

2.9.1 Replica Exchange Molecular Dynamics

One such technique is replica exchange molecular dynamics (REMD) [114, 115]. This method relies on the principle that when a system, such as a biomolecule is at two close temperatures, there is an overlap in the conformations that are accessible. For example, take two identical proteins, one at temperature T_1, the other at at a slightly higher temperature T_2. Since there is an overlap in the temperatures, there would be an overlap in the energies of both proteins (fig 2.7). As a result, there would be conformations that could be found in both T_1 and T_2.

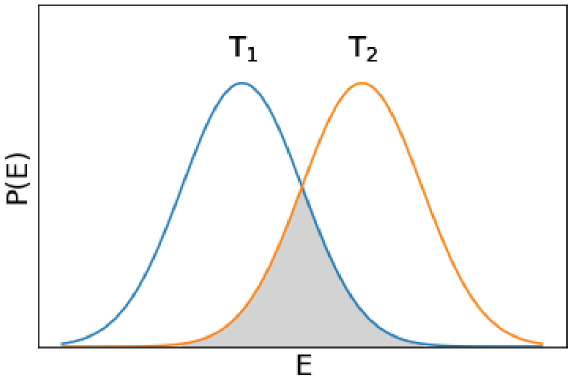

Figure 2.7: One system at two temperatures T_1 and T_2 have a distribution of energies, and thus conformations, which can be found in both T_1 and T_2 (shaded grey area).

REMD simulations then run a number of identical MD simulations, called replicas, simultaneously with each one at a different temperature. The temperatures are exchanged between the replicas so that the lower temperature replicas can sample more of the conformational landscape. Exchanges between two replicas are attempted periodically, with exchanges accepted or rejected based upon how close the temperatures and energies are. The probability that an exchange occurs is based upon equation 2.30 [115].

$$P(exchange) = exp\left[\left(\frac{1}{k_BT_i} - \frac{1}{k_BT_j}\right)(E(q_m) - E(q_n)\right] \tag{2.30}$$

Where P, is the probability of exchange, k_B, the Boltzmann constant, T the temperature of each system, i and j, E the energy and q the coordinate vectors of each replica, here replica m and n.

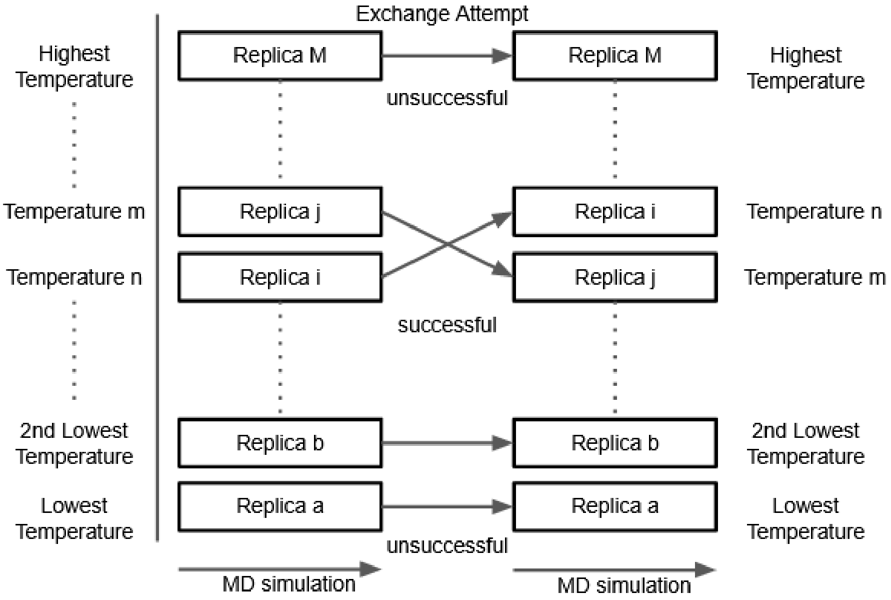

Figure 2.8: Figure demonstrating Replica Exchange Molecular Dynamics technique.

If the exchange is rejected then the MD simulation carries on unchanged until it reaches the next exchange attempt, as seen in figure 2.8. On the other hand, if the exchange is accepted then the temperatures are exchanged between the two replicas. The kinetic energy and velocities are then rescaled to reflect the new temperature as in equation 2.31 [116].

$$v_{new} = V_{old}\sqrt{T_{new}/T_{old}} \tag{2.31}$$

The benefit of REMD simulations is that it can sample inaccessible conformations in lower temperatures, by overcoming the energetic barriers. In lower temperature replicas, the conformational landscape can be quite rugged, compared with the higher temperature replicas which have much flatter conformation landscapes, due to the extra thermal energy. When a low temperature replica has exchanged with one of high temperature, the MD simulation can move through the conformations, unobstructed by the barrier found in the low temperature replica. When the replica moves down to the lower temperature, the replica can now access the conformations on the other side of the barrier. Shown in figure 2.9.

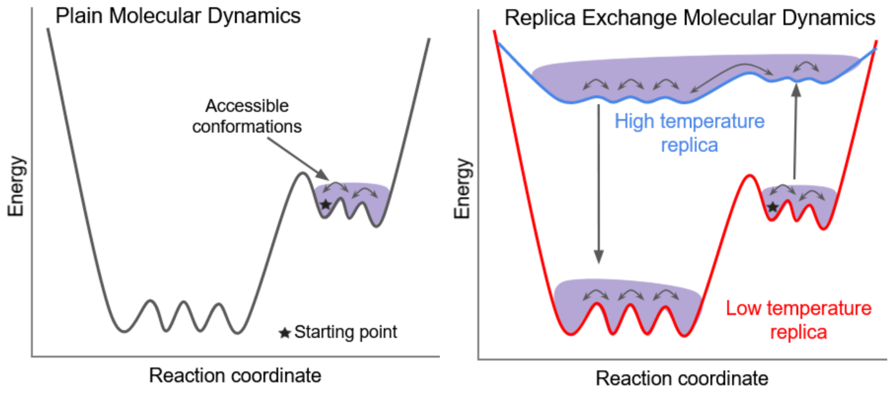

Figure 2.9: Difference in accessible conformations in plain MD simulations compared to replica exchange MD simulations. *Figure based upon figure 1 Corbett et al. 2021 [6]*

In REMD simulations, the choice of the replicas' temperatures is clearly important. To start with, a greater temperature range allows more conformational sampling as the higher temperature replicas have more energy. The more energy the system has, the more it can overcome energetic barriers and therefore sample more conformations, not found in lower energy or temperature systems. There is also the choice of the temperature difference between the replicas, as the temperature difference affects the rate of exchanges between the replicas. The smaller the temperature difference, the more likely a temperature exchange will be accepted. However, this increases the number of replicas needed in order to reach the higher temperatures. The more replicas required, the greater the computational cost.

The choice of temperature range, temperature differences, and number of replicas is a balance to get the optimum conformational sampling for the smallest computational cost.

The number of replicas is also dependent on the degrees of freedom of the system. The larger the system, the more degrees of freedom it has, and so the greater the number of replicas required. REMD can be used with larger molecules provided only a small number of atoms are of interest. In this case, the rest of the molecule can be fixed with positional restraints to reduce the degrees of freedom [117, 118].

2.10 Analysis of Simulations

There are a variety of different techniques for analysing MD simulations. All of the following methods were to analyse the Rep and BPV E1 helicase simulations conducted for this thesis. These, along with many, many others, can be found in the AMBER manuals [97]. Although they were not used in this thesis, there are also many other programs that used, whether in the browser or downloaded from the internet. CURVES+ [119] and Prody [120] were two such programmes that were used for initial research but did not make it into the final thesis.

2.10.1 Hydrogen bonds

Hydrogen bonds are important in biomolecular processes. In DNA, it is hydrogen bonds that keep the base pairs, and therefore two DNA strands, together. Measuring the number of hydrogen bonds between a base pair can determine whether the two strands are together, partially separated or fully separated. In a G-C pair, the strands are together when there are 3 hydrogen bonds between the pair, 1 or 2 when the bases are coming apart, and 0 when the bases have separated. Hydrogen bonds not only exist in DNA but between DNA and proteins. Here, the hydrogen bonds or salt bridges show whether or not an amino acid in the protein is interacting with the DNA.

A hydrogen bond forms between a hydrogen, which is covalently bonded to an electronegative atom, such as nitrogen or oxygen, and another electronegative atom. In MD simulations, the hydrogen bonds can be found by specifying the heavy atoms involved and the geometric cut-off between these atoms. For example, the atoms involved in the hydrogen bonds between guanine and cytosine for instance are, O6 and N4, N1 and N3, and N2 and O2. The geometric cutoffs are the distances between the heavy atoms, and angle between the them and the hydrogen, with the distance 3.5Å and angle as 120°. A hydrogen bond had formed if the chosen heavy atoms were separated by 3.5Å or less, or had an angle of 120° or more.

Salt bridges are important interactions involving hydrogen bonds. There are two definitions of salt bridge, one where two like charges form hydrogen bonds with an ion, and the other where hydrogen bonds form between two opposite charges, which is what occurs here. In this case, a salt bridge is a type of interaction between a negatively and positively charged atom, which can include hydrogen bonds. An example of a salt bridge would be an oxygen, on the phosphate backbone of DNA, forming a bond with the nitrogen of the side-chain amine of lysine.

Salt bridges can be determined by using the same distance cut-off of 3.5Å but removing the angle cut-off, as salt bridges can have any angle between them. There can be multiple hydrogen bonds in a salt bridge, depending on the number of hydrogens bonded to the heavy atom, making them stronger than single atom hydrogen bonds. The hydrogen bonds and salt bridges are determined each frame using the analysis tool of AMBER, cpptraj [121].

2.10.2 Potential Energy calculation

The potential energy of a biomolecule can be used to measure its energetic state. If a biomolecule is sufficiently equilibrated then the positions, bonds and angles of all the atoms will result in a lower potential energy, as determined by the force-fields. However, should the atoms not be in an equilibrated position, with atoms too close to one another, then the potential energy will be higher. In AMBER, the potential energy can be measured using a programme called *esander*, as part of the *cpptraj* package [121].

2.10.3 DNA conformations

DNA has a number of physical parameters that can be measured in order to quantify the change in its structure. Two such parameters are the twist, the angle between one base pair and the next, and the minor groove width, the distance between the two DNA backbones in the minor groove. In MD simulations these can be measured using a programme called CURVES+ [119]. The distribution of the DNA parameters over the simulation can then be displayed, in order to see how they are different in different states. The frames where there are 3 or 0 hydrogen bonds between the last base pair can also be removed to see the differences with respect to the separation of the last base pair.

The ssDNA in the E1 collar domain chamber was stretched between the point of separation and the channel in the motor domain. The stretching was measured by taking the distance between the C5′ atom on the last dsDNA nucleotide on the active strand, and the C5′ on the third ssDNA nucleotide.

2.10.4 Protein conformations

A few methods were used to quantify the global conformations of Rep. The first was the root-mean-squared deviation (rmsd). The rmsd is a measure of how far each amino acid has moved from its starting position [122]. Measuring it required the rmsd function in *cpptraj* [121]:

$$rmsd = \sqrt{\frac{1}{N}\sum x_i(t)^2 - x_0^2} \tag{2.32}$$

where N is the number of atoms in the amino acid, x_0 the starting position of each atom, and $x_i(t)$ the position of each atom in each frame.

The radius of gyration was used to measure the compactness of Rep. Like before it used *cpptraj* with the radgyr analysis function.

$$R_g = \sqrt{\frac{1}{N}\sum r_i^2} \tag{2.33}$$

where N is the number of atoms and r_i the distance from the atom to the centre of mass (COM) of the molecule.

The final observable was the angle by which subdomain 2B rotated around 2A. This could be measured by defining it as a dihedral angle between the centre of mass (COM) of each domain as well as the hinge. So the angle was between the COM of 2A (residue 388), the hinge (residues 374 to 373) and COM of 2B (569). The residues closest to the the COM of each subdomain were calculated using a python script written by Dr George Watson [123]. While the dihedral was measured using the *cpptraj* command of the same name.

Chapter 3

Plain Molecular Dynamics Simulations of the E1 helicase

3.1 Introduction

The crystallographic structure, obtained by Prof Fred Antson and Dr Vladimir Levdikov, located the DNA inside the E1 helicase, but certain areas lacked resolution. The DNA that could be resolved well, were the entrance and exit of the DNA. Entering the E1 helicase from a side channel, the dsDNA passes into the helicase and is separated above the collar domain (CD) as seen in figure 3.1. One ssDNA strand is sterically excluded by the CD and exits via another side channel. Meanwhile, the other strand passes through the CD, into a chamber, and then on through the hairpin-lined channel in AAA+ domain.

Often the most interesting parts of the protein are the ones that move, which has the unfortunate effect of being the most difficult to see by crystallography. In this case, the two ssDNA residues after the DNA fork, produced low electronic density and low resolution in this area, due to having conformational flexibility. As a result, there was little information as to the behaviour of the DNA in the

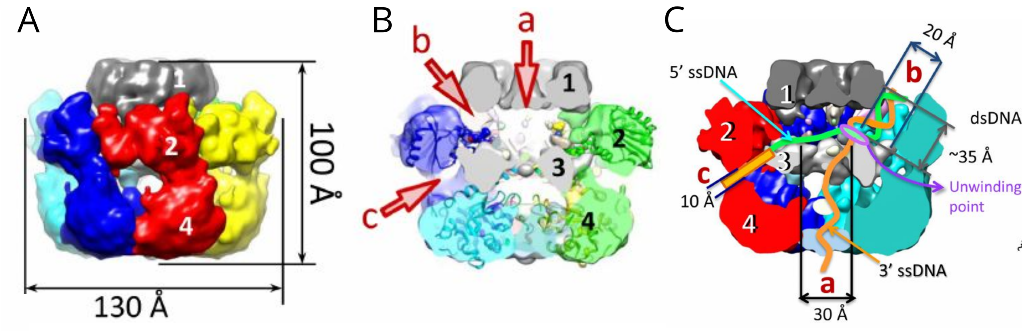

Figure 3.1: The full structure of the BPV E1 helicase (A) with a cross-section showing three side channels a, b and c (B) where the DNA enters and exits (C). *1 - The N-terminal domain, 2 - DNA binding domain, 3 - Collar Domain 4 - AAA+ domain. Figures from figure 3 of Chaban et al 2015 [3]*

chamber. Molecular dynamic simulations could solve this problem, by modelling the likely behaviour of the molecule, the crystallographic structure, shown in figure 3.2.

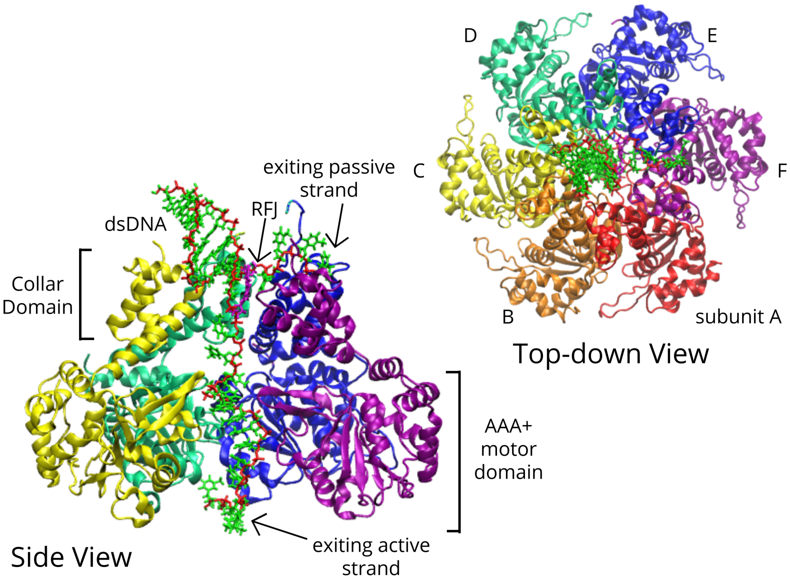

Figure 3.2: The starting structure for the MD simulations of the BPV E1 helicase domain and DNA *Structure from Prof Fred Antson and Dr Vladimir Levdikov.*

The aim of the simulations was to provide information on the DNA in the chamber. Preliminary simulations showed that the two ssDNA nucleotides could take a variety of conformations, so it was theorised that the chamber could act as an adapter for the DNA. It could provide the space for the DNA to transition from ten base pairs a turn in dsDNA, to six nucleotides a turn when the ssDNA passes through the chamber. To test this hypothesis, the rotary wave was modelled to see how the DNA might adapt in each rotary state.

3.2 Methods

3.2.1 Modelling the E1 states

Whether in scientific endeavours or normal life, it is important for the tool to match the problem. It would be quite absurd to paint a wall with a small watercolour brush, and likewise the appropriate method for simulating the E1 helicase needs to be found. In an ideal situation, the whole rotary wave could be simulated from the first step, starting with the hydrolysis of ATP, which enacts the conformational changes in the AAA+ domain, and translocating the DNA. However, with the current technology this would require a variety of advanced modelling techniques. A simpler approach is to physically manipulate the protein, and recreate the AAA+ states from each rotary wave step. This is much simpler and less computationally expensive than advanced techniques, while achieving a similar result.

The method of creating the new structures was to rotate the AAA+ domain with the ssDNA inside. This would make it appear to the CD and dsDNA as if a conformational wave had occurred (see figure 3.3). A simulation was run for each state, each created from the original crystallographic structure (State 1). The AAA+ subunits, ssDNA in the chamber, and the ADP were rotated and reattached to the unrotated CD, dsDNA and ssDNA in the chamber, resulting in the name "cut and stick" method, as described in figure 3.4. The "Frankenstein structure" was then minimised, solvated, and equilibrated, followed by a 50ns production run.

The internal energy of the DNA in the chamber was checked using cpptraj -

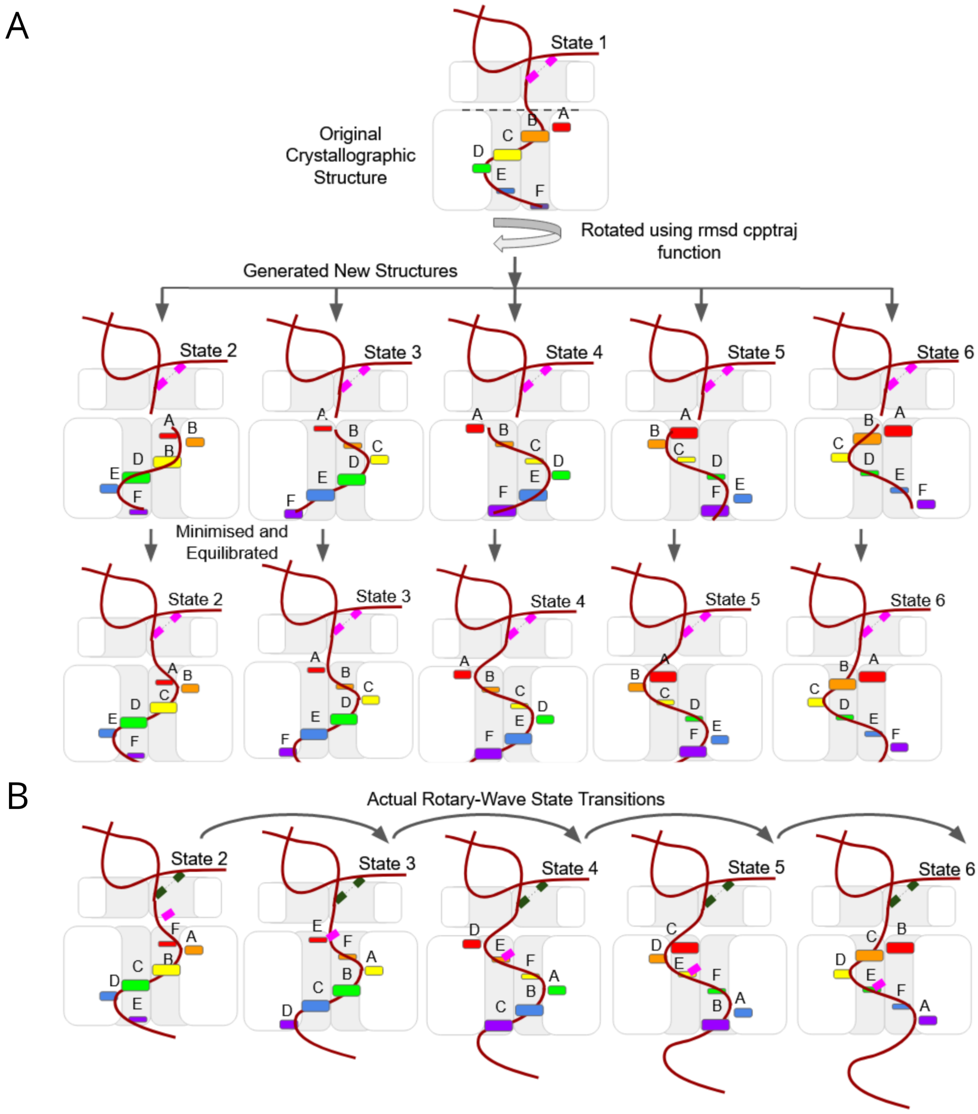
A
State 1
Original Crystallographic Structure
Rotated using rmsd cpptraj function
Generated New Structures
State 2
State 3
State 4
State 5
State 6
Minimised and Equilibrated
B
Actual Rotary-Wave State Transitions
A
B
C
D
E
F

Figure 3.3: (A) "Cut and stick method" of structure generation of the different states of the rotary wave of the AAA+ domain. State 1, from the original structure, is rotated six times. The original collar domain, dsDNA and two ssDNA nucleotides in the chamber are combined with the rotated AAA+ domains, ssDNA inside and ADP. These structures are then minimised and equilibrated, to allow for a reasonable rejoin between the DNA and protein parts, from which MD simulations then start. (B) The structures of the states as they would be in reality, and which the structures in A are modelling.

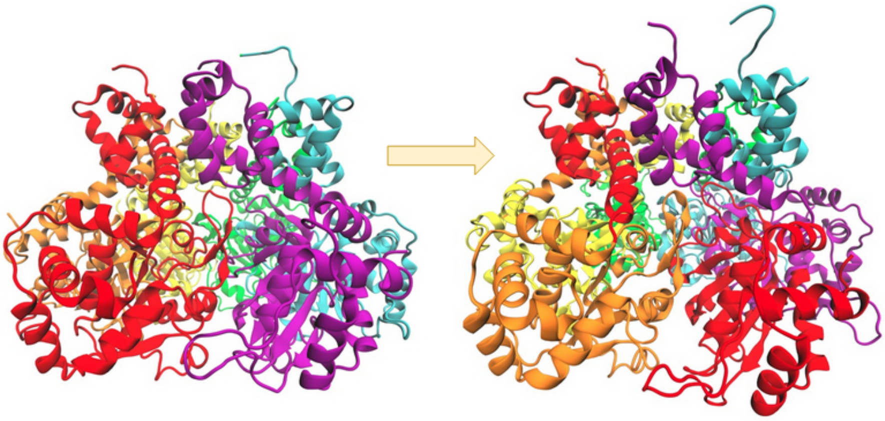

Figure 3.4: Left. Original crystallographic structure and state 1 of the BPV helicase domain. Right. First rotated state, state 2, where the subunits in the motor domain have been moved around the collar domain by one turn. *Red - subunit A, orange - subunit B, yellow - subunit C, green - subunit D, blue - subunit E, purple - subunit F.*

esander from AMBER. This was to ensure that the rotations were not increasing the energy of the DNA, and pushing it into higher energy states. For completeness, this process was repeated four times, making four different sets of simulations, each with six states (as shown in Table 1 in the Appendix). Each set had a different base for the second ssDNA nucleotide, the next one to be grabbed, in order to see if the simulations would be possible no matter the base.

3.2.2 Structure preparation

Before the E1 states could be made, the crystallographic structure from the Antson Lab needed to be prepared for MD simulations. Firstly, the crystallographic structure did not have the bound hydrogen atoms, so the correct protonation state of each amino acid was determined using the Virginia Tech H++ web application [83]. Since the protonation states are dependent on the environment, the environmental conditions were chosen to match cellular conditions, where the E1 helicase would be found, so a pH of 7.0 and a salinity of 0.15M. The lysines and histidines, both found in the DNA-binding hairpins, were found to be fully protonated in the lysines, and epsilon protonated in the histidines.

After protonation, the new structures were generated as described in the method before, and each was solvated with explicit solvent in a 15Å octahedral periodic box. The water was modelled using TIP3P water [96], along with 0.15M of sodium and chlorine ions, which also neutralised the protein and DNA. The resultant topology was made using the ff14SB force field for the helicase [89], the parmbsc1 for DNA [94], dangions for the ions [110] and AMBER16 ADP [124]and MG force fields [125] all in AMBER16.

3.2.3 Molecular Dynamics

In order for the simulations to be run from a stable energetic state, the system was minimised to situate it in a local energy minima. The solvent was minimised first for 10,000 steps, keeping the E1 helicase and DNA restrained with 50kcal/mol positional restraints. Then the helicase and DNA was minimised along with the solvent for another 10,000 steps. The system was gradually heated up over 80ps to 300K, with positional restraints on the helicase and DNA which were slowly reduced. Finally, there was 50ns of production run in the NPT ensemble with a Berendsen thermostat and Berendsen coupling barostat to maintain the temperature and pressure respectively. The simulations converged rapidly, with only very small deviations from the starting structures (around 3Å), as can be seen in Appendix figure A1.

3.2.4 DNA naming

For the ease of analysis, the DNA nucleotides used from the crystallographic structure were named using a system that denoted their position and strand as demonstrated in figure 3.5. The dsDNA nucleotides were labelled ds1 to ds8, where ds1 is the last dsDNA base pair, and the pair just before the point of separation. Those on the active strand were then labelled as1 to as14, with as1 the first ssDNA of the newly separated strands. The same was for those on the passive strand, which only had four nucleotides, ps1 to ps4, corresponding to as1 to as4. The first two ssDNA nucleotides of the active strand, as1 and as2, were located inside the large chamber, with nucleotides as3 to as8 bound to their corresponding hairpins in the motor domain. Finally, the nucleotides as9 to as14 exited the helicase.

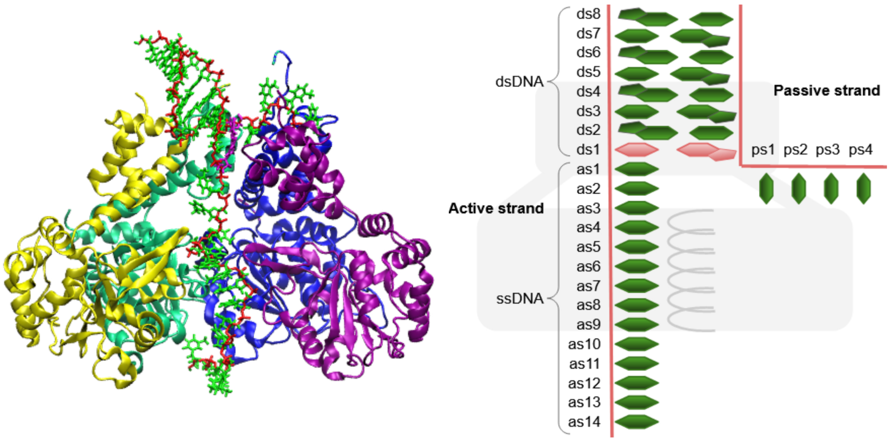

Figure 3.5: Left. The crystallographic structure of the E1 helicase with the DNA inside obtained by Prof Fred Antson and Dr Vladimir Levdikov. Right. The labelling of the nucleotides. The sequence is GCGCGC in the dsDNA and only T in the ssDNA, with the exception of as2 which was changed to each DNA base.

3.2.5 Base Replacement at as2

To see that the "cut and stick" method would work for any base, four sets of simulations were created. Each set had a different base at as2, since the as2 nucleotide was the most likely to lead to instabilities with the "cut and stick" method. The as2 nucleotide was the closest nucleotide to the hairpins, and therefore had the highest chance of clashes with other atoms. The as2 nucleotide is also of particular interest as it was found to move more than the other ssDNA and had the greatest variability in conformations. If the "cut and stick" method was a success, then changing the base at as2 could lead to simulations that would give insight into whether the hairpins grab new nucleotides differently depending on the base.

The as2 base was changed from thymine in the original structure to adenine, guanine and cytosine using the AMBER topology builder subprogram, tLeap. The atoms in thymine that were not found in the other bases were removed, with tLeap adding the necessary missing atoms.

3.3 Results

3.3.1 Energy of the Rejoin

The newly created states could produce high potential energy in the connecting regions. Firstly, there was the possibility that the adjustments needed to reconnect the ssDNA in the collar to those in the channel, could produce covalent bonds that were too long or short, or had energetically unfavourable angles within the DNA. Secondly, as the DNA is forced into different conformations in the chamber, it could present conformations that were particularly close to the chamber walls or the hairpins. Consequently this might again result in high energies that would not be physical or stable. If the energies in the rotated states closely resemble and sufficiently overlap with those of state 1, for that as2-nucleotide, such as the mean being within 3σ of the state 1 mean, then it is likely that the observed conformations are physical.

The energy in the chamber was calculated to see whether the rotated structures (state 2 to 6) were reasonable. As can be seen in figure 3.6, the energies are within 1 to 2σ of the energy of state 1 for each base replacement in all but one case. The only exception is state 5 in as2-adenine, which is within 3σ. The energies demonstrate the validity of the "cut and stick" method, as none were significantly high.

Looking at the results in more detail, the energies are different depending on which base is at as2. This makes intuitive sense since each base has different atomic arrangements, which would effect its internal energy. The lowest energies

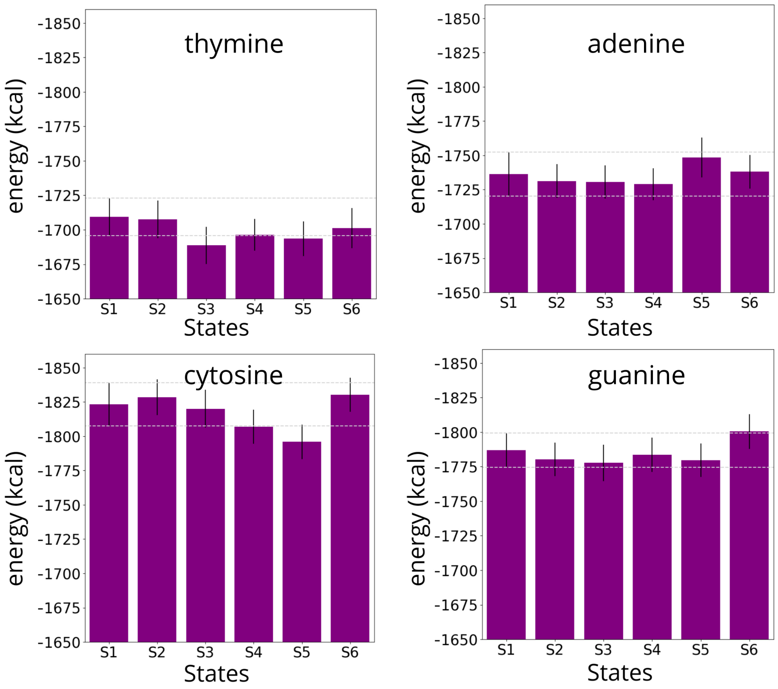

Figure 3.6: Average internal energy of the DNA nucleotides around the rejoin between as2 and s1 (d7 to s1) and top 3 hairpins over the first 50ns from all molecular dynamics simulations with as2 mutations. The grey lines are ± 1σ from the mean of state 1.

are seen where thymine occupies as2, as in the original structure. This is expected, since it was the only base to be found as it was in the original structure, so it has had more time to find its lowest energy state compared to the others.

Overall, the energies show that the model is plausible. It can therefore be concluded that the method of rotating the structure to model a change in rotary state, does not push the DNA into unrealistic energies.

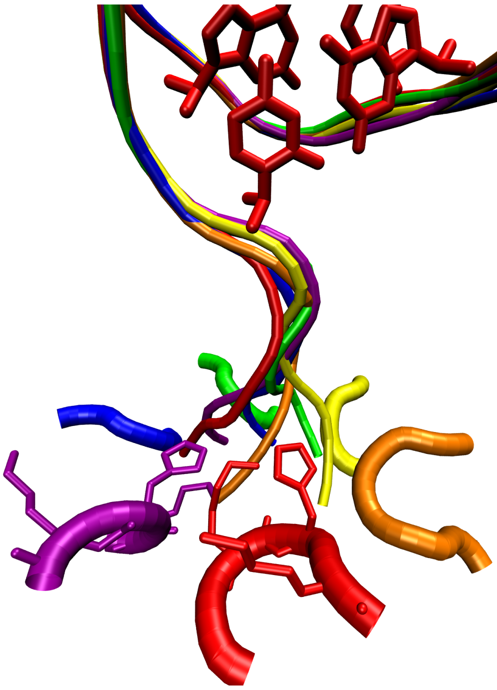

Figure 3.7: DNA backbone and top hairpin of each state as found at the start of the plain MD simulations. There is sufficient space within the E1 helicase chamber (between the CD where the dsDNA resides and the motor channel where the hairpins are located) to allow for many different backbone conformations. *State 1 - red, state 2 - orange, state 3 - yellow, state 4 - green, state 5 - blue, state 6 - purple. Last two dsDNA base pairs shown in red. K506, H507 and K508 shown for subunit A and F.*

3.3.2 DNA Conformations inside the Chamber

The hypothesis of the CD chamber as an adaptor was accepted and deemed to be true, as the space in the CD chamber did allow each state generated by the "cut and stick" or "Frankenstein" method to rejoin to the as3 nucleotide, with an energetically reasonable structure. Not only this, but the rejoined ssDNA backbones demonstrated that many conformations could be allowed in the chamber, as shown in figure 3.7. These conformations bridged the 10 base pair per turn dsDNA and 6 base pair per turn ssDNA conformations in a number of ways. Ultimately proving that the CD could certainly act like a "DNA adapter", with the DNA in the chamber taking a variety of conformations.

3.3.3 Maintained Hairpin Interactions

Besides energy, another check that our model was reasonable was whether the interactions between the translocating hairpins and the ssDNA were maintained. In the E1 structure, all but the top hairpin have bound a ssDNA nucleotide. Each hairpin binds the ssDNA using the first lysine (K506) and the histidine (H507), with the histidine binding to the adjacent nucleotide, and the lysine to the one below [39]. So hairpin 2 binds to as3 with histidine and as4 with its first lysine, hairpin 3 to as4 with histidine and as4 with its first lysine, and so on. Since the method for modelling the different rotary states rotates a part of the ssDNA and causes it to adjust its conformation, it is possible that the bonds between the hairpins could be disrupted.

The model structure was shown to be reasonable, as the interactions between the hairpins and DNA remained mostly unchanged in the rotated states, in comparison to the experimentally observed interactions, and the simulation for state 1. As can be seen in figures 3.8, 3.9, 3.10, and 3.11, the interactions between hairpins 3 to 6 are, in most cases, maintained throughout the whole simulation for the rotated states and state 1. The only DNA-hairpin interaction of these four hairpins not to be maintained 100% of the time is the last hairpin interaction, between K506 of hairpin 6 and as8. However, this can be expected, since it would be this hairpin that would next disengage from the DNA and move up the channel to collect a new nucleotide. This interaction is also likely to be weaker than the others, since the nucleotide is only held by one lysine, in contrast to the others which are held by both a lysine and histidine.

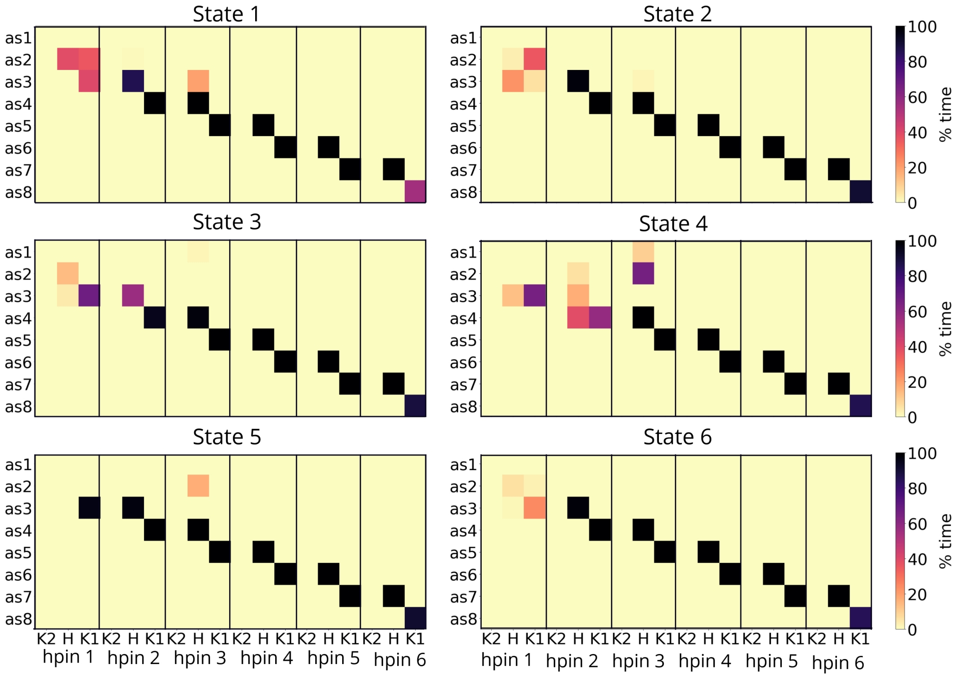

Figure 3.8: Heatmap of the hydrogen bonds of each hairpin for each state with the original base of thymine as as2. Hydrogen bonds are between the phosphate of as1 to as8 with the amino group of K506 and K508, labelled K1 and K2 respectively, and between the main chain amine and side chain epsilon2 of the histidine (H). DNA-hairpin interactions are maintained from as4 to as8, with most as3 interactions maintained.

3.3.4 New hairpin interactions

While the interactions between hairpins 3 to 6, and ssDNA nucleotides as4 to as8, would be the same across the different states, those between the hairpins close to chamber and the ssDNA inside, would be different. Since in each state new hairpin positions and different ssDNA conformations would create new DNA-hairpin interactions. In more robust simulations (see chapter 4), these new interactions would be the first ones between the protein and the DNA, and could provide

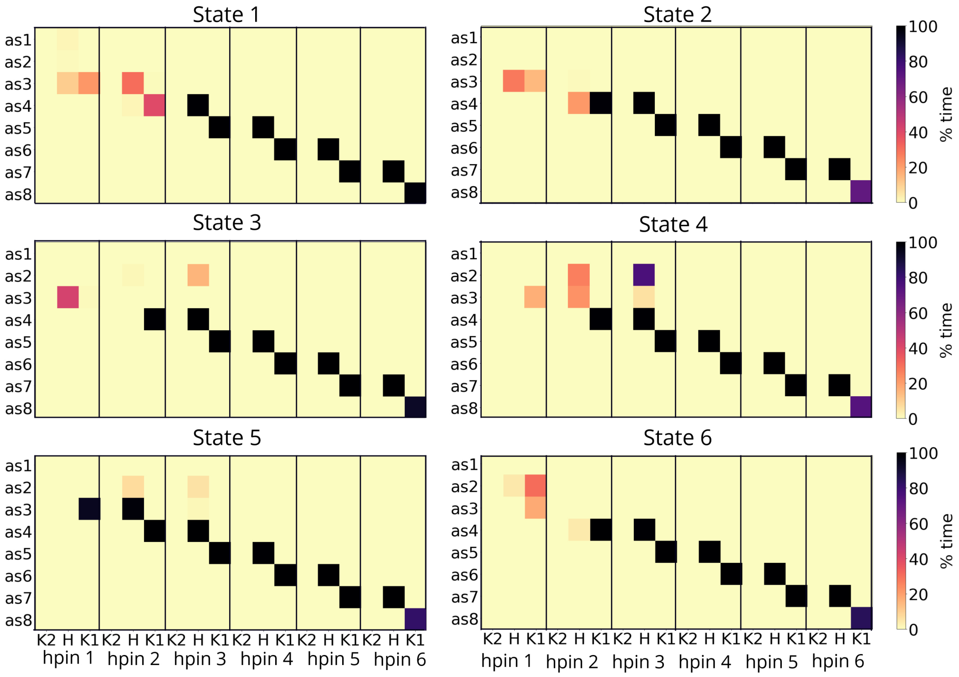

Figure 3.9: Heatmap of the hydrogen bonds of each hairpin for each state with adenine as as2. Hydrogen bonds are between the phosphate of as1 to as8 with the amino group of K506 and K508, labelled K1 and K2 respectively, and between the main chain amine and side chain epsilon2 of the histidine (H). DNA-hairpin interactions are maintained from as4 to as8, with most as3 interactions maintained.

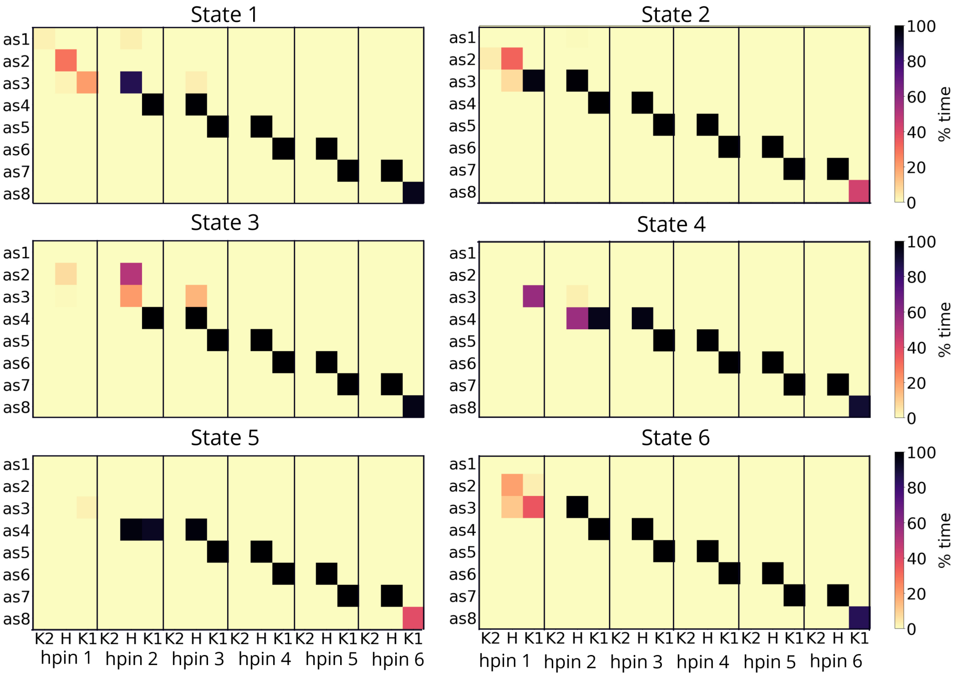

Figure 3.10: Heatmap of the hydrogen bonds of each hairpin for each state with cytosine as as2. Hydrogen bonds are between the phosphate of as1 to as8 with the amino group of K506 and K508, labelled K1 and K2 respectively, and between the main chain amine and side chain epsilon2 of the histidine (H). DNA-hairpin interactions are maintained from as4 to as8, with most as3 interactions maintained.

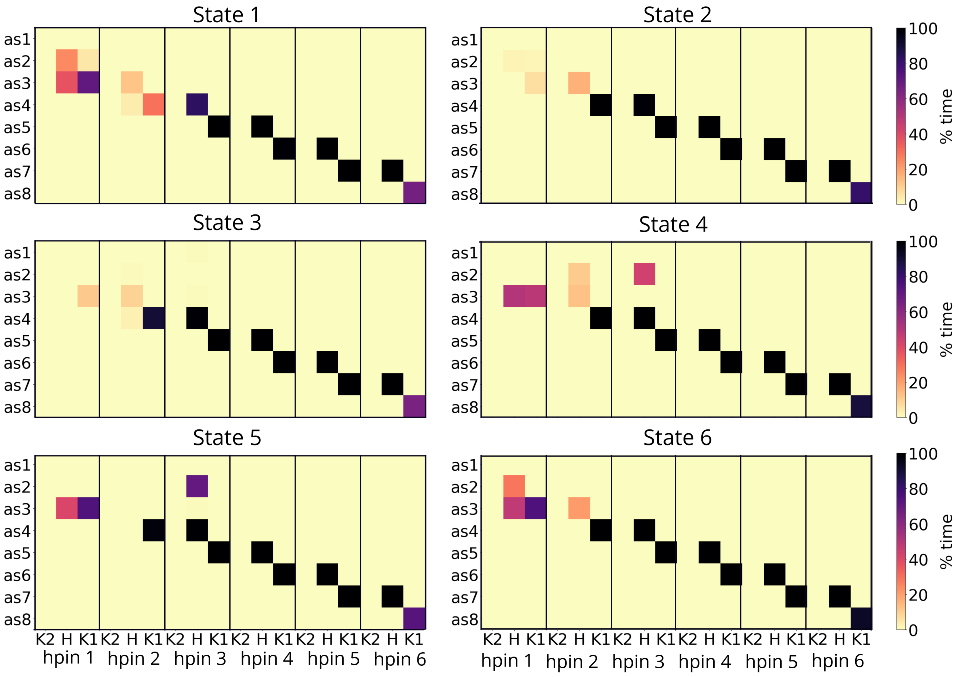

Figure 3.11: Heatmap of the hydrogen bonds of each hairpin for each state with guanine as as2. Hydrogen bonds are between the phosphate of as1 to as8 with the side chain amino group of K506 and K508, labelled K1 and K2 respectively, and between the main chain amine and side chain epsilon2 of the histidine (H). DNA-hairpin interactions are maintained from as4 to as8, with most as3 interactions maintained.

insight into how the hairpins grab the ssDNA.

In line with the experimental structure, the histidine, H507, of hairpin 1 should bind to as2 (h1-H507:as2), while the lysine, K506, binds to as3 (h1-K506:as3). These interactions were observed in state 1, with the h1-K506:as3 interaction present in all four as2 bases. Meanwhile, the interaction between h1-H507:as2 was present in three out of the four simulations in state 1, with it missing only in as2-adenine. The next state to have the most interactions was state 4, with all simulations having the h1-K506:as3 interaction but not h1-H507:as2. The same was seen in state 5, but with no interactions from hairpin 1 seen in as2-cytosine. Other states had a mix of both h1-H507:as2 and h1-K506:as3, just h1-K506:as3 or neither.

As for hairpin 2, it interacted with the DNA far more frequently compared to hairpin 1, having interactions between the H507 and as3 (h2-H507:as3), and K506 and as4 (h2-K506:as4). Presumably because, while as3 and as4 have their position maintained through stacking interactions with the bases below, as well as being constrained in the motor domain channel, as2, which hairpin 1 binds to, is much freer to move. This would make it more difficult to obtain the necessary orientation for a hydrogen bond. As with hairpin 1, the translocating interactions were observed most often in state 1, where all as2-simulations had both h2-H507:as3 and h2-K506:as4 interactions. In all the other states, the h2-H507:as3 interaction was missed in at least one simulation, with the h2-H507:as3 interaction appearing over fewer frames of the simulations even in state 1. When the h2-H507:as3 was absent (or in as2-thymine state 4, weaker) the H507 histidine would occasionally interact with as4, rather than its corresponding nucleotide as3.

There were no interactions between K508 (ie K2) and the ssDNA. However this was expected as K508 has not been shown to interact with the ssDNA in the chamber experimentally. The analysis included K508 because is reasonably close to the ssDNA so that, should it undergo a conformational change, it could move closer to the ssDNA in order to interact with it. Although this was not observed in these simulations it does not rule out the possibility that it could interact with the

DNA. If the simulations were run for longer, in a less viscous solvent such as an implicit solvent, or with a form of accelerated MD, then the possibility still exists that K508 could be involved in hairpin grabbing interactions.

3.3.5 DNA Separation

In addition to testing the validity of the "cut and stick" method, the MD simulations showed separation of the last base pair. Some of the simulations showed DNA separation where the hydrogen bonds between the last base pair of the dsDNA were broken. Although there was little to no DNA separation in the original structure, as2-thymine, nor adenine simulations, DNA separation was observed in as2-cytosine and as2-guanine. In as2-cytosine there were DNA separation events in state 4, and a few occurrences in state 3. Many more were seen in as2-guanine, with separation occurring in state 2, state 4 and state 6. No correlation was found between the separation of the DNA, the hairpin interactions, or the the states.

3.4 Discussion

Since the simulations were set up to test the validity of the "cut and stick" method they were relatively short and thus had limited sampling. As a result the hairpin interactions and DNA separation events can not be used to accurately describe the mechanisms of the helicase. Besides, no pattern was found in the DNA hairpin interactions, and there was no correlation between the hairpin interactions and the DNA separation. However, the fact that base pair separation was observed in these simulations indicates that DNA separation could occur from thermal motion, as described by a passive mechanism.

All in all, it would seem that the energetic landscape of the DNA and the protein is rough, and that these simulations were stuck in local minima. Thus it would require increased sampling to overcome these barriers, and sample the conformations necessary to observe a mechanism of DNA grabbing.

3.5 Conclusion

The "cut and stick" method produced energetically reasonable structures in all five of the rotated states. These states also maintained the hairpin-DNA interactions from the crystallographic structure, with many of the simulations also forming new translocating interactions between the top hairpins and the DNA. As a result, the plain MD simulations of the E1 helicase showed that the "cut and stick" method was reasonable, and could be used to investigate the mechanism of the E1 helicase.

The validity of the structures also proved the hypothesis of the collar domain being an "adapter". The hypothesis was that the collar domain provided the space for the ssDNA to adapt its conformation from 10 base pair per turn dsDNA, to the 6 nucleotides per turn ssDNA in the AAA+ domain, in each of the states. The simulations did observe this, as each of the five rotated states formed different energetically reasonable structures. The different states also showed different conformations in the chamber with all bases able to fit and adjust into energetically favourable states.

These simulations revealed that the hydrogen bonds between the last base pair could be broken from thermal motion. However, they were not accurate enough to say whether this is certainly the case. Further simulations using advanced sampling techniques will be required in order to overcome the energetic barriers between the conformations of the rough energetic landscape, and sample the dynamics of the E1 helicase DNA separation and hairpin grabbing mechanisms.

Chapter 4

Replica Exchange Simulations of the E1 Helicase

4.1 Introduction

The plain MD simulations demonstrated that the rotary wave could be modelled using a "cut and stick" method, and that these simulations produced base pair separation. However, the plain MD simulations were limited by low sampling. Thus the simulations were not accurate enough to see how the hairpins grab the DNA, or how the strands are separated. So, in order to sample more of the conformational landscape, and provide more accurate dynamics, a form of accelerated molecular dynamics would be needed, such as replica exchange molecular dynamics (REMD).

REMD simulations are able to sample more conformations by using thermal energy to overcome energetic barriers. As the name suggests, the method works upon the exchange of information between multiple replicas running in parrallel, each at a different temperature. The temperature difference is small enough that there would be an overlap between the conformations of one simulation and another.

The temperatures of two simulations could then be exchanged, allowing lower temperatures to sample conformations found at higher temperatures.

The temperatures can only be exchanged if there is sufficient overlap between the energies of both simulations. The energies, and conformations, have a greater chance of overlap if the degrees of freedom are small. As a result, REMD works best with only a small number of atoms. The atoms in the solvent are included in the degrees of freedom, making implicit solvent a better choice. In implicit solvent, the water molecules and salt are calculated through a continuum approach, without explicitly stated atoms. As a result, implicit solvent not only reduces the degrees of freedom, but also the number of calculations, speeding up the simulations, as well as the rate of the sampling. As for the atoms in the protein and DNA, the dynamics of DNA separation and hairpin grabbing only include a small number of atoms, those of the top three hairpins and DNA at the point of separation. Therefore, the positions of the atoms outside the area of interest can be fixed, with only those of interest free to move, in order to reduce the degrees of freedom [117, 118].

REMD simulations of the DNA fork and hairpins could answer questions about the helicase mechanism of the BPV E1 helicase. Questions such as whether the dsDNA separation is caused purely by thermal energy or an active process by the helicase? And by what mechanism do the hairpins grab the free nucleotides in the chamber?

4.2 Method

4.2.1 Simulation setup

Although the solvation method was different in the REMD simulations, compared to the plain MD ones, the starting structure was the same. Both starting structures took the coordinates and atoms from the crystallographic structure, once it had been protonated using the H++ web server. The simulations also had the same salt concentration of 0.15M. However, the temperature in the REMD simulations was now maintained using a Langevin thermostat, with a collision frequency of 0.01ps^{-1}. The low frequency produced a lower viscosity in the implicit solvent, allowing more conformations to be sampled. As can be seen in Appendix figure A2, the REMD simulations converged, and did not move far from the beginning structure.

The residues of interest were those in the chamber, so only the dsDNA separation point and top hairpins were free to move, with all others positionally restrained. This would allow the DNA to separate at the last base pair, and allow conformational changes at the second to last, should they be needed for base pair separation. Only the top three hairpins were chosen to be free to move as only these three interacted with the ssDNA in the chamber. Although the lysine K506 and histidine H507 interact with the DNA, the adjacent K508 and R505 are involved in staircasing and could affect movement of the hairpin. The method of restraining part of the system in order to reduce the degrees of freedom, and thus improve exchanges.

4.2.2 REMD parameterisation

Small REMD simulations were conducted with different temperature ranges to find the best range needed for successful REMD simulations. The lowest temperature replica, was room-temperature, 300K, in order to match in vitro experiments. However, the highest temperature, as well as the difference in temperature between the replicas needed to be decided. The temperature difference would determine how regularly the replicas could be exchanged, and thus the breadth of conformations that the lower temperatures could sample. Meanwhile, the range of temperatures needed to be great enough for there to be sufficient energy to break the bonds between the last base pair. It was hoped that the temperature range would be sufficiently small, or the temperature differences sufficiently high, that no more that twelve replicas would be needed. The reason being, that twelve replicas would run efficiently on four GPUs, the number on a single node of JADE. More replicas would not only run slower, taking more time, but possibly require more GPUs, increasing the computational cost and wait time in order to get access to more GPUs.

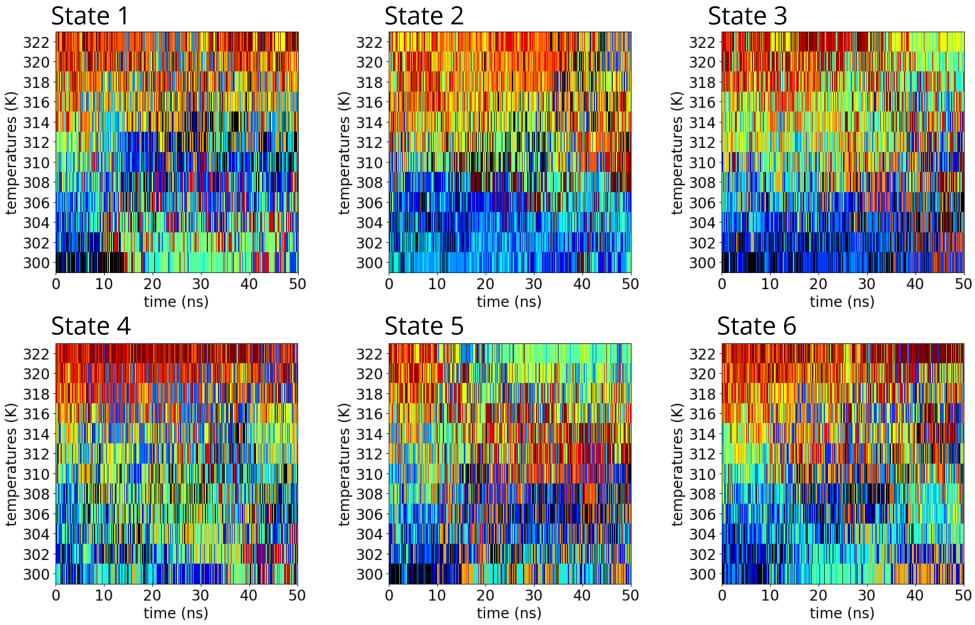

Figure 4.1: The exchanges between the twelve different replicas, ranging from T= 300K to T=322K for states 1 to 6. Each colour shows a different replica. *Black - replica 1, starting temperature of 300K, replicas starting at the cooler temperatures in blue, with the ones starting at higher temperatures in red.*

In order to determine the optimum parameters, several small REMD simulations were run at different final and replica temperatures. It was decided that a final temperature of 322K with a temperature difference of 2K between replicas would be optimum, as it gave a reasonable probability of exchange of 0.35 [126]. Meanwhile, the highest temperature of 322K enabled DNA base pair separation. As can be seen in figure 4.1, there were differences in the exchanges in the different states. A thorough exchange rate would show all replicas sampling all temperatures. So the replicas starting in the lower temperatures (blue) would be seen sampling the higher temperatures, while those replicas starting in higher temperatures (red) would also exchange with the lower temperatures. Most of the simulations did have thorough sampling of the different temperatures.

Looking closely at the row showing the replicas that exchanged temperatures with one at 300K, the most thorough sampling was with those for state 1, 5 and 6 having a variety of replicas sample 300K. Simulations for state 2 and 3 had the least thorough exchanges, with few exchanges between high and low temperatures. As can be seen later, these two states showed the least amount of base pair separation.

4.3 Results

The REMD simulations did just as was expected, with clear movement of the DNA and hairpins. The last base pair of the dsDNA had particularly exciting dynamics. Here, the base pairs broke apart, breaking some, or all, of the three hydrogen bonds that kept the last base pair together. This breaking sometimes left only one hydrogen bond, or none at all. In most cases, the two bases having separated, would come back together, before separating again, with many separation events throughout the simulation.

As can be seen in figure 4.2, the different states had different degrees of separation. State 2 and 3 showed the least base pair separation, with the bases staying connected throughout. State 5 showed the most separation with the base pairs completely separating. The base pairs in state 4 did separate, but stacked one on top of the other, resulting in hydrogen bonds forming vertically between the two. Finally, both state 1 and state 6 had separation events, the bases separating and rejoining over the simulations. In total, there were 10,817 events out of a possible 30,000, where there were no hydrogen bonds between the two bases.

Along with DNA separation, there was also grabbing of the DNA by the hairpins. The hairpin grabbing events were correlated with the simulations that had

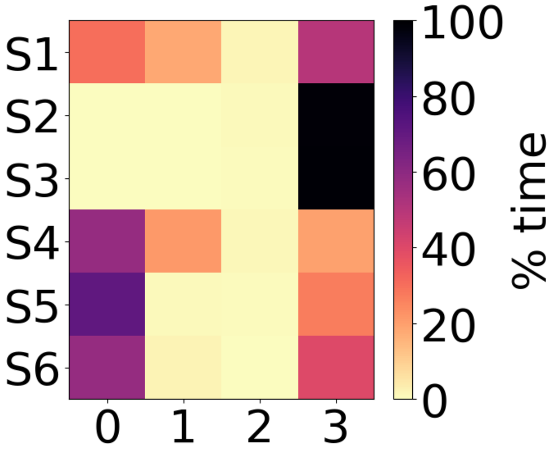

Figure 4.2: The percentage of time during the 50ns simulation that the last base pair (ds1) of each state had 0, 1, 2 or 3 hydrogen bonds between the base pair.

DNA separation events. Clearly, analysis of each state was required to see the mechanisms of each one, and how they might indicate an overall E1 helicase mechanism.

4.3.1 State 1

One brilliant side effect of molecular dynamic simulations is that you can visualise the results, rather than just relying on graphical analysis. The simulations showed that after the first couple of nanoseconds, the second lysine of the top hairpin, h1-K508, started to interact with the phosphate backbone of as2. The last base pair then broke apart when h1-K508 interacted with the DNA backbone of as2, and came back together when the lysine disengaged. It seemed that the h1-K508 grabbed the as2 phosphate, and pulled the ds1-ds1 nucleotides apart. However, as the timing of the events was very fast, it was unclear if the h1-K508 was grabbing the DNA and pulling it apart, or if the hairpin was grabbing at the same time or after the DNA had separated.

To see whether the DNA separation or hairpin pulling came first, the time after a h1-K508:as2 interaction and a base separation event was measured. As can

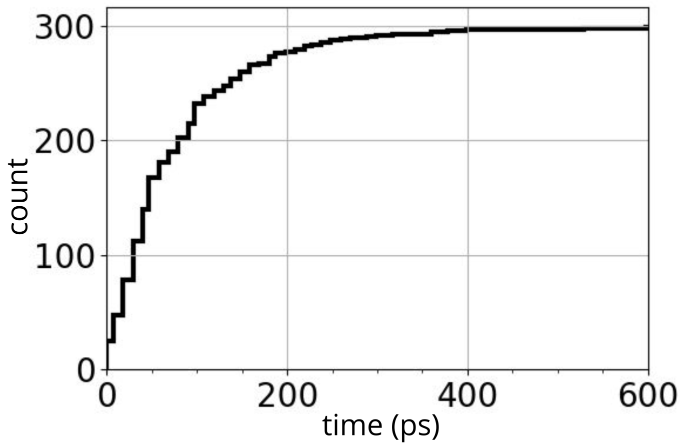

Figure 4.3: Cumulative histogram of the time between a salt bridge forming between the phosphate of as2 and the h1-K508 of state 1, and the last base pair going from three hydrogen bonds to zero, provided that the salt bridge did not break.

be seen in figure 4.3, of the 314 separation events, where the the last base pair went from having three to zero hydrogen bonds, 96% occurred within 400ps of an h1-K508:as2 interaction. The 4% that did not was during a 1ns stretch (t=12.580ns to t=13.550ns). Here, the h1-K508 disengaged from as2 20ps (2 frames) before the DNA separates, and did not re-engage, although the bases repeatedly came together and separated. Therefore, in the case of state 1, it seems that although the DNA can separate without an h1-K508:as2 interaction, in the majority of cases the interaction precedes and thus is the likely cause of the DNA separation.

4.3.2 State 2 and 3

Unlike state 1, state 2 showed very little dynamic behaviour and essentially no DNA separation events. However, in the last 15ns of the simulation h1-K508 started to engage with as2, as can be seen in figure 4.4. The behaviour appeared very similar to what was seen at the beginning of state 1. In both cases, the as2 phosphate moves towards the lysine end chain amino group as the amino group moves towards as2. This behaviour changed the conformation of the DNA as it rotated to have the phosphate orientated, so that the two protruding oxygens face the lysine. Since this behaviour preceded the DNA separation events in state 1, it is possible that should the simulation continue past 50ns, that DNA separation might be observed.

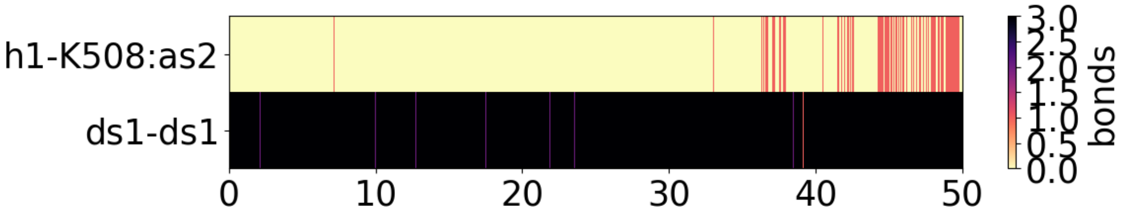

Figure 4.4: Interactions over the 50ns state 2 simulation between K508 of hairpin 1 (h1-K508) and the phosphate oxygens in as2 of the ssDNA (top). Along with the base pair interactions between ds1, the last dsDNA base pair (bottom). *The last base pair have a maximum of 3 hydrogen bonds (black), whereas the h1-K508 forms a salt-bridge (pink) or not (cream).*

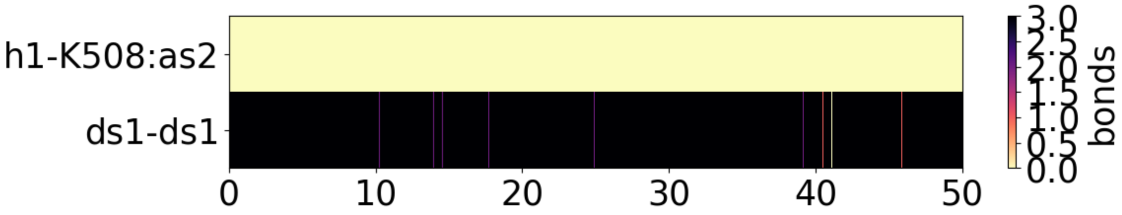

Figure 4.5: Interactions over the 50ns state 3 simulation between K508 of hairpin 1 (h1-K508) and the phosphate oxygens in as2 of the ssDNA (top). Along with the base pair interactions between ds1, the last dsDNA base pair (bottom). *The last base pair have a maximum of 3 hydrogen bonds (black), whereas the h1-K508 forms a salt-bridge (pink) or not (cream).*

State 3 had much more dynamic behaviour than state 2, with lots of movement of the top three hairpins and ssDNA in the chamber. However, this behaviour did not translate into DNA separation events. In fact, it had little to no separation of the DNA, and no h1-K508:as2 interactions, as can be seen in figure 4.5. The reason could lie in the positioning of h1-K508. The lysine seemed to be too far below as2 to form any interactions. The protruding oxygens from the as2 phosphate then never aligned with end amino group of h1-K508, as observed in states 1 and 2.

4.3.3 State 4

Compared to state 2 and 3, the DNA base pair separation in state 4 was dramatic, since there was separation of both the last and penultimate base pair. During the simulation the minor groove collapses, with the active strand moving towards the passive strand. The effect on the dsDNA is considerable, as it pushes out not just one, but two base pairs. Both of the last two base pairs are pushed out into the major groove, with the second to last base on the passive strand popping out completely, as can be seen in figure 4.6A. After separating, the last base pairs stack on top of each other, a conformation that is maintained through the majority of the simulation.

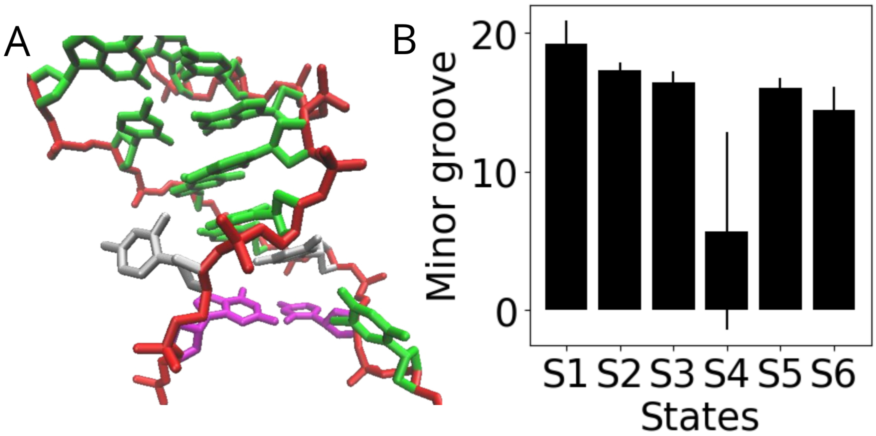

Figure 4.6: (A) The most common structure in state 4 showing the shortened distance between ds1 of the active strand and ds6 of the passive strand, and the bases opening into the major groove. (B) The average distance between ds1 of the active strand and ds6 of the passive strand (a measurement for minor groove distance) over the whole REMD simulation (T=300K) for each state. *DNA: backbone - red, bases - green, last base pair -pink, second to last base pair - white.*

The collapsing of the minor groove was significant compared to other states, as seen in 4.6B. It seems the minor groove collapse is the cause of the DNA separation for this state, unlike state 1, where the h1-K508:as2 interaction seemed to bring about the separation. However, the h1-K508 still interacts with the as2 phosphate and, like the states before, seems to pull the active strand away from the passive one, and towards the hairpin. When this occurs, the bases on both strands are not pulled further apart but with the stacked pair moving together. If the last bases had not stacked upon each other, it seems likely that the interaction of h1-K508 with as2 would have pulled the base pairs further apart.

4.3.4 State 5 and 6

In state 5, dynamics occurred which was not seen in the other simulations. Unlike the other states where the last two bases separated and came together many times, the last base pair completely separated, with the nucleotide moving further towards the hairpins. The cause of this was from the same DNA grabbing amino acid seen before in the other states, h1-K508. Here, the top hairpin K508, finally pulls the two strands away from each other. By pulling the nucleotide down towards the hairpins, it can no longer rejoin to its complementary base, and remains separated, as can be seen in figure 4.7 and 4.8A.

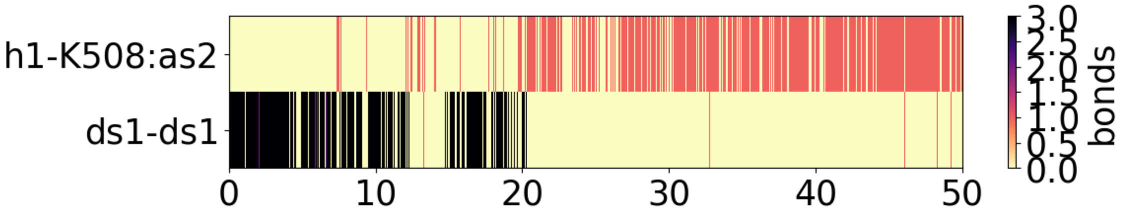

Figure 4.7: Interactions over the 50ns state 5 simulation between K508 of hairpin 1 (h1-K508) and the phosphate oxygens in as2 of the ssDNA (top). Along with the base pair interactions between ds1, the last dsDNA base pair (bottom). *The last base pair have a maximum of 3 hydrogen bonds (black), whereas the h1-K508 forms a salt-bridge (pink) or not (cream).*

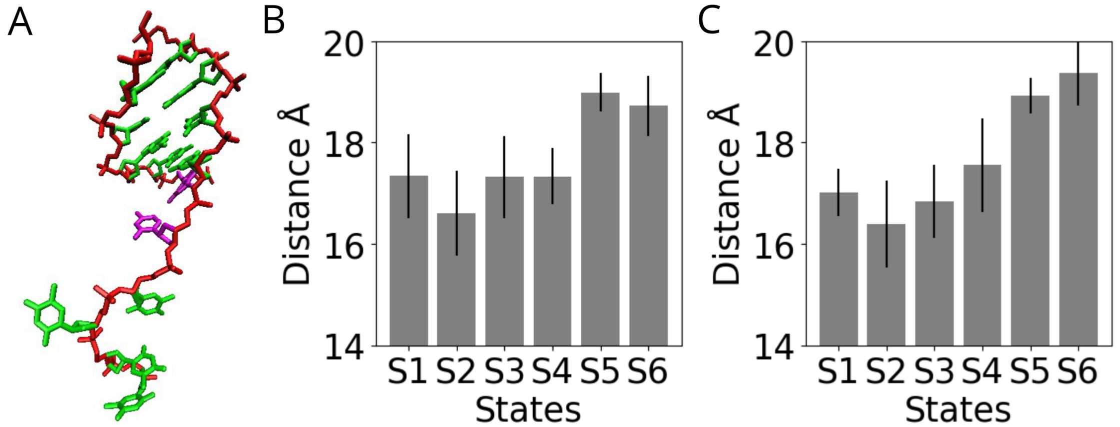

Figure 4.8: (A) The most common structure in state 5 showing the stretched distance between ds1 and as3 of the dsDNA. The average distance between ds1 and as3 for each state when there were (B) 3 hydrogen bonds between the last base pair and (C) less than 3 hydrogen bonds between the last base pair. *DNA: DNA backbone - red, bases - green, last base pair (ds1) - pink.*

In the final state, state 6, there were separation events between the last base pair. Unlike state 5, the h1-K508 did not grab as2 as the ssDNA seemed too far away for them to interact. As a result, the two strands were not pulled apart. As to why state 6 had DNA separation without h1-K508, the answer might lie in the DNA conformation. The DNA in the chamber was found to be stretched in state 6, as well as state 5, as seen in figure 4.8. Presumably, this made it easier for thermal motion to destabilise the base pair in state 6, and made it easier for h1-K508 to pull the strands apart in state 5.

4.3.5 Twist and DNA separation

As can be seen in figure 4.9, all of the states, except state 4, demonstrated a clear reduction in the twist of the last base pair, when the pair was separated. In all the states that showed separation of the last base pair (states 1, 4, 5 and 6), two peaks arose for twist of the last base pair. When the dsDNA is together, the twist is similar to the commonly found twist of 36°. While when the DNA is apart, the last

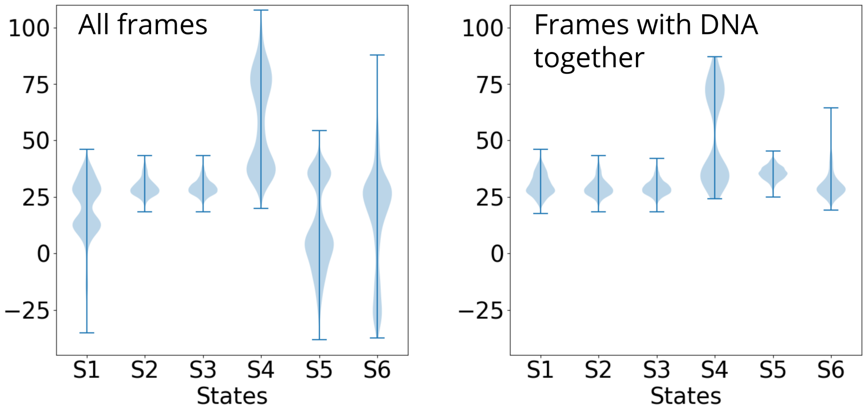

Figure 4.9: The distribution of twist of the last base pair over the 50ns REMD simulation for each state for all frames (Left) and only frames where the DNA was together with 3 hydrogen bonds (Right).

base pairs are undertwisted. State 4 shows slightly different results to the others, as both the last base pair and second to last had separation events. As a result, since the twist of one base pair is related to that of the next, the twist would be widely different should the second to last base pair have remained intact, as it was in all the other states.

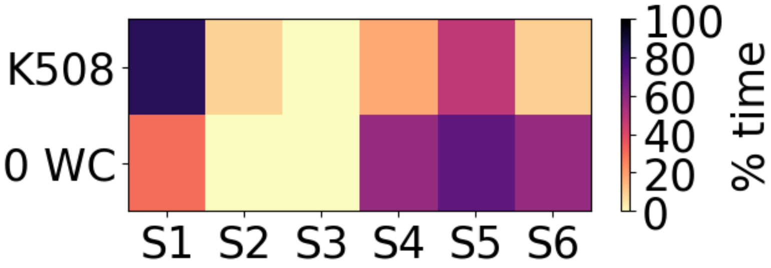

Figure 4.10: The percentage of time during the 50ns simulation that K508 formed a salt bridge with the phosphate of as2, and the percentage time the last base pair (ds1) of each state had 0 hydrogen bonds between the base pair.

4.3.6 K508 across states

It was clear that due to the different relative positions of the hairpins to the last base pair, each state had a different mechanism for DNA separation. However, the interaction of h1-K508 was consistent across the states. Figure 4.10 revealed that when the DNA had separated, K508 from hairpin 1 interacts with as2, such as in states 1, 4, 5 and 6. Likewise, when there is no separation, such as in state 3, there is also no interaction from h1-K508. State 2 is the only exception, as it has no separation, but still had interactions between h1-K508 and the DNA. However, as figure 4.4 showed previously, the interactions between h1-K508 and as2 occur in the last 15ns of the simulation. If the simulation was longer, perhaps DNA separation would be seen.

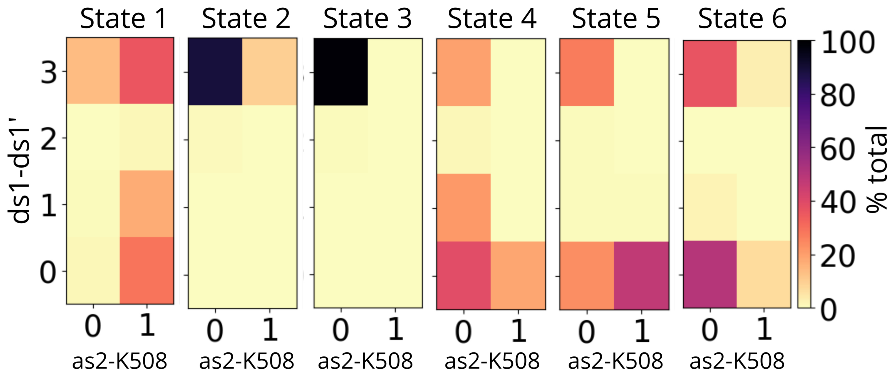

Figure 4.11: The percentage time each frame has 0, 1, 2 or 3 WC interactions of the last base pair, at the same time as having a salt bridge between as2 and K508.

The breakdown of the h1-K508:as2 interactions in conjunction with separation of the last base pair can be seen in figure 4.11. It seems clear that there are different variations in interactions of h1-K508 and the DNA. However, there is still an "on-off" mechanism between DNA separation and h1-K508. For states 4, 5 and 6 there is no h1-K508:as2 interaction when the last base pair is together. Only when the DNA separates is there an interaction.

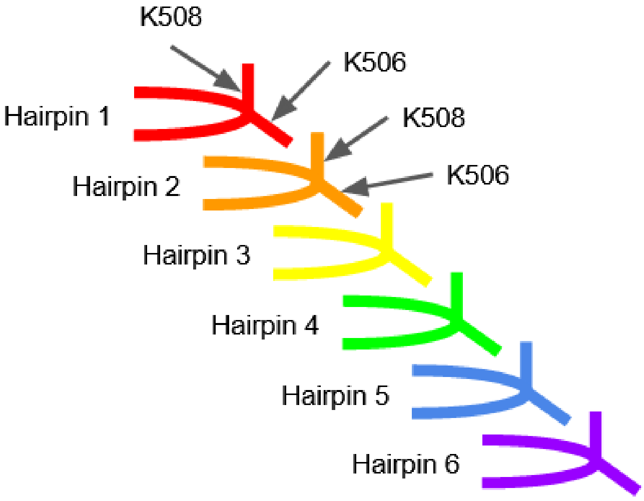

Figure 4.12: Representation of the stair-casing interactions between lysine K508 and lysine K506 between each hairpin. The end chain epsilon amino group of K506 forms a salt bridge with the main chain nitrogen of K508 from the hairpin below.

4.3.7 Hairpin stacking and K508

Previously, h1-K508 was only found to have interactions with the other hairpins through staircasing interactions. Here, the side chain amino group of K506 interacts with the main chain amino group of the K508 of the hairpin below, as represented in figure 4.12. As the interactions between K508 and DNA had not been observed before, it was unknown whether it worked alongside the stair-casing. As can be seen in figure 4.13, two states had interactions between the top hairpin K506 and second to top K508; states 1 and 4. So it seems that the interactions between h1-K508 and the DNA, do not destabilise the stair-casing interactions, but could possibly increase them.

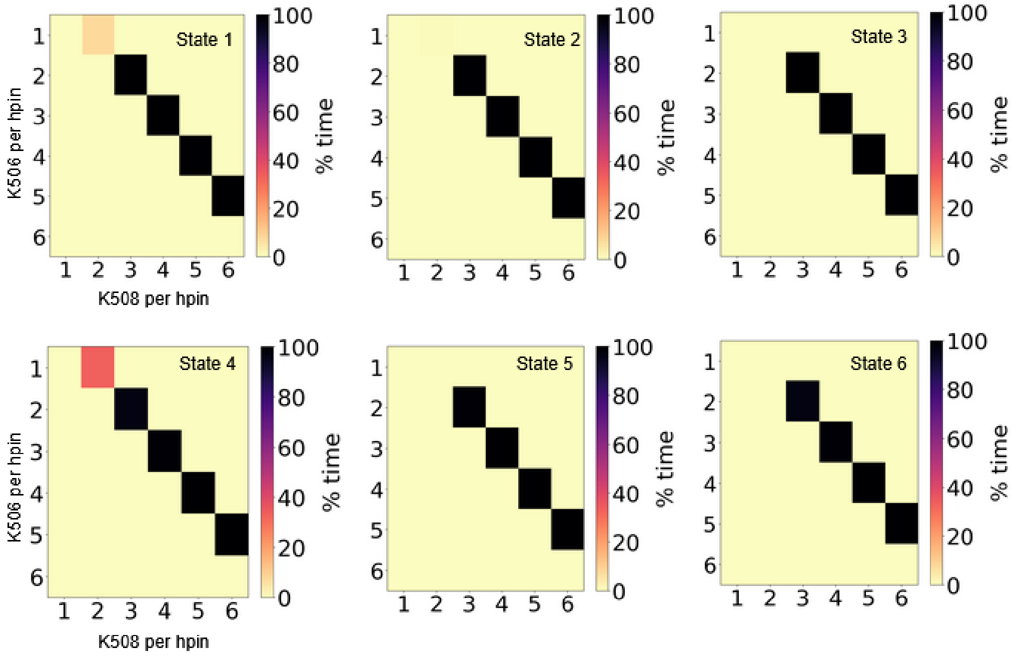

Figure 4.13: The percentage of time there is a stair-case interaction between the end chain K506 lysine of one hairpin, and the main chain K508 lysine of the hairpin below, for each state.

4.3.8 H507 and K506

It did not seem like H507 or K506 had any significant effect on DNA separation. However, they could still have supporting roles. Looking at figure 4.14, the K506 of every hairpin, including hairpin 1, interacts with its translocating nucleotide (eg. h1-K506 with as3, h2-K506 with as4, etc) in every state. Not only this, but the interaction occurs for the entire duration of the simulation. The fact the bond is maintained 100% of the time means there is little movement of the K506 from the DNA. It seems then that K506 is able to grab its nucleotide easily, and does so before DNA separation, and before the interactions of H507.

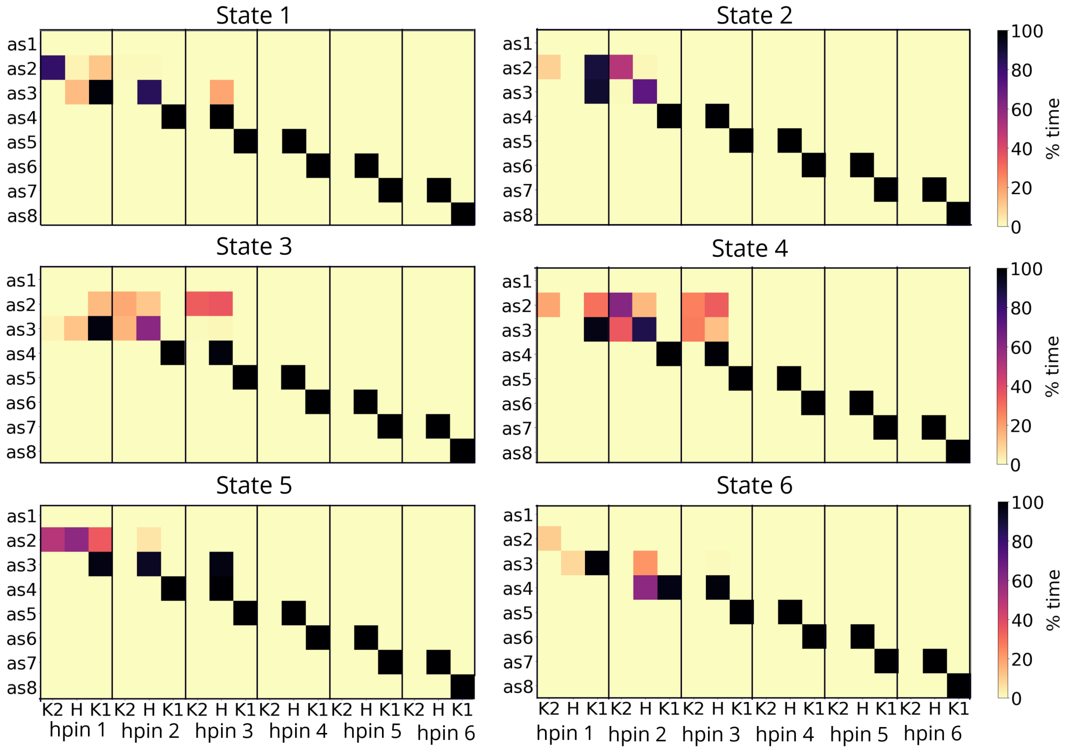

Figure 4.14: Heatmap of the hairpin-DNA interactions for each state in the REMD simulations. Hydrogen bonds are between the phosphate of as1 to as8 with the amino group of K506 (K1) and K508 (K2), and between the main chain amine and side chain epsilon of the histidine (H).

The histidine, H507, is a lot less consistent. Although, h2-H507 does interact with its translocating nucleotide, as3, in all of the states, these are not maintained for the whole simulation. Accompanying this, only h1-H507 interacts with its translocating nucleotide, as2, in state 4. This state is special compared to the others, as it is the only one where the two strands are fully separated. Along with this, as2 is pulled down towards the hairpins, where it would seem it is now close enough to interact with h1-H507. This indicates that H507 can only grab its translocating nucleotide once it moves further into the chamber.

4.4 Discussion

The REMD simulations had much higher sampling of the conformational space, however not all of the space was observed. State 2 and 3 showed little to no separation over the 50ns, most likely due to fewer exchanges between the low and high temperature replicas. As a result, the two simulations were stuck in local minima, and unable to access the DNA-separated states. Exchanges are accepted when the energies, and conformations, of two replicas sufficiently overlap. For states 2 and 3, this overlap with the higher and lower temperature replicas seemed to happen less frequently than the others. However, more thorough exchanges did occur near the end of the 50ns. So better sampling could occur for these states should the simulations be extended. Extending the simulations could also test the hypothesis of K508 as the key hairpin grabber. Only state 5 observed K508 pulling the DNA into the channel. However, it may be possible for this to occur in all states given enough simulation time.

The simulations generated possible helicase mechanisms, however further validation is needed to say for certain if this is true. Further work on states 2 to 6 are especially required, since these were generated from the "cut and stick" method. Rotating the AAA+ domain and rejoining it to the CD, as well as the two DNA pieces, may have introduced inaccuracies. The method also relies on the states being conformationally identical eg. subunit A with the hairpin at the top is identical to subunit B when its hairpin is at the top. The results for state 1 are likely to be the most accurate, and thus have a high degree of confidence, but experiments would still be required to validate these predictions.

Experiments with mutations of the E1 helicase could determine the role of K508 without observing the mechanism directly. Should K508 be replaced with a non-electrostatic residue such as glycine, where the side chain is a single hydrogen atom, then comparisons could be made with the original. If the G508 mutation is less processive than K508, then it would prove K508 has a role in DNA translocation.

Should the experiments confirm the simulations, then this work will have found a new interaction between E1 and the DNA, which is key to helicase activity. The interaction between the K508 lysine of the top hairpin, and the second ssDNA nucleotide, would demonstrate an active mechanism of strand separation. Meanwhile, a passive mechanism would also be observed in separating the base pairs via thermal activity and certain DNA conformations. This would demonstrate for the first time, that the helicase activity arises from a combination of both active and passive mechanisms.

4.5 Conclusion

The REMD simulations provided new insights into how the E1 helicase separates strands of DNA. The base pairs can be separated by a combination of a passive and active mechanisms. The passive mechanism destabilises the base pairs through thermal motion and DNA deformation, while the active mechanism is due the top hairpin lysine K508, which can pull the strands apart. The top hairpin lysine K506, known previously to translocate the DNA, grabs the third ssDNA nucleotide first. Then the lysine K508 secures the second ssDNA residue, orientating the DNA into the "hands" of the hairpin, as per the "croupier model".

The H507 was not found to have any significant involvement in DNA base pair separation or DNA grabbing, and thus is likely to be the last hairpin residue to bind to the DNA. This, in all likelihood, is due to the shape of the histidine molecule compared with lysine, and the method by which both molecules bind with the DNA. In the case of lysine, the rod-like structure with a charged point can cleanly interact with the DNA, as firstly there are no bonded atoms around it to obstruct its movement, and secondly, it is able to interact with atoms further away. Histidine on the other hand, despite having a charged point on its ring, cannot interact with DNA to the same extent, and binds to DNA via the main chain amino group. There is undoubtedly an evolutionary purpose for the histidine's location on the hairpin, however this was not observed in the simulations. Instead, the two flanking lysines were found to have superior capabilities, and were the most important, in these simulations, to the helicase function.

Overall, the results signify a K506-K508-H507 DNA grabbing mechanism. The K506 grabs first, followed by K508 which pulls the nucleotide into the channel. The histidine then binds last. This is the first time this has been observed, and answers the research question posed "What is the mechanism by which the helicase grabs the DNA?". The results also demonstrate that this mechanism can aid in separating the strands of DNA. The function of the helicase.

Chapter 5

Molecular Dynamics Simulations of the Rep Helicase

5.1 Introduction

There are two known structures of the Rep helicase; the open form and the closed. These two structures had been observed using x-ray crystallography, however the dynamics between these two structures were unknown. Working with experimentalists, the aim was to combine FRET data and molecular dynamic simulations, to determine if there were intermediate states between the open and closed structures. The FRET data would determine the number of transition states, by categorising the conformations from the FRET efficiency of dyes placed on the 1B and 2B subdomains. The conformations sampled in the molecular dynamics simulations could then be compared with the FRET data, to see if they complimented each other, and if so, what was occurring to the conformations of Rep.

Due to a 2017 paper on UvrD [60], the experimentalists chose to study Rep at two salt conditions; 0.01M (low salt) and 0.5M (high salt). The paper found that UvrD, a structurally and functionally similar protein to Rep has salt dependency, and also took four conformations. The Leake group therefore decided to see if the same was true of Rep.

5.2 Methods

5.2.1 Completing structure

The starting structure for the simulations came from the 1UAA Rep structure on the protein data bank [4, 5]. The structure had both the open and closed conformations bound to a strand of ssDNA. However, since the experiments took place in the absence of DNA, the DNA was removed, with simulations conducted using only the open or closed structure.

Both of the Rep conformations had missing residues in the 2A-2B hinge. The closed structure was missing eight residues (M539, M540, E541, R542, G543, E544, S545, E546), while the open structure was only missing three (G543, E544, S545). Since the open structure had more of the loop, the existing sequence was used to fill a part of the hole. The open structure was superimposed onto the closed so that the edges of the gap aligned, and the missing residues could be "cut and stuck" back in (see figure A3 in the appendix for the full sequence of Rep and UvrD).

As for the remaining three residues, the loop was filled using UvrD, obtained from the 2IS2 structure in the protein data bank [37, 127]. The structures can be used to fill the gaps as it is structurally similar. Despite the structural similarity, there are differences in the amino acid sequence, and so these were changed to match the sequence found in Rep.

5.2.2 Molecular Dynamics

The MD simulations were set up to have similar conditions to the FRET experiments, in order for better comparability. As such, both the MD and FRET experiments were conducted in 0.01M, for a low salt and 0.5M, for a high salt concentration. The pH in the FRET experiments was around 7.5. The MD simulations do not have a pH environment, but the protonation states of the protein were determined for a pH of 8.0. In order to sample more of the conformational landscape, implicit solvent was used, as well as four replicas, identical simulations with different starting velocities. Each replica ran for 50ns in one of the four conditions: high salt and low salt, starting in open and closed, totalling 16 simulations and 800ns.

The Rep simulations had very little convergence. As can be seen in Appendix figure A4, the only condition where all the replicas converged, and did so at a similar point, was those starting from the closed structure in low salt (0.01M). All of the replicas in the other conditions had differences in their RMS values. These values were much greater than for E1, with the plain MD E1 simulations converging around 3Å, and the REMD simulations converging around 0.4Å. Even the Rep simulations that converged did so around 9Å, three times greater than for the plain MD simulations of E1.

5.3 Results

The simulations showed very different results depending on their starting structure and ionic environment. Those starting in the open conformation showed the most changes in the structure. The simulations starting in the closed conformation, in low salt, showed the least deviation from the starting structure. Overall, the structures moved further from their starting structures, when they were in a high salt concentration.

5.3.1 Results from the Open Structure

Figure 5.1 shows how one of the replicas starting in the open conformation in low salt evolved in the simulation. The global conformation changes considerably, moving away from the open conformation. The biggest changes are as a result of movement in the 1B and 2B subdomains, which begin to move away from each other between 10 to 15ns. They return at the 20ns mark, but move apart again for the duration of the simulations. The separation of these subdomains seems to be mostly from the 1B subdomain moving away, though there is also movement of 2B. Similar results were seen in the other replicas, with movement away from the open conformation, contributed mostly from the 1B domain. Over the course of the simulation all replicas move further away from the closed crystal structure, starting with a difference of around 15Å when the production run starts, and increasing to a maximum of around 30Å in the case of replica 0.

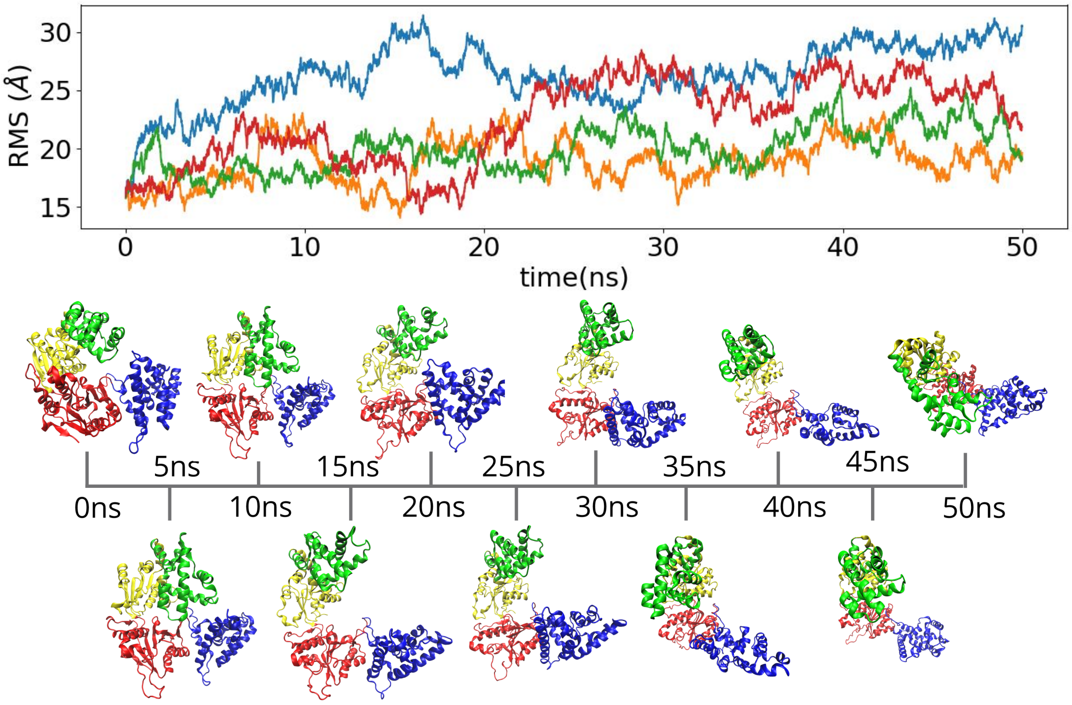

Figure 5.1: Top. RMS over time of the Rep simulations starting in the open conformation in high salt, compared with the closed crystal structure. *Replica 0 - blue, replica 1 - orange, replica 2 - green, replica 3 - red.* Bottom. Snapshots of the simulation starting in the open conformation in low salt (replica 2 of 4) *1A - yellow, 1B - green, 2A - red, 2B - blue.*

As can be seen in figure 5.2 similar results were seen in the high salt, but on a shorter time-scale. Again, the 1B subdomain was the most dynamic, but instead of moving away after 10ns, it took only the first 5ns. The 2B subdomain also showed movement, both compared to 1B and 2A, with possible rotation about the 2A-2B hinge. It was unclear whether the rotation was the same as appeared in low salt. Whether in high or low salt, all the replicas moved considerably over the 50ns duration. Again, these movements are away from both the closed crystal structure, and the open one from which the simulations began. While replicas 0 and 2, remain closer to the starting open conformation, replicas 1 and 3 diverge more, surpassing the RMS of those in high salt by the end of the 50ns simulations.

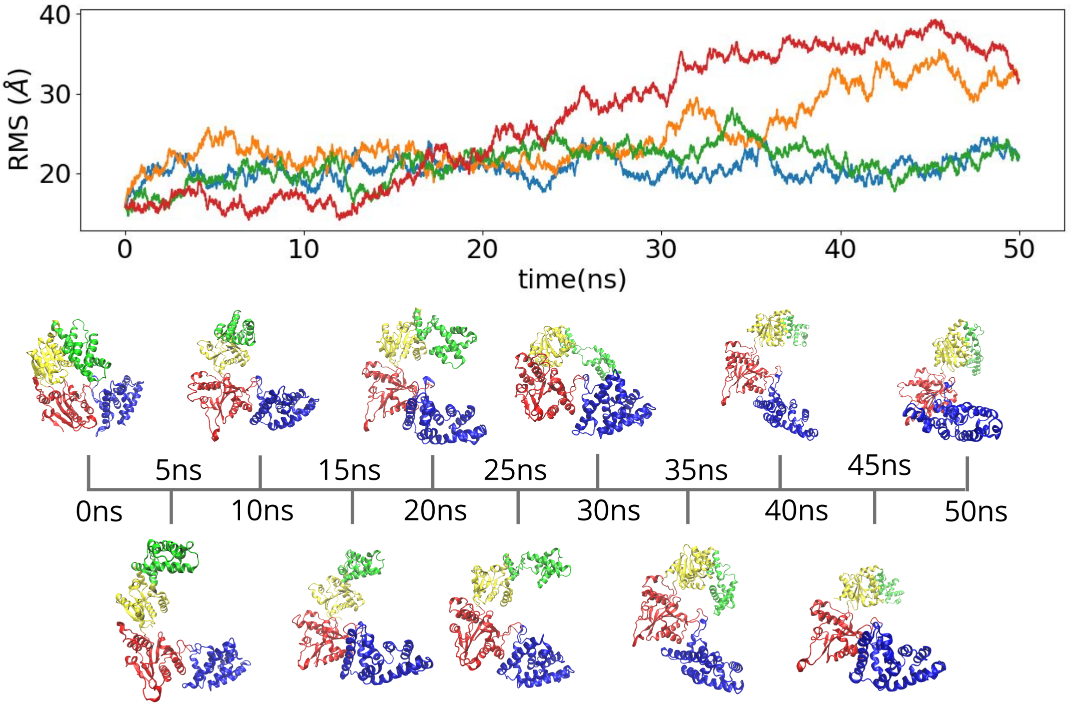

Figure 5.2: Top. RMS over time of the Rep simulations starting in the open conformation in low salt, compared with the closed crystal structure. *Replica 0 - blue, replica 1 - orange, replica 2 - green, replica 3 - red.* Bottom. Snapshots of the simulation starting in the open conformation in high salt (replica 1 of 4) *1A - yellow, 1B - green, 2A - red, 2B - blue.*

5.3.2 Results from the Closed Structure

The simulations starting in the closed structure had less movement than those from the open. Even so, there was still movement away from the starting closed structure, mostly from the 1B subdomain, as can be seen in figure 5.3. One of the replicas (replica 1) did show similar dynamics to those in open, with the 1B and 2B domains moving further apart. However most of the simulations remained relatively compact, like those in figure 5.3. These replicas (0, 2 and 3) did not move considerably towards or away from the open crystal structure, though replica 2 moved a small amount of towards the open structure at the beginning, before returning back. Replica 1 appeared to move closer to the open structure for the first 15ns, and the deviated further away for the remainder of the simulation.

The closed conformation in low salt showed the least movement across all the simulations. The simulations did not seem to move considerably away from their starting structure, maintaining roughly the same shape over the 50ns duration. The result was consistent across all the replicas. As a result, the replicas moved neither closer or further away from the open structure, as seen in figure 5.4.

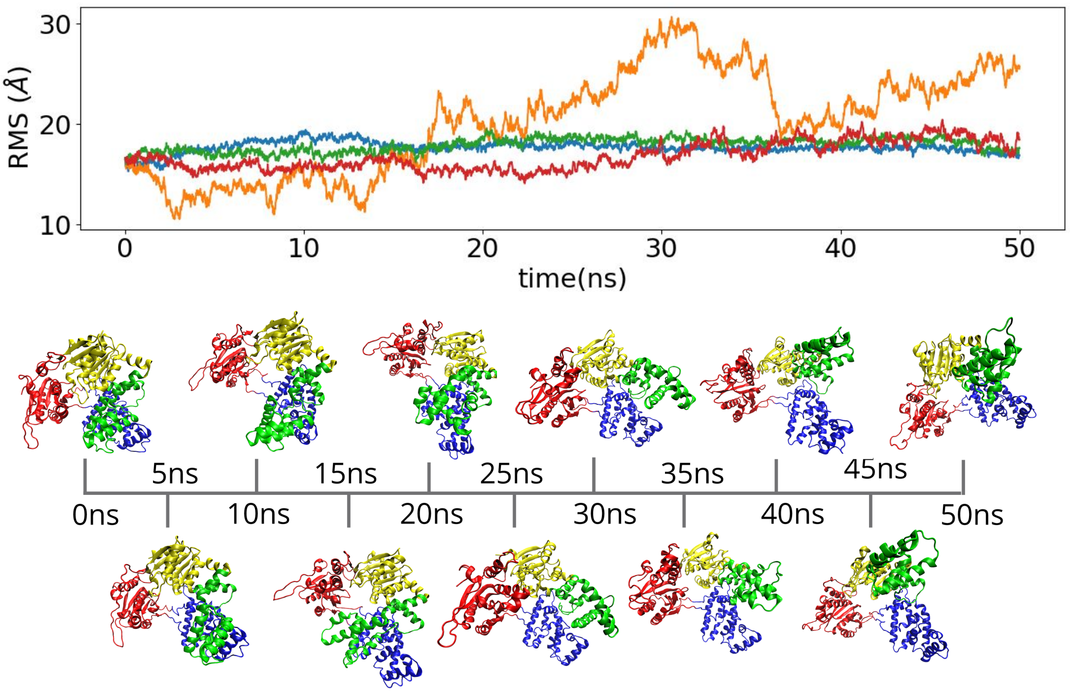

Figure 5.3: Top. RMS over time of the Rep simulations starting in the open conformation in low salt, compared with the closed crystal structure. *Replica 0 - blue, replica 1 - orange, replica 2 - green, replica 3 - red.* Bottom. Snapshots of the simulation starting in the closed conformation in high salt (replica 2 of 4) *1A - yellow, 1B - green, 2A - red, 2B - blue.*

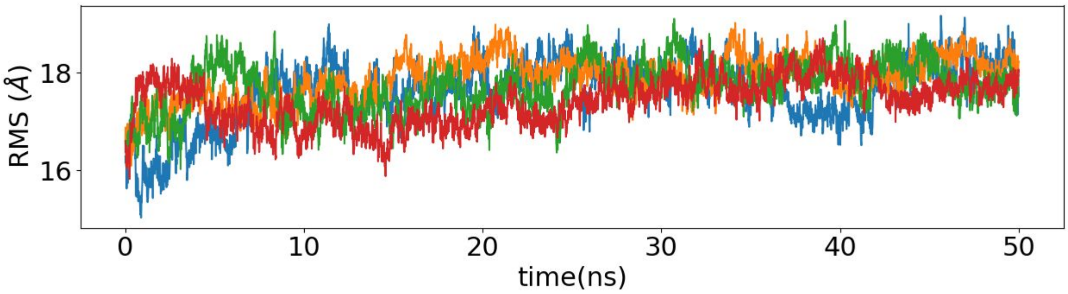

Figure 5.4: Top. RMS over time of the Rep simulations starting in the open conformation in low salt, compared with the closed crystal structure. *Replica 0 - blue, replica 1 - orange, replica 2 - green, replica 3 - red.*

5.3.3 RMS displacement

The structural changes observed were quantified by measuring the RMSD of each residue from its starting structure (open or closed). The averages of each replica were measured, and these were again averaged to show the overall changes per residues over all the replicas, shown in figure 5.5. The RMSD analysis showed the conformational flexibility of the protein, with an average RMSD of approximately 10Å. It also confirmed that 1B moved more than the other subdomains, a difference of approximately 25Å, whether starting in open or closed. This is unlike the previous studies, where 2B showed the greatest movement.

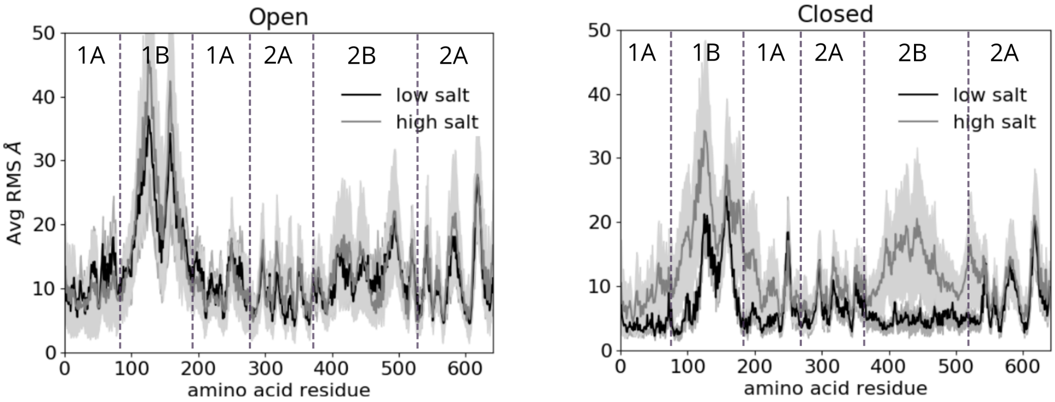

Figure 5.5: The average of the average RMSD of each amino acid residue from its starting structure in each replica Left. starting from the open conformation and Right. from the closed. *Shaded area denotes the error*

Starting in the open conformation, the salt concentration showed little difference in the overall RMSD. The same could not be said when starting from closed. Here, the simulations in high salt showed considerably more movement in 2B, with an RMSD of roughly 20Å compared with 5Å in low salt. The error in 2B was also the greatest error. The reason was the differences in the replicas. Replica 1 and replica 2 had a greater RMSD in 2B than the other replicas. The other two replicas having RMSD similar to those in low salt. The smallest deviation was seen in the closed starting structure in low salt, as the replicas had similar RMSD.

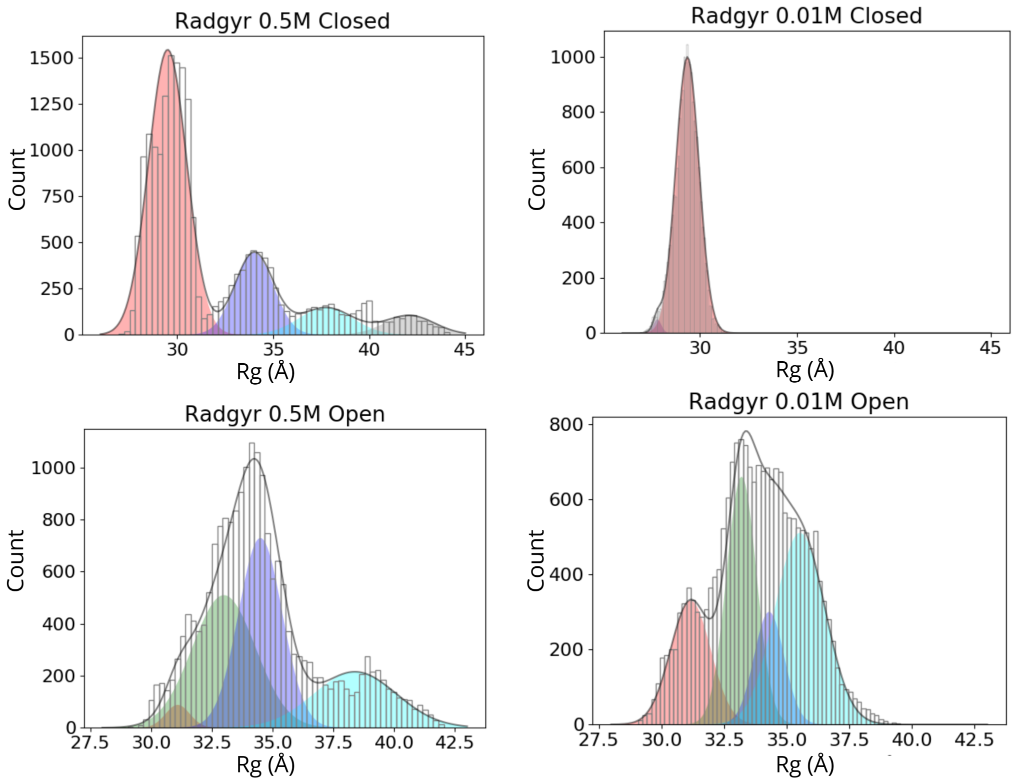

Figure 5.6: Histogram of the radius of gyrations from the replicas in each of the four conditions. *Similar peaks of radius of gyration (Rg) in the same colour: state 1 (red) Rg = 31.1Å, state 2 (green) Rg = 33.1Å, state 3 (blue) Rg = 34.3Å, state 4 (cyan) Rg = 35.6Å, state 5 (grey) Rg = 42.1Å and state 6 (purple) Rg = 27.8Å. Black reflects the overall fit.*

5.3.4 Radius of gyration

The radius of gyration was measured in order to gain insight into the global conformations. Radius of gyration (Rg) is a measure of compactness, and was used to quantify the movement of the 1B and 2B seen in the simulations. Figure 5.6 shows that all the simulations became more spread out than the closed structure, with Rg = 26.75Å, and open structure, Rg = 28.81Å. It also shows that the Rg could be grouped into a number of states with Gaussian distributions. The centre of the distributions was determined by Dr Steve Quinn, who fitted the peaks using Origin. Four peaks were chosen due to the FRET data being fitted with four peaks previously, as in figure A5. The peak positions were constrained, with the peak height and width unconstrained. Each peak had a coefficient of determination (R^2) of 0.99.

The first state, state 1 (red), appeared to be the closest to the closed state, with Rg = 31.1Å. It was the main state in the simulations starting in the closed structure in both high and low salt. A smaller peak was also observed in the simulations starting in the open structure, with a higher peak in low salt.

The next state (green) had a Rg of 33.1Å. This state only appeared in the open simulations, and was absent from the closed. However the third main state, state 3 (blue), was found in the open simulations and the high salt closed simulation, with Rg = 34.3Å. The last state, state 4 (cyan) to be found in both the open and closed simulations had Rg = 35.6Å. There were two much smaller states, both in the closed simulations. The one in the high salt (grey) had Rg = 42.1Å. In the low salt there was a very small peak (purple) at Rg = 27.8Å.

5.3.5 Rotation of 2B subdomain

In the MD simulations, the angle between 2B and 2A was measured using the centre of mass of each subdomain (as described previously in Chapter 2) with the results shown in figure 5.7. Both in low and high salt, the closed simulations

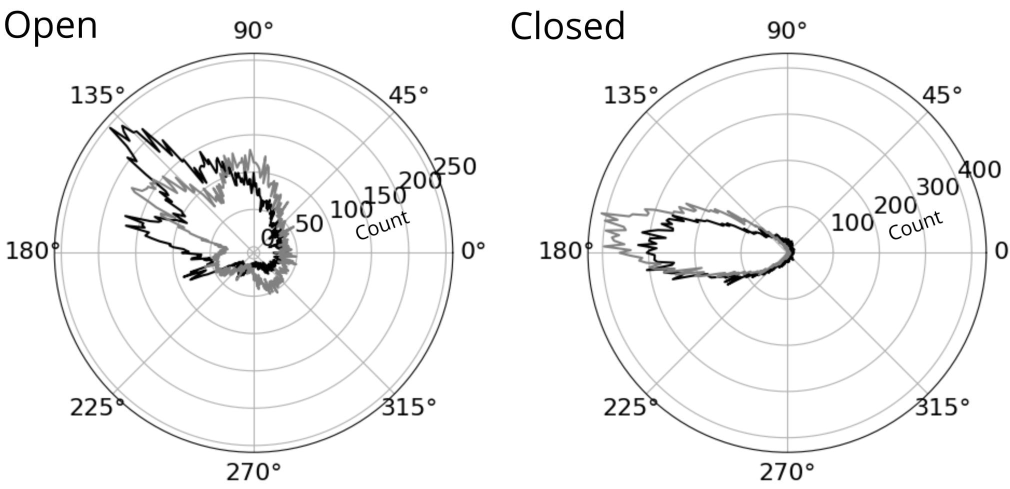

Figure 5.7: Angle between 2B and 2A subdomains *0.5M salt concentration (black), 0.01M salt concentration (grey).*

peaked around 174°, with no peaks at other angles. This is consistent with the other results, suggesting the structures were stuck in local energetic minima. The open simulations on the other hand, showed multiple peaks, roughly four peaks for low salt, at angles of approximately 95°, 146°, 200°, and 300°. Meanwhile, the high salt peaked around 120°, 137°, 167°, and 202°.

5.3.6 MD vs Experiment

The four state prediction observed in the radius of gyration and 2B hinge rotation, was also seen in the FRET results obtained by Dr Steve Quinn, Dr Jamieson Howard, Dr Ben Ambrose, Dr Tim Craggs and Prof Mark Leake (Appendix. Figure A5). Here, the experiments placed a dye on the 1B (residue 97) and 2B (residue 473) subdomains (Appendix. Figure A6), in the absence of DNA, in either 0.5M or 0.01M NaCl concentration. The shortest time-scale measurement by FRET was 1ms, and while this is much longer than the MD simulations, the use of implicit solvent with reduced viscosity would have increased the MD time dramatically. Since it is not known if the time scale of the MD simulation and Rep simulation coincide, it is important to look at the overall results rather than the details.

The experiments found that the most common state was the closed conformation, suggesting it is a stable structure, which can maintain its shape for long periods. Three other structures, with 2B and 1B further apart were found including the open conformation. The two new conformations had a FRET efficiency between the open and closed.

The FRET results also showed that the conformations of Rep are affected by the salt concentration. More open conformations were observed in the 0.5M solution (high salt) compared with 0.01M solution (low salt).

5.4 Discussion

Both the experiments and MD simulations found multiple states, however these might not be the same structures. The FRET experiment was obtained over the second time-scale, whereas MD was over nanoseconds. Even so, the results could be comparable when considering the low viscosity of the implicit solvent. Reducing the viscosity can increase the rate of sampling so the simulated time is likely much longer than 50ns. Whether it is as great as seconds is difficult to tell, as there is no common method for adjusting the simulation time.

In all the simulations, the structures moved away from their starting structures, resulting in different radius of gyration than either the open or closed crystallographic structure. The simulations were run without DNA, and it is possible that when Rep binds to DNA, it maintains the proteins structure. Simulations with Rep bound to DNA could be run to see if this is the case.

Although it seemed to be the 1B subdomain that moved, rather than the 2B subdomain in the literature [4], the MD results could still match with those from the FRET experiments. In these experiments, the FRET efficiency changes with respect to the distance between the 1B and 2B domain. Therefore it could be the

1B subdomain moving away from 2B, which caused the change, rather than 2B from 1B. Further experiments would be required to see which moved more, such as moving the dyes to 1B and 2A for example, to see how the subdomains moved with respect to each other.

5.5 Conclusion

The Leake group surmised that a four state model similar to UvrD, seems to have been observed in Rep. Both MD simulations and experiments found that structures of Rep could be classified into four groups. The group saw four peaks in the FRET efficiency, and four main peaks in the radius of gyration distribution from the simulations.

Both experiments and simulations indicated that the closed structure is more energetically favourable than other structures, as it is maintained for longer than the other states. In the MD simulations, the closed simulation in 0.01M salt concentration stayed within one conformation, with the radius of gyration and 2B rotation having no significant changes. Meanwhile, the radius of gyration in the 0.5M salt concentration, was predominately the same as in 0.01M. Confirming the results from the simulations, the experiments found that the most common FRET efficiency was in the closed structure.

The conformations and dynamics of Rep was found to be dependent on the salt concentration of its environment. Again, both the MD simulations and experiments were in agreement. The results revealed that Rep is more likely to be in its more compact states, such as the closed structure, in 0.01M salt concentration. Correspondingly, Rep takes a more expanded form in 0.5M salt concentration. The salt concentration can effect protein conformation through Debye shielding, whereby charged ions can shield or screen other charged particles. In molecules such as proteins, this can allow conformations to change, as the effect from the charged atoms is felt over a smaller range with a greater amount of charged ions.

Chapter 6

Discussion

For the first time, simulations of the E1 and Rep helicases have been conducted. Not only this, but a novel approach for investigating multimeric motors was designed, and implemented, via the "cut and stick" or "Frankenstein" approach. Together, these approaches deduced a mechanism for base pair separation in the E1 BPV helicase never seen before, as well as conformational changes of Rep conformers in the absence of DNA. Both of these results, in turn, expanding the field of biochemistry by adding the next layer to the picture of these two magnificent motors; their dynamics. Until now, both proteins had been observed statically at the atomic level [4, 39], or dynamically at the molecular level [2, 70, 71, 72]. However, by using molecular dynamics, this gap has been filled by allowing the dynamics on an atomistic scale to be seen.

6.1 Papillomavirus E1 Helicase

The "cut and stick" method, along with enhanced modelling techniques, discovered the inner workings of the E1 helicase, not seen before. The chamber was found to be key in the function of E1, providing the space for the DNA to adapt its conformation, from ten bases per turn to six. It was also the location for two important functions, the separation of the two strands of DNA, and the grabbing of the newly separated strand. The hairpins grab the DNA first by K506, then by K508, a new interaction which was found to pull the DNA strand, with H507 interacting last.

Of course, these results are purely from the simulations, and so real-world evidence will be needed to see if this is the case in reality. One criticism of the simulations, is that they could be considered short at 50ns, and that the plain MD simulations had no replicas. Should this research be repeated, then replicas or longer simulation time could be used. However, there is already a high confidence that the results produced in this thesis would agree with the experiments. Firstly, the RMSD convergence plot plateaued for both the plain MD and REMD simulations, signifying that both had found local minimas. These plateaus occurred within a small RMSD, implying that the structures did not move significantly away from the original structure. Therefore the simulations did not dramatically change or distort the protein's structure, resolved using x-ray crystallography.

As it is important to see whether K508 does indeed assist DNA base pair separation in the E1 helicase, experiments need to be conducted. Dr Cyril Sanders, a collaborator from Sheffield university, is due to mutate the K508 residue to a more inert amino acid like glycine, to see how this affects the rate of DNA unwinding. Should the mutated helicase perform worse than the wild-type, then it could be presumed that K508 does impact DNA base pair separation.

The work from this thesis found that the chamber in the E1 helicase could be important for DNA base pair separation, so it is possible that this chamber could be a target for anti-virals; for BPV, and potentially HPV too. For example, finding molecules that could block the chamber or hinder the movement of the hairpins, could in theory prevent the helicase from functioning, and therefore stop the virus from replicating. One method of analysing the druggability of the helicase is via DruGUI and Prody [128, 120]. These are programmes developed in Python that can test where certain molecules might bind to a protein.

There is currently no crystallographic structure for the E1 HPV helicase, so before any computational analysis of druggability can be performed on the E1 HPV helicase, a structure needs to be found. Since there are already known sequences of HPV [129], this can be done using Alpha Fold [130], which predicts protein structure based upon known sequences and structures. Once this structure has been obtained, any number of analysis could be performed to not only discover its druggability but answer other questions about the helicase's mechanism.

The "cut and stick" method designed for the E1 BPV helicase could also be used to probe the internal mechanisms of the E1 HPV helicase and comparable proteins, such as other hexameric helicases, like T7gp4 and MCM. These proteins are multimeric ring-like structures that surround DNA, and translocate it with DNA-binding hairpin loops, similar to the E1 helicase. This method then, could be used to advance the understanding of proteins across the helicase superfamilies.

The aim of this research on the E1 was to decipher a mechanism of base pair separation, however there is more about this helicase that can be explored. It is not yet known how the conformational changes of the subunits take place, how the rotary wave of ATP-hydrolysis circulates around the protein, and the affect of ATP. The answer could lie in other computational techniques, such as an Elastic Network Model (ENM) [131], which can probe dynamics on time-scales far greater than molecular dynamics, using connections of the alpha-carbon atoms.

6.2 E. Coli Rep Helicase

The simulations of Rep corroborated the experimental results that Rep is affected by the salt concentration of its environment. Though it would be mindful to take into account the convergence plots. Unlike the simulations of E1, the RMS convergence plots of Rep displayed a large degree of movement away from the original structures. And, although replicas were produced, many of them did not converge. However, this may be in-line with experimental results, and due to the conditions, notably the absence of DNA, in both the experiments and simulations.

The original crystallographic structure had two proteins bound to ssDNA, one in the open structure, one closed. The simulations were run on just one structure, with the other, and the DNA, removed. The decision was due to wanting the simulations to most closely match those of the FRET experiments. However, the presence of DNA, and the other molecule, may be required to maintain stable conformations. This could be what occurs in the FRET experiments, since the predicted FRET values of the simulations did coincide with the experimental values. However, more replicas would likely be needed, along with AFM, to see how the Rep structures in vitro and in silico align.

Another limitation of these simulations is the time scale. The smallest measured time step of the FRET experiments is around 1ms, with a standard time resolution of tens-of-milliseconds [132]. In comparison, the time step of the MD simulations was 2 femtoseconds [95], with a total simulation time of 50ns, several orders of magnitude smaller than those observed in the FRET experiments. However, there are ways of accelerating the sampling in MD simulations.

MD simulations can be "sped up" by reducing the viscosity of the solvent (in implicit solvent), allowing more conformations to be reached in a shorter time. This was the case in the Rep simulations (and also in the REMD of E1). Should this work be repeated, then a better tool may have been ENM, which can predict dynamics over longer time-scales. However, in this case, the Rep group desired molecular dynamic simulations. So the simulations used a lower viscosity in the aim of reaching conformations observed in the experiments. There is no way to currently measure how the simulation time relates to real time. Therefore, it is not known if the simulations were sampling on the same time scale as those observed with FRET. Further work would be required to check if this were true, such as by comparing the FRET values over time with the predicted FRET values of the simulations.

The work produced here was conducted on the Rep helicase in the absence of DNA. Although these results may not be able to describe the mechanism of Rep, both as a helicase and a remover of road-blocks, it could be used to discover how Rep binds to the DNA. Different computational tools, such as ligand-docking programmes, could further develop this line of research. As for the function of Rep as a DNA processor, other advanced MD simulations could be used, such as Collective Molecular Dynamics [133], which could predict how Rep moves between the open and closed structures, and the mechanism of the 2B rotation. This would not only apply to Rep, but to similar proteins such as UvrD and PcrA.

Appendix

	Sim. Type	Protein/DNA	Start Str.	Salt Conc. (M)	Time (ns)	Replicas
3	MD	E1 w/DNA	Crystal Str.	0.15	50	0
		States 1 to 6	mutG(as2)	0.15	50	0
			muA(as2)	0.15	50	0
			mutC(as2)	0.15	50	0
4	REMD	E1 w/DNA	Crystal Str	0.15	50	12
		States 1 to 6				
5	MD	Rep	Closed	0.01	50	4
			Closed	0.5	50	4
			Open	0.01	50	4
			Open	0.5	50	4

Table 1: Table showing the details of the simulations produced for this thesis, including the chapter number, simulation type, the solute, starting structure, salt concentration, simulation time and number of replicas.

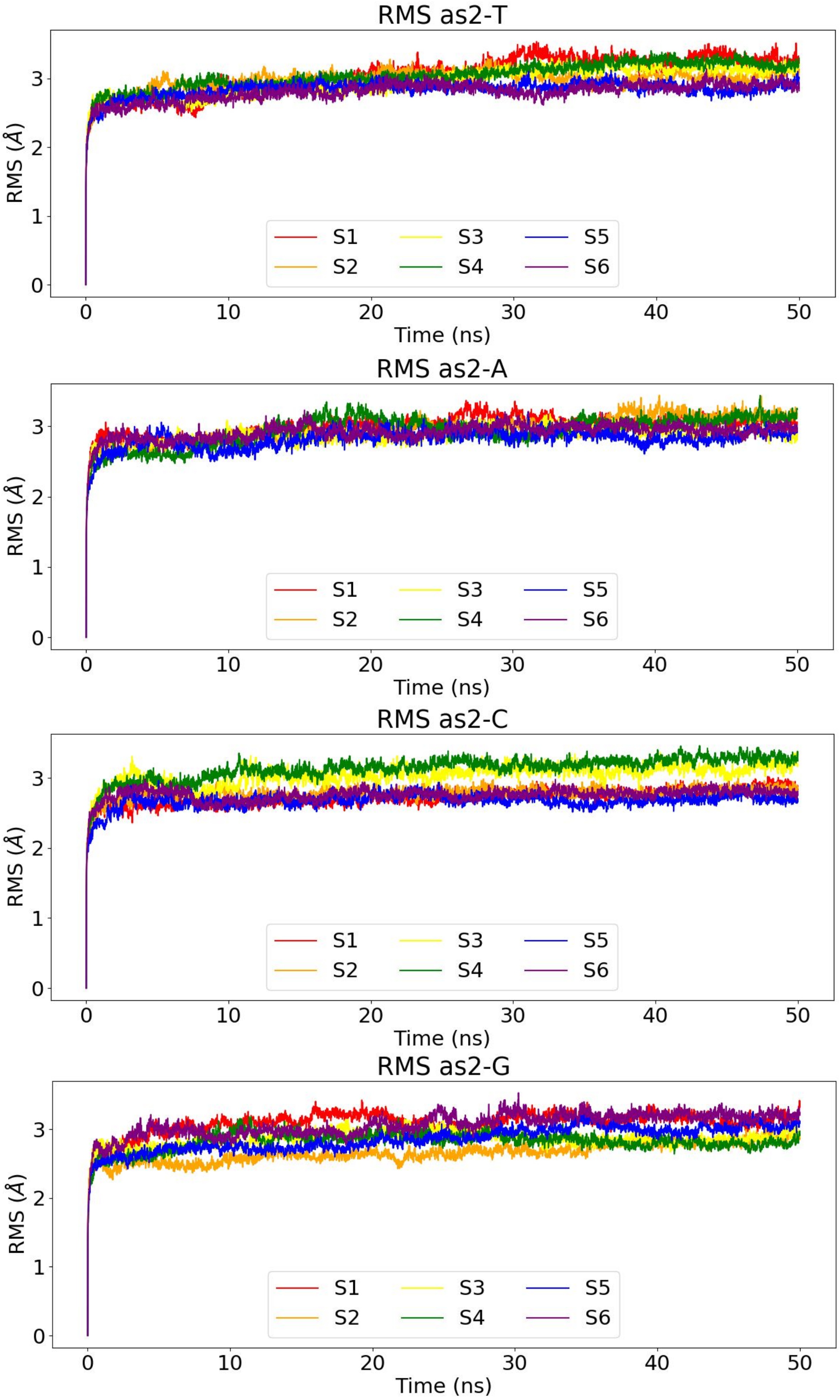

Figure A1: Convergence plot (RMS over time) of the plain MD simulations of the E1 helicase for each state and as2 base.

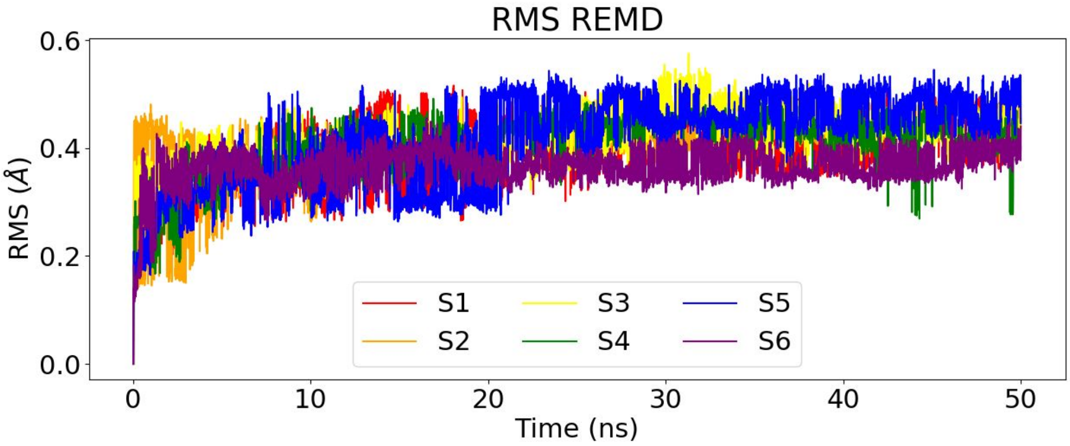

Figure A2: Convergence plot (RMS over time) of the REMD simulations of the E1 helicase for each state.

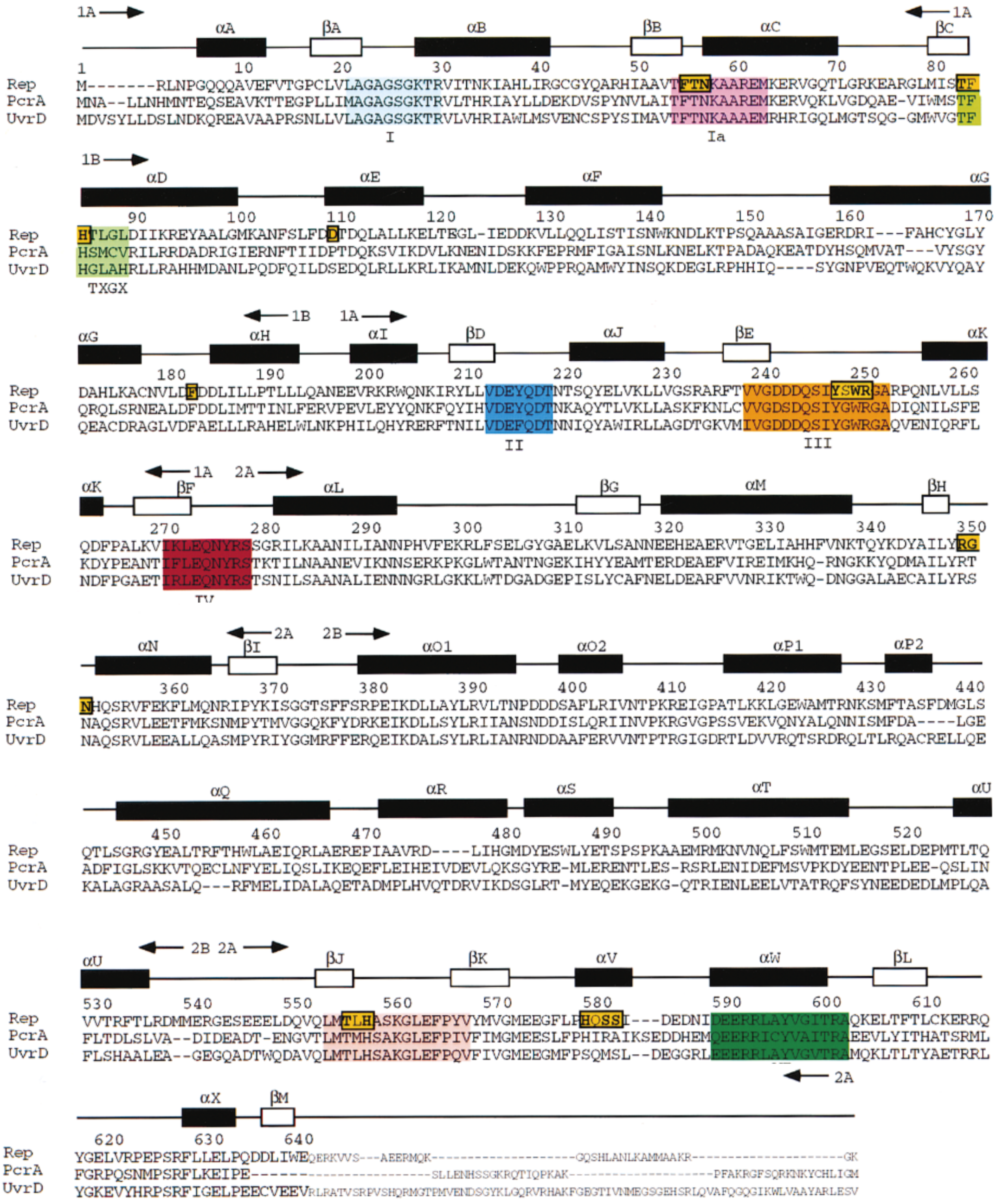
1A
αA
βA
αB
βB
αC
1A
βC
1 10 20 30 40 50 60 70 80
Rep M-------RLNPGQQQAVEFVTGPCLVLAGAGSGKTRVITNKIAHLIRGCGYQARHIAAVTFTNKAAREMKERVGQTLGRKEARGLMISTF
PcrA MNA--LLNHMNTEQSEAVKTTEGPLLIMAGAGSGKTRVLTHRIAYLLDEKDVSPYNVLAITFTNKAAREMKERVQKLVGDQAE-VIWMSTF
UvrD MDVSYLLDSLNDKQREAVAAPRSNLLVLAGAGSGKTRVLVHRIAWLMSVENCSPYSIMAVTFTNKAAAEMRHRIGQLMGTSQG-GMWVGTF
I
Ia
1B
αD
αE
αF
αG
90 100 110 120 130 140 150 160 170
Rep HTLGLDIIKREYAALGMKANFSLFDDTDQLALLKELTEGL-IEDDKVLLQQLISTISNWKNDLKTPSQAAASAIGERDRI---FAHCYGLY
PcrA HSMCVRILRRDADRIGIERNFTIIDPTDQKSVIKDVLKNENIDSKKFEPRMFIGAISNLKNELKTPADAQKEATDYHSQMVAT---VYSGY
UvrD HGLAHRLLRAHHMDANLPQDFQILDSEDQLRLLKRLIKAMNLDEKQWPPRQAMWYINSQKDEGLRPHHIQ----SYGNPVEQTWQKVYQAY
TXGX
1B
1A
αG
αH
αI
βD
αJ
βE
αK
180 190 200 210 220 230 240 250 260
Rep DAHLKACNVLDFDDLILLPTLLLQANEEVRKRWQNKIRYLLVDEYQDTNTSQYELVKLLVGSRARFTVVGDDDQSIYSWRGARPQNLVLLS
PcrA QRQLSRNEALDFDDLIMTTINLFERVPEVLEYYQNKFQYIHVDEYQDTNKAQYTLVKLLASKFKNLCVVGDSDQSIYGWRGADIQNILSFE
UvrD QEACDRAGLVDFAELLLRAHELWLNKPHILQHYRERFTNILVDEFQDTNNIQYAWIRLLAGDTGKVMIVGDDDQSIYGWRGAQVENIQRFL
II
III
1A
2A
αK
βF
αL
βG
αM
βH
270 280 290 300 310 320 330 340 350
Rep QDFPALKVIKLEQNYRSSGRILKAANILIANNPHVFEKRLFSELGYGAELKVLSANNEEHEAERVTGELIAHHFVNKTQYKDYAILYRG
PcrA KDYPEANTIFLEQNYRSTKTILNAANEVIKNNSERKPKGLWTANTNGEKIHYYEAMTERDEAEFVIREIMKHQ-RNGKKYQDMAILYRT
UvrD NDFPGAETIRLEQNYRSTSNILSAANALIENNNGRLGKKLWTDGADGEPISLYCAFNELDEARFVVNRIKTWQ-DNGGALAECAILYRS
IV
2A
2B
αN
βI
αO1
αO2
αP1
αP2
360 370 380 390 400 410 420 430 440
Rep NHQSRVFEKFLMQNRIPYKISGGTSFFSRPEIKDLLAYLRVLTNPDDDSAFLRIVNTPKREIGPATLKKLGEWAMTRNKSMFTASFDMGLS
PcrA NAQSRVLEETFMKSNMPYTMVGGQKFYDRKEIKDLLSYLRIIANSNDDISLQRIINVPKRGVGPSSVEKVQNYALQNNISMFDA----LGE
UvrD NAQSRVLEEALLQASMPYRIYGGMRFFERQEIKDALSYLRLIANRNDDAAFERVVNTPTRGIGDRTLDVVRQTSRDRQLTLRQACRELLQE
αQ
αR
αS
αT
αU
450 460 470 480 490 500 510 520
Rep QTLSGRGYEALTRFTHWLAEIQRLAEREPIAAVRD----LIHGMDYESWLYETSPSPKAAEMRMKNVNQLFSWMTEMLEGSELDEPMTLTQ
PcrA ADFIGLSKKVTQECLNFYELIQSLIKEQEFLEIHEIVDEVLQKSGYRE-MLERENTLES-RSRLENIDEFMSVPKDYEENTPLEE-QSLIN
UvrD KALAGRAASALQ---RFMELIDALAQETADMPLHVQTDRVIKDSGLRT-MYEQEKGEKG-QTRIENLEELVTATRQFSYNEEDEDLMPLQA
αU
2B 2A
βJ
βK
αV
αW
βL
530 540 550 560 570 580 590 600 610
Rep VVTRFTLRDMMERGESEEELDQVQLMTLHASKGLEFPYVYMVGMEEGFLPHQSSI---DEDNIDEERRLAYVGITRAQKELTFTLCKERRQ
PcrA FLTDLSLVA--DIDEADT-ENGVTLMTMHSAKGLEFPIVFIMGMEESLFPHIRAIKSEDDHEMEERRICYVAITRAEEVLYITHATSRML
UvrD FLSHAALEA--GEGQADTWQDAVQLMTLHSAKGLEFPQVFIVGMEEGMFPSQMSL--DEGGRLEEERRLAYVGVTRAMQKLTLTYAETRRL
2A
αX
βM
620 630 640
Rep YGELVRPEPSRFLLELPQDDLIWEQERKVVS---AEERMQK------------------GQSHLANLKAMMAAKR-------------------GK
PcrA FGRPQSNMPSRFLKEIPE-------------------SLLENHSSGKRQTIQPKAK------------------PFAKRGFSQRKNKYCHLIGM
UvrD YGKEVYHRPSRFIGELPEECVEEVRLRATVSRPVSHQRMGTPMVENDSGYKLGQRVRHAKFGEGTIVNMEGSGEHSRLQVAFQGQGIKWLVAAYARLESV

Figure A3: Sequence alignment of Rep, UvrD and PcrA *Figure 4 in Korolev et al 1997 [4]*

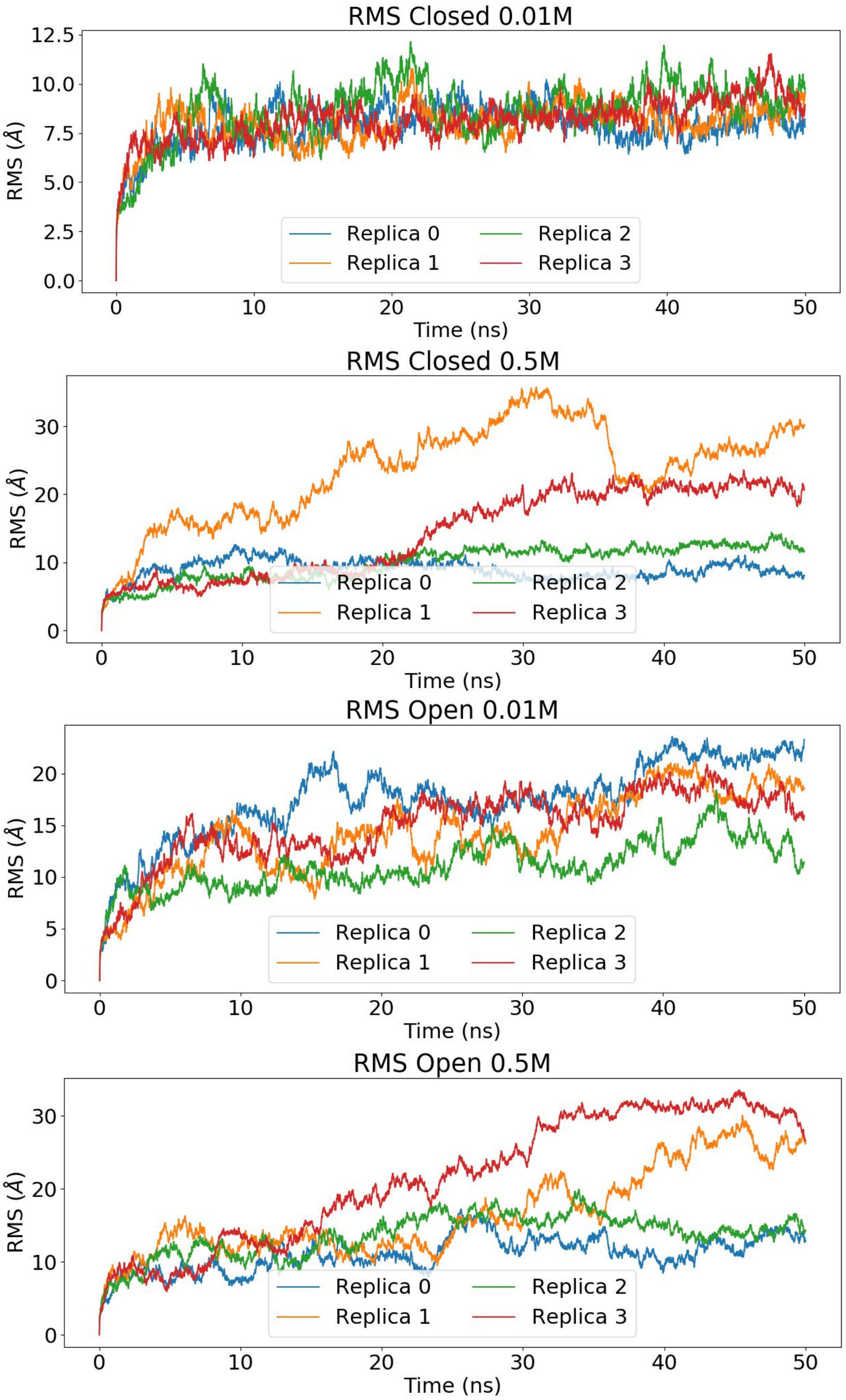

Figure A4: Convergence plot (RMS over time) of each Rep MD simulation.

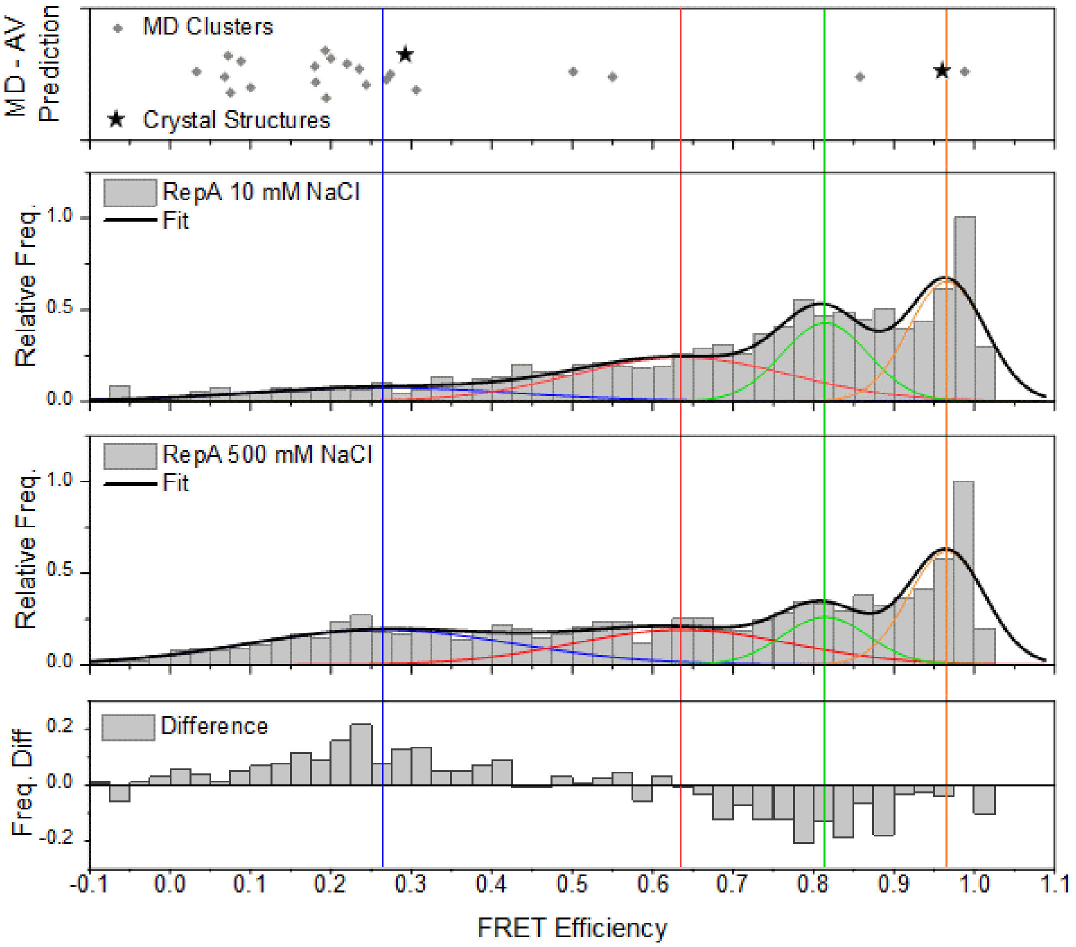

Figure A5: FRET efficiciency histogram with dyes placed on the 1B and 2B subdomains. *From Howard et al - in preparation*

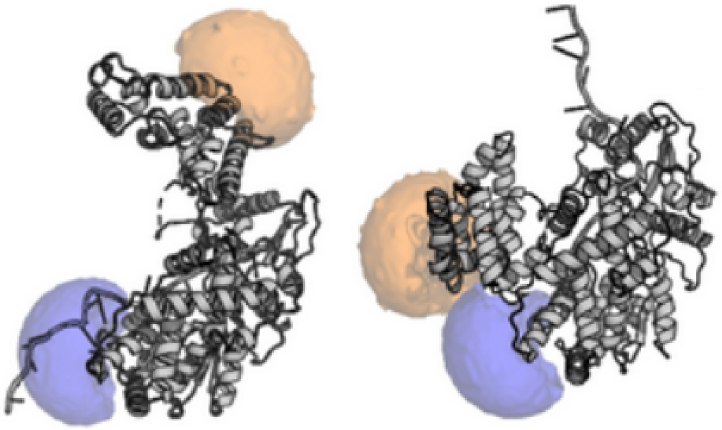

Figure A6: Location of the two FRET volumes (the predicted space that the dyes can occupy) on residues 97 and 473 of Rep, as determined by Dr Ben Ambrose. Left. Position of dyes on the open conformation. Right. Positions on the closed. *Blue - dye on residue 97 (1B subdomain). Orange - dye on residue 473 (2B subdomain). From Howard et al - in preparation*

Bibliography

[1] Martin R Singleton, Mark S Dillingham, and Dale B Wigley. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.*, 76:23–50, 2007.

[2] Seung-Jae Lee, Salman Syed, Eric J Enemark, Stephen Schuck, Arne Stenlund, Taekjip Ha, and Leemor Joshua-Tor. Dynamic look at dna unwinding by a replicative helicase. *Proceedings of the National Academy of Sciences*, 111(9):E827–E835, 2014.

[3] Yuriy Chaban, Jonathan A Stead, Ksenia Ryzhenkova, Fiona Whelan, Ekaterina P Lamber, Alfred Antson, Cyril M Sanders, and Elena V Orlova. Structural basis for dna strand separation by a hexameric replicative helicase. *Nucleic acids research*, 43(17):8551–8563, 2015.

[4] Sergey Korolev, John Hsieh, George H Gauss, Timothy M Lohman, and Gabriel Waksman. Major domain swiveling revealed by the crystal structures of complexes of e. coli rep helicase bound to single-stranded dna and adp. *Cell*, 90(4):635–647, 1997.

[5] Sergey Korolev and Gabriel Waksman. E. coli rep helicase/dna complex. https://www.rcsb.org/structure/1UAA, 1997.

[6] Karen M Corbett, Colin W Pouton, and David K Chalmers. Temperature replica exchange molecular dynamics simulations of cyclic peptide conformation. *Australian Journal of Chemistry*, 2021.

[7] Christopher Charles Whiston Taylor et al. *The atomists, Leucippus and Democritus: fragments: a text and translation with a commentary*, volume 5. University of Toronto Press, 2010.

[8] Steven B. Heymsfield, ZiMian Wang, Richard N Baumgartner, and Robert Ross. Human body composition: advances in models and methods. *Annual review of nutrition*, 17(1):527–558, 1997.

[9] Kenneth J Ellis. Human body composition: in vivo methods. *Physiological reviews*, 80(2):649–680, 2000.

[10] Zi-Mian Wang, Richard N Pierson Jr, and Steven B Heymsfield. The five-level model: a new approach to organizing body-composition research. *The American journal of clinical nutrition*, 56(1):19–28, 1992.

[11] Carl Ivar Branden and John Tooze. *Introduction to protein structure*. Garland Science, 2012.

[12] Georg E Schulz and R Heiner Schirmer. *Principles of protein structure*. Springer Science & Business Media, 2013.

[13] Javier A Velázquez-Muriel, Manuel Rueda, Isabel Cuesta, Alberto Pascual-Montano, Modesto Orozco, and José-María Carazo. Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC structural biology*, 9(1):1–14, 2009.

[14] Aron Marchler-Bauer, Myra K Derbyshire, Noreen R Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y Geer, Renata C Geer, Jane He, Marc Gwadz, David I Hurwitz, et al. Cdd: Ncbi's conserved domain database. *Nucleic acids research*, 43(D1):D222–D226, 2015.

[15] Ken A Dill, S Banu Ozkan, M Scott Shell, and Thomas R Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37:289–316, 2008.

[16] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[17] Gerald Platzer, Mark Okon, and Lawrence P McIntosh. ph-dependent random coil 1 h, 13 c, and 15 n chemical shifts of the ionizable amino acids: a guide for protein pk a measurements. *Journal of biomolecular NMR*, 60(2-3):109–129, 2014.

[18] Shenhui Li and Mei Hong. Protonation, tautomerization, and rotameric structure of histidine: a comprehensive study by magic-angle-spinning solid-state nmr. *Journal of the American Chemical Society*, 133(5):1534–1544, 2011.

[19] Cyrus Chothia, Michael Levitt, and Douglas Richardson. Structure of proteins: packing of alpha-helices and pleated sheets. *Proceedings of the National Academy of Sciences*, 74(10):4130–4134, 1977.

[20] K Gunasekaran, C Ramakrishnan, and P Balaram. Beta-hairpins in proteins revisited: lessons for de novo design. *Protein engineering*, 10(10):1131–1141, 1997.

[21] E James Milner-White and Ron Poet. Four classes of β-hairpins in proteins. *Biochemical Journal*, 240(1):289–292, 1986.

[22] Robert M Hughes and Marcey L Waters. Model systems for β-hairpins and β-sheets. *Current opinion in structural biology*, 16(4):514–524, 2006.

[23] BL Sibanda and JM Thornton. β-hairpin families in globular proteins. *Nature*, 316(6024):170–174, 1985.

[24] John W Pelley and Edward F Goljan. Rapid review biochemistry e-book. 2010.

[25] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

[26] Peter Y Chou and Gerald D Fasman. Empirical predictions of protein conformation. *Annual review of biochemistry*, 47(1):251–276, 1978.

[27] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic acids research*, 34(2):564–574, 2006.

[28] Rosalind E Franklin and Raymond George Gosling. The structure of sodium thymonucleate fibres. i. the influence of water content. *Acta Crystallographica*, 6(8-9):673–677, 1953.

[29] Rosalind E Franklin and Raymond George Gosling. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature*, 172(4369):156–157, 1953.

[30] James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[31] RE Franklin and RG Gosling. The structure of sodium thymonucleate fibres. iii. the three-dimensional patterson function. *Acta Crystallographica*, 8(3):151–156, 1955.

[32] Richard E Dickerson. Definitions and nomenclature of nucleic acid structure components. *Nucleic acids research*, 17(5):1797–1803, 1989.

[33] Richard Wing, Horace Drew, Tsunehiro Takano, Chris Broka, Shoji Tanaka, Keiichi Itakura, and Richard E Dickerson. Crystal structure analysis of a complete turn of b-dna. *Nature*, 287(5784):755–758, 1980.

[34] William S Klug, Michael R Cummings, et al. *Essentials of genetics.* Number Ed. 3. Prentice-Hall Inc., 1999.

[35] Kathleen A Marquis, Briana M Burton, Marcelo Nollmann, Jerod L Ptacin, Carlos Bustamante, Sigal Ben-Yehuda, and David Z Rudner. Spoiiie strips proteins off the dna during chromosome translocation. *Genes & development*, 22(13):1786–1795, 2008.

[36] Eckhard Jankowsky, Christian H Gross, Stewart Shuman, and Anna Marie Pyle. Active disruption of an rna-protein interaction by a dexh/d rna helicase. *Science*, 291(5501):121–125, 2001.

[37] Jae Young Lee and Wei Yang. Uvrd helicase unwinds dna one base pair at a time by a two-part power stroke. *Cell*, 127(7):1349–1360, 2006.

[38] Sameer S Velankar, Panos Soultanas, Mark S Dillingham, Hosahalli S Subramanya, and Dale B Wigley. Crystal structures of complexes of pcra dna helicase with a dna substrate indicate an inchworm mechanism. *Cell*, 97(1):75–84, 1999.

[39] Eric J Enemark and Leemor Joshua-Tor. Mechanism of dna translocation in a replicative hexameric helicase. *Nature*, 442(7100):270, 2006.

[40] Artem Y Lyubimov, Melania Strycharska, and James M Berger. The nuts and bolts of ring-translocase structure and mechanism. *Current opinion in structural biology*, 21(2):240–248, 2011.

[41] Nathan D Thomsen and James M Berger. Running in reverse: the structural basis for translocation polarity in hexameric helicases. *Cell*, 139(3):523–534, 2009.

[42] Lutz Gissmann, Lutz Wolnik, Hans Ikenberg, Ursula Koldovsky, Hans Georg Schnürch, and Harald Zur Hausen. Human papillomavirus types 6 and 11 dna sequences in genital and laryngeal papillomas and in some cervical cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 80(2):560, 1983.

[43] Harald zur Hausen. Papillomaviruses in the causation of human cancers—a brief historical account. *Virology*, 384(2):260–265, 2009.

[44] Hans-Ulrich Bernard. The clinical importance of the nomenclature, evolution and taxonomy of human papillomaviruses. *Journal of clinical virology*, 32:1–6, 2005.

[45] Shona Hilton, Kate Hunt, Mairi Langan, Helen Bedford, and Mark Petticrew. Newsprint media representations of the introduction of the hpv vaccination programme for cervical cancer prevention in the uk (2005–2008). *Social science & medicine*, 70(6):942–950, 2010.

[46] F T Cutts, S Franceschi, S Goldie, X Castellsague, S De Sanjose, G Garnett, WJ Edmunds, P Claeys, KL Goldenthal, DM Harper, et al. Human papillomavirus and hpv vaccines: a review. *Bulletin of the World Health Organization*, 85:719–726, 2007.

[47] World Health Organization et al. Human papillomavirus vaccines: Who position paper. *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, 84(15):118–131, 2009.

[48] Tim Palmer, Lynn Wallace, Kevin G Pollock, Kate Cuschieri, Chris Robertson, Kim Kavanagh, and Margaret Cruickshank. Prevalence of cervical disease at age 20 after immunisation with bivalent hpv vaccine at age 12-13 in scotland: retrospective population study. *BMJ*, 365, 2019.

[49] Van G Wilson, Michael West, Kelly Woytek, and Dandapani Rangasamy. Papillomavirus e1 proteins: form, function, and features. *Virus genes*, 24(3):275–290, 2002.

[50] Giuseppe Borzacchiello and Franco Roperto. Bovine papillomaviruses, papillomas and cancer in cattle. *Veterinary research*, 39(5):1, 2008.

[51] Yukiko Maeda, Tomoyuki Shibahara, Yoshihiro Wada, Koichi Kadota, Toru Kanno, Ikuo Uchida, and Shinichi Hatama. An outbreak of teat papillomatosis in cattle caused by bovine papilloma virus (bpv) type 6 and unclassified bpvs. *Veterinary microbiology*, 121(3-4):242–248, 2007.

[52] Sihwa Joo, Bong H Chung, Mina Lee, and Tai H Ha. Ring-shaped replicative helicase encircles double-stranded dna during unwinding. *Nucleic acids research*, 47(21):11344–11354, 2019.

[53] John Doorbar. The papillomavirus life cycle. *Journal of clinical virology*, 32:7–15, 2005.

[54] Juhan Sedman and Arne Stenlund. The papillomavirus e1 protein forms a dna-dependent hexameric complex with atpase and dna helicase activities. *Journal of virology*, 72(8):6893–6897, 1998.

[55] Maria Manosas, Xu Guang Xi, David Bensimon, and Vincent Croquette. Active and passive mechanisms of helicases. *Nucleic acids research*, 38(16):5518–5526, 2010.

[56] Shixin Liu, Gheorghe Chistol, and Carlos Bustamante. Mechanical operation and intersubunit coordination of ring-shaped molecular motors: insights from single-molecule studies. *Biophysical journal*, 106(9):1844–1858, 2014.

[57] Cyril M Sanders, Oleg V Kovalevskiy, Dmytro Sizov, Andrey A Lebedev, Michail N Isupov, and Alfred A Antson. Papillomavirus e1 helicase assembly maintains an asymmetric state in the absence of dna and nucleotide cofactors. *Nucleic acids research*, 35(19):6451–6457, 2007.

[58] Howard C Berg. *E. coli in Motion*. Springer Science & Business Media, 2008.

[59] Nicola E Farthing, Rachel C Findlay, Jan F Jikeli, Pegine B Walrad, Martin A Bees, and Laurence G Wilson. Simultaneous two-color imaging in digital holographic microscopy. *Optics express*, 25(23):28489–28500, 2017.

[60] Binh Nguyen, Yerdos Ordabayev, Joshua E Sokoloski, Elizabeth Weiland, and Timothy M Lohman. Large domain movements upon uvrd dimerization and helicase activation. *Proceedings of the National Academy of Sciences*, 114(46):12178–12183, 2017.

[61] Yerdos A Ordabayev, Binh Nguyen, Alexander G Kozlov, Haifeng Jia, and Timothy M Lohman. Uvrd helicase activation by mutl involves rotation of its 2b subdomain. *Proceedings of the National Academy of Sciences*, 116(33):16320–16325, 2019.

[62] Yerdos A Ordabayev, Binh Nguyen, Anita Niedziela-Majka, and Timothy M Lohman. Regulation of uvrd helicase activity by mutl. *Journal of molecular biology*, 430(21):4260–4274, 2018.

[63] Liisa T Chisty, Christopher P Toseland, Natalia Fili, Gregory I Mashanov, Mark S Dillingham, Justin E Molloy, and Martin R Webb. Monomeric pcra helicase processively unwinds plasmid lengths of dna in the presence of the initiator protein repd. *Nucleic acids research*, 41(9):5010–5023, 2013.

[64] Colin P Guy, John Atkinson, Milind K Gupta, Akeel A Mahdi, Emma J Gwynn, Christian J Rudolph, Peter B Moon, Ingeborg C van Knippenberg, Chris J Cadman, Mark S Dillingham, et al. Rep provides a second motor at the replisome to promote duplication of protein-bound dna. *Molecular cell*, 36(4):654–666, 2009.

[65] Samuel G Mackintosh and Kevin D Raney. Dna unwinding and protein displacement by superfamily 1 and superfamily 2 helicases. *Nucleic acids research*, 34(15):4154–4159, 2006.

[66] Timothy M Lohman, K Chao, JM Green, S Sage, and GT Runyon. Large-scale purification and characterization of the escherichia coli rep gene product. *Journal of Biological Chemistry*, 264(17):10139–10147, 1989.

[67] Sua Myong, Ivan Rasnik, Chirlmin Joo, Timothy M Lohman, and Taekjip Ha. Repetitive shuttling of a motor protein on dna. *Nature*, 437(7063):1321–1325, 2005.

[68] Wei Cheng, John Hsieh, Katherine M Brendza, and Timothy M Lohman. E. coli rep oligomers are required to initiate dna unwinding in vitro. *Journal of molecular biology*, 310(2):327–350, 2001.

[69] Sinan Arslan, Rustem Khafizov, Christopher D Thomas, Yann R Chemla, and Taekjip Ha. Engineering of a superhelicase through conformational control. *Science*, 348(6232):344–347, 2015.

[70] Katherine M Brendza, Wei Cheng, Christopher J Fischer, Marla A Chesnik, Anita Niedziela-Majka, and Timothy M Lohman. Autoinhibition of escherichia coli rep monomer helicase activity by its 2b subdomain. *Proceedings of the National Academy of Sciences*, 102(29):10076–10081, 2005.

[71] Wei Cheng, Katherine M Brendza, George H Gauss, Sergey Korolev, Gabriel Waksman, and Timothy M Lohman. The 2b domain of the escherichia coli rep protein is not required for dna helicase activity. *Proceedings of the National Academy of Sciences*, 99(25):16006–16011, 2002.

[72] Jan Gert Brüning, Jamieson A.L. Howard, Kamila K. Myka, Mark S. Dillingham, and Peter McGlynn. The 2b subdomain of rep helicase links translocation along dna with protein displacement. *Nucleic Acids Research*, 46(17):8917–8925, 2018.

[73] Haifeng Jia, Sergey Korolev, Anita Niedziela-Majka, Nasib K Maluf, George H Gauss, Sua Myong, Taekjip Ha, Gabriel Waksman, and Timothy M Lohman. Rotations of the 2b sub-domain of e. coli uvrd helicase/translocase coupled to nucleotide and dna binding. *Journal of molecular biology*, 411(3):633–648, 2011.

[74] Theodor Förster and O Sinanoglu. Modern quantum chemistry. *Academic Press, New York*, 3:93–137, 1965.

[75] Everett A Lipman, Benjamin Schuler, Olgica Bakajin, and William A Eaton. Single-molecule measurement of protein folding kinetics. *Science*, 301(5637):1233–1235, 2003.

[76] Jinglei Hu, Mingcheng Yang, Gerhard Gompper, and Roland G Winkler. Modelling the mechanics and hydrodynamics of swimming e. coli. *Soft matter*, 11(40):7867–7876, 2015.

[77] Berni Julian Alder and Thomas Everett Wainwright. Phase transition for a hard sphere system. *The Journal of chemical physics*, 27(5):1208–1209, 1957.

[78] Aneesur Rahman. Correlations in the motion of atoms in liquid argon. *Physical review*, 136(2A):A405, 1964.

[79] Aneesur Rahman and Frank H Stillinger. Molecular dynamics study of liquid water. *The Journal of Chemical Physics*, 55(7):3336–3359, 1971.

[80] Frank H Stillinger and Aneesur Rahman. Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics*, 60(4):1545–1557, 1974.

[81] A Rahman, FH Stillinger, and HL Lemberg. Study of a central force model for liquid water by molecular dynamics. *The Journal of Chemical Physics*, 63(12):5223–5230, 1975.

[82] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[83] Ramu Anandakrishnan, Boris Aguilar, and Alexey V Onufriev. H++ 3.0: automating p k prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research*, 40(W1):W537–W541, 2012.

[84] John Edward Jones. On the determination of molecular fields.—i. from the variation of the viscosity of a gas with temperature. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 106(738):441–462, 1924.

[85] John Edward Jones. On the determination of molecular fields.—ii. from the equation of state of a gas. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 106(738):463–477, 1924.

[86] John E Lennard-Jones. Cohesion. *Proceedings of the Physical Society (1926-1948)*, 43(5):461, 1931.

[87] Charles Augustin de Coulomb. Premier mémoire sur l'electricité et le magnétisme. *Histoire de l'Académie Royale des Sciences*, 569:569–577, 1785.

[88] Charles Augustin de Coulomb. Second mémoire sur l'electricité et le magnétisme. *Histoire de l'Académie Royale des Sciences*, 579:578–611, 1785.

[89] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

[90] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

[91] SJ Weiner, PA Kollman, DT Nguyen, and DA Case. Force field calculations in computational chemistry. *J. Comput. Chem*, 7:230, 1986.

[92] Péter Várnai and Krystyna Zakrzewska. Dna and its counterions: a molecular dynamics study. *Nucleic acids research*, 32(14):4269–4280, 2004.

[93] Alberto Pérez, Iván Marchán, Daniel Svozil, Jiri Sponer, Thomas E Cheatham III, Charles A Laughton, and Modesto Orozco. Refinement of the amber force field for nucleic acids: improving the description of α/γ conformers. *Biophysical journal*, 92(11):3817–3829, 2007.

[94] Ivan Ivani, Pablo D Dans, Agnes Noy, Alberto Pérez, Ignacio Faustino, Adam Hospital, Jürgen Walther, Pau Andrio, Ramon Goñi, Alexandra Balaceanu, et al. Parmbsc1: a refined force field for dna simulations. *Nature methods*, 13(1):55, 2016.

[95] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with

constraints: molecular dynamics of n-alkanes. *Journal of computational physics*, 23(3):327–341, 1977.

[96] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[97] D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, III T.E. Cheatham, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York, and P.A. Kollman. Amber 2018 reference manual. *University of California: San Francisco, CA, USA*, pages 1–927, 2018.

[98] Juan Jose Nogueira. Chapter 3: Periodic boundary conditions, temperature and pressure, 2020.

[99] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*, volume 1. Elsevier, 2001.

[100] Bojan Zagrovic and Vijay Pande. Solvent viscosity dependence of the folding rate of a small protein: distributed computing study. *Journal of computational chemistry*, 24(12):1432–1436, 2003.

[101] Hai Nguyen, Alberto Perez, Sherry Bermeo, and Carlos Simmerling. 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, 2015.

[102] Yen-Lin Lin, Alexey Aleksandrov, Thomas Simonson, and Benoit Roux.

An overview of electrostatic free energy computations for solutions and proteins. *Journal of chemical theory and computation*, 10(7):2690–2709, 2014.

[103] Chris Cramer. Solvation (condensed phase) models - implicit solvent - electronstatics, 2014.

[104] Modesto Orozco and F Javier Luque. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chemical Reviews*, 100(11):4187–4226, 2000.

[105] Hai Nguyen, Daniel R Roe, and Carlos Simmerling. Improved generalized born solvent model parameters for protein simulations. *Journal of chemical theory and computation*, 9(4):2020–2034, 2013.

[106] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[107] WF Van Gunsteren and Herman JC Berendsen. Algorithms for macromolecular dynamics and constraint dynamics. *Molecular Physics*, 34(5):1311–1327, 1977.

[108] Ross Walker. Amber tutorials: Running minimization and molecular dynamics. https://ambermd.org/tutorials/basic/tutorial1/section5.htm, 2015. Accessed: 2021-08-03.

[109] David A Case, RM Betz, DS Cerutti, TE Cheatham III, TA Darden, RE Duke, TJ Giese, H Gohlke, AW Goetz, N Homeyer, et al. Amber 2016 reference manual. *University of California: San Francisco, CA, USA*, pages 1–923, 2016.

[110] David E Smith and Liem X Dang. Computer simulations of nacl association in polarizable water. *The Journal of Chemical Physics*, 100(5):3757–3766, 1994.

[111] Rodrigo Galindo-Murillo, James C Robertson, Marie Zgarbova, Jiri Sponer, Michal Otyepka, Petr Jurecka, and Thomas E Cheatham III. Assessing the

current state of amber force field modifications for dna. *Journal of chemical theory and computation*, 12(8):4114–4127, 2016.

[112] Yi Isaac Yang, Qiang Shao, Jun Zhang, Lijiang Yang, and Yi Qin Gao. Enhanced sampling in molecular dynamics. *The Journal of chemical physics*, 151(7):070902, 2019.

[113] Rafael C Bernardi, Marcelo CR Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

[114] Robert H Swendsen and Jian-Sheng Wang. Replica monte carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607, 1986.

[115] Ayori Mitsutake, Yuji Sugita, and Yuko Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Peptide Science: Original Research on Biomolecules*, 60(2):96–123, 2001.

[116] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

[117] Lucía Álvarez, Ariel Lewis-Ballester, Adrián Roitberg, Darío A Estrin, Syun-Ru Yeh, Marcelo A Marti, and Luciana Capece. Structural study of a flexible active site loop in human indoleamine 2, 3-dioxygenase and its functional implications. *Biochemistry*, 55(19):2785–2793, 2016.

[118] Xiaorong Liu, Zhiguang Jia, and Jianhan Chen. Enhanced sampling of intrinsic structural heterogeneity of the bh3-only protein binding interface of bcl-xl. *The Journal of Physical Chemistry B*, 121(39):9160–9168, 2017.

[119] Richard Lavery, M Moakher, John H Maddocks, D Petkeviciute, and Krystyna Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *Nucleic acids research*, 37(17):5917–5929, 2009.

[120] Bahar I Bakan A, Meireles LM. Prody: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577.

[121] Daniel R Roe and Thomas E Cheatham III. Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, 9(7):3084–3095, 2013.

[122] Vladimir N Maiorov and Gordon M Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of molecular biology*, 235(2):625–634, 1994.

[123] George Daniel Watson. *Atomistic simulation of interactions between DNA and integration host factor*. PhD thesis, University of York, 2021.

[124] Kristin L Meagher, Luke T Redman, and Heather A Carlson. Development of polyphosphate parameters for use with the amber force field. *Journal of computational chemistry*, 24(9):1016–1025, 2003.

[125] Olof Allner, Lennart Nilsson, and Allessandra Villa. Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.*, 8(4):1493–1502, 2012.

[126] Guillem Portella and Modesto Orozco. Multiple routes to characterize the folding of a small dna hairpin. *Angewandte Chemie*, 122(42):7839–7842, 2010.

[127] Jae Young Lee and Wei Yang. Crystal structure of uvrd-dna binary complex. https://www.rcsb.org/structure/2IS2, 2007.

[128] Zhang Y Kaya C Kaynak B Mikulska-Ruminska K Doruker P Li H Bahar I Zhang S, Krieger JM. Prody 2.0: Increased scale and scope after 10 years of protein dynamics modelling with python. *Bioinformatics*, page 187, 2021.

[129] Rabbiah Manzoor Malik. *Elastic Network Modeling of Human Papilloma Virus Proteins*. PhD thesis, CAPITAL UNIVERSITY, 2021.

[130] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[131] Ali Rana Atilgan, SR Durell, Robert L Jernigan, Melik C Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[132] Ben Ambrose Maria Dienerowitz-Antoinette Alevropoulos-Borrill Agnes Noy Michael Borsch Mahmoud Abdelhamid Timothy D. Craggs Jamieson A. L. Howard, Steven D. Quinn and Mark C. Leake. Molecular-level insights reveal the rapid structural remodelling of single rep helicases. *Manuscript in Preparation*.

[133] Mert Gur, Jeffry D Madura, and Ivet Bahar. Global transitions of proteins explored by a multiscale hybrid methodology: application to adenylate kinase. *Biophysical journal*, 105(7):1643–1652, 2013.

[134] Mao W Liu Y Chennubhotla C Lezon TR Bahar I Bakan A, Dutta A. Evol and prody for bridging protein sequence evolution and structural dynamics. *Bioinformatics*, 30(18):2681–2683.

[135] Paul A Tucker and László Sallai. The aaa+ superfamily—a myriad of motions. *Current opinion in structural biology*, 17(6):641–652, 2007.

[136] Sandrine Castella, Gregg Bingham, and Cyril M Sanders. Common determinants in dna melting and helicase-catalysed dna unwinding by papillomavirus replication protein e1. *Nucleic acids research*, 34(10):3008–3019, 2006.

[137] Sandrine Castella, David Burgin, and Cyril M Sanders. Role of atp hydrolysis in the dna translocase activity of the bovine papillomavirus (bpv-1) e1 helicase. *Nucleic acids research*, 34(13):3731–3741, 2006.

[138] Smita S Patel and Kristen M Picha. Structure and function of hexameric helicases. *Annual review of biochemistry*, 69(1):651–697, 2000.

[139] Panos Soultanas and Dale B Wigley. Dna helicases:'inching forward'. *Current opinion in structural biology*, 10(1):124–128, 2000.

[140] Alison Burgess Hickman and Fred Dyda. Binding and unwinding: Sf3 viral helicases. *Current opinion in structural biology*, 15(1):77–85, 2005.

[141] Jan P Erzberger and James M Berger. Evolutionary relationships and structural mechanisms of aaa+ proteins. *Annu. Rev. Biophys. Biomol. Struct.*, 35:93–114, 2006.

[142] Aimee H Marceau. Functions of single-strand dna-binding proteins in dna replication, recombination, and repair. In *Single-Stranded DNA Binding Proteins*, pages 1–21. Springer, 2012.

[143] Arne Stenlund. Initiation of dna replication: lessons from viral initiator proteins. *Nature reviews Molecular cell biology*, 4(10):777, 2003.

[144] Donald J Crampton, Sourav Mukherjee, and Charles C Richardson. Dna-induced switch from independent to sequential dttp hydrolysis in the bacteriophage t7 dna helicase. *Molecular cell*, 21(2):165–174, 2006.

[145] Sam Yoshua, George Watson, Jamiesson Howard, Victor Velasco-Berrelleza, Mark C Leake, and Agnes Noy. Integration host factor bends and bridges dna in a multiplicity of binding modes with varying specificity. *bioRxiv*, pages 2020–04, 2021.

[146] Shane C Dillon and Charles J Dorman. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nature Reviews Microbiology*, 8(3):185–195, 2010.

[147] Remus T Dame, Fatema-Zahra M Rashid, and David C Grainger. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nature Reviews Genetics*, 21(4):227–242, 2020.

[148] Isaac Wong and Timothy M Lohman. Allosteric effects of nucleotide cofactors on escherichia coli rep helicase&dna binding. *Science*, 256(5055):350–355, 1992.

[149] Casey P Kelly, Christopher J Cramer, and Donald G Truhlar. Aqueous solvation free energies of ions and ion- water clusters based on an accurate value for the absolute aqueous solvation free energy of the proton. *The Journal of Physical Chemistry B*, 110(32):16066–16081, 2006.

[150] Adam Hospital, Ivan Ivani, Pablo D. Dans, Guillem Portella, Modesto Orozco, and Carlos González. How accurate are accurate force-fields for B-DNA? *Nucleic Acids Research*, 45(7):4217–4230, 2017.

[151] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[152] Carl S Ewig, Rajiv Berry, Uri Dinur, Jörg-Rüdiger Hill, Ming-Jing Hwang, Haiying Li, Chris Liang, JON Maple, Zhengwei Peng, Thomas P Stockfisch, et al. Derivation of class ii force fields. viii. derivation of a general quantum mechanical force field for organic compounds. *Journal of computational chemistry*, 22(15):1782–1800, 2001.

[153] Agnes Noy and Ramin Golestanian. The chirality of dna: elasticity cross-terms at base-pair level including a-tracts and the influence of ionic strength. *The Journal of Physical Chemistry B*, 114(23):8022–8031, 2010.

[154] George C Shields, Charles A Laughton, and Modesto Orozco. Molecular dynamics simulations of the d (t a t) triple helix. *Journal of the American Chemical Society*, 119(32):7463–7469, 1997.

[155] Thana Sutthibutpong, Christian Matek, Craig Benham, Gabriel G Slade, Agnes Noy, Charles Laughton, Jonathan P K. Doye, Ard A Louis, and Sarah A Harris. Long-range correlations in the mechanics of small dna circles under topological stress revealed by multi-scale simulation. *Nucleic acids research*, 44(19):9121–9130, 2016.

[156] Pablo D Dans, Linda Danilāne, Ivan Ivani, Tomáš Dršata, Filip Lankaš,

Adam Hospital, Juergen Walther, Ricard Illa Pujagut, Federica Battistini, Josep Lluis Gelpí, et al. Long-timescale dynamics of the drew–dickerson dodecamer. *Nucleic acids research*, 44(9):4052–4066, 2016.

[157] Wei Zhang, Chun Wu, and Yong Duan. Convergence of replica exchange molecular dynamics. *The Journal of chemical physics*, 123(15):154105, 2005.

[158] David L Beveridge and Kevin J McConnell. Nucleic acids: theory and computer simulation, y2k. *Current opinion in structural biology*, 10(2):182–196, 2000.