In Vitro, *In Silico* and *In Vivo* Studies of the Structure and Conformational Dynamics of DNA Polymerase I

A Thesis Presented by

Marko Sustarsic
Magdalen College

in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

University of Oxford
Michaelmas Term 2015

To delo posvečam svoji mami,

za njeno brezpogojno ljubezen.

Abstract

DNA polymerases are a family of molecular machines involved in high-fidelity DNA replication and repair, of which DNA polymerase I (Pol) is one the best-characterized members. Pol is a strand-displacing polymerase responsible for Okazaki fragment synthesis and base-excision repair in bacteria; it consists of three protein domains, which harbour its 5'-3' polymerase, 3'-5' exonuclease and 5' endonuclease activities.

In the first part of the thesis, we use a combination of single-molecule Förster resonance energy transfer (smFRET) and rigid-body docking to probe the structure of Pol bound to its gapped-DNA substrate. We show that the DNA substrate is highly bent in the complex, and that the downstream portion of the DNA is partly unwound. Using all-atom molecular dynamics (MD) simulations, we identify residues in the polymerase important for strand displacement and for downstream DNA binding. Moreover, we use coarse-grained simulations to investigate the dynamics of the gapped-DNA substrate alone, allowing us to propose a model for specific recognition and binding of gapped DNA by Pol.

In the second part of the thesis, we focus on the catalytically important conformational change in Pol that involves the closing of the 'fingers' subdomain of the protein around an incoming nucleotide. We make use of the energy decomposition method (EDM) to predict the stability-determining residues for the closed and open conformations of Pol, and test their relevance by site-directed mutagenesis. We apply the unnatural amino acid approach and a single-molecule FRET assay of Pol fingers-closing, to show that substitutions in the stability-determining residues significantly affect the conformational equilibrium of Pol.

In the final part of the thesis, we attempt to study Pol in its native environment of the living cell. We make use of the recently developed method of internalization by electroporation, and optimize it for organically labelled proteins. We demonstrate the internalization and single-molecule tracking of Pol, and provide preliminary data of intra-molecular FRET in Pol, both at the single-cell and single-molecule levels. Finally, by measuring smFRET within an internalized gapped-DNA construct, we observe DNA binding and bending by endogenous Pol, confirming the physiological relevance of our *in vitro* Pol-DNA structure.

Statement of Authorship

All of the experimental work underlying this thesis is my own, except for the contributions indicated below and at the end of each chapter. The pronoun 'we' is used throughout the thesis instead of the pronoun 'I' for stylistic reasons, although the latter would be more accurate.

The three major contributions to this thesis include:
(i) the work of Tim Craggs towards Pol-DNA structure determination, including sample preparation, and collection and analysis of single-molecule FRET data (Chapter 4)
(ii) the work of Majid Mosayebi, Hendrik Kaju and Jonathan Doye in coarse-grained DNA simulations (Chapter 4)
(iii) the work of Meli Massimilliano and Giorgio Colombo in energy decomposition analysis (Chapter 6)

I have opted to integrate the results from these contributions into the thesis, both because I have been intellectually involved in their production and/or analysis, and because they are intricately connected to the results of my work.

I have authored all of the text in this thesis. Parts of it have previously been published, as indicated at the end of the relevant chapters. All of the figures are my work, unless indicated otherwise; any figure adaptations or reproductions are noted in the figure legends.

Marko Sustarsic, December 2015

Acknowledgements

I would like to thank Achilles Kapanidis, for his supervision over the years, his insightful advice and his creative ideas. I am grateful for his enthusiasm, optimism and faith in my ability, which helped me to get through the difficult times of my DPhil. I also thank Tim Craggs and Louise Aigrain, for being great teammates in the *in vitro* and *in vivo* projects, respectively. Tim, for being involved and enthusiastic, and for caring deeply about my projects, often calling me up during his holiday to discuss yesterday's results. Louise, for introducing me to the world of fluorescent bacteria and providing immense support with the challenges that I encountered there, for cheering me up with art and photography, and for co-starring with me in our very first scientific movie.

Great thanks go to Anne Plochowietz, for her help with microscopy and data analysis, and generally all things physics. I also thank David Bauer, for helping with the unnatural amino acid modifications and for his advice on polymerase-specific issues. Further thanks go to Peter May, for help with R_0 determination, Rob Crawford, for advice on in-cell FRET, Alexandra Tomescu, Nicole Robb and Javier Periz, for help with various biochemical techniques, Geraint Evans and Johannes Hohlbein, for sharing their knowledge of polymerase structure and dynamics, Federico Garza de Leon, Mathew Stracy and Stephan Uphoff, for maintaining microscope set-ups, Pawel Zawadzki, for efforts with the *PolA*- strain, Nicolae Solcan, for providing the recoded *E. coli* strain, and Florence Wagner, for providing competent cells. I also thank Sarah Matthews for being a very helpful technician, as well as all the other 'Kapanidians' for the sociable and playful atmosphere that made spending time in the lab always fun.

On the simulation side, I am thankful to Phil Biggin, for taking in an MD-illiterate student and venturing into the him-unknown world of DNA polymerases. I would also like to thank Laura Domicevica, for introducing me to MD simulations, and Maria Musgaard, for being willing to stay at my computer as long as necessary to solve all my problems with Gromacs and VMD. I am also generally grateful to all members of the SBCB group, for providing both a great working environment and unhealthy amounts of cake.

I thank Giorgio Colombo, Meli Massimilliano, Jonathan Doye, Majid Mosayebi and Hendrik Kaju, for the fruitful collaborations.

Special thanks go to Tim, Maria, Phil, David and Louise, for reading through my thesis chapters and providing useful comments.

Finally, I am grateful to the Wellcome Trust for the generous financial support, without which none of this work would have been possible.

Contents

1 Introduction

DNA polymerase I (Pol) was discovered in 1956 and was the first known of any type of polymerase [1]. It has since served as a model for understanding the general mechanisms of DNA polymerases in both prokaryotes and eukaryotes. A plethora of biochemical and mutagenesis approaches, X-ray crystallography and fluorescence methods have been used to illuminate Pol function in DNA replication and repair pathways, its three-domain architecture, and the high-resolution structure of its DNA-binding and catalytic sites [2, 3]. In addition, Pol conformational dynamics and reaction kinetics have been thoroughly studied, along with the mechanisms that ensure high fidelity of polymerization [4–6]. Nevertheless, important questions remain that have evaded conventional approaches. These include the mechanisms underlying DNA-substrate recognition and enzyme translocation along the DNA, the workings of strand-displacement synthesis, and the structural and functional communication between the different catalytic domains of Pol.

Many of these questions have been left outstanding because they can only be studied through a synthesis of different biochemical, biophysical and computational approaches. Crystal structures can provide high-resolution snapshots of Pol, and the use of stabilizing substitutions and artificial substrate analogues can aid in capturing their transient conformations at different stages of the reaction cycle. However, static structures can only give hints on protein conformational dynamics, and cannot study reaction kinetics, stressing

the importance of complementary approaches such as nuclear magnetic resonance (NMR) [7] and single-molecule fluorescence. Even with the help of these methods, it remains difficult to study residue-specific dynamics at nano- or microsecond time resolution, necessitating the application of the 'computational microscope', such as provided by *in silico* molecular dynamics simulations. In this thesis, we address some of the outstanding questions in understanding Pol mechanisms by making extensive use of the wealth of the available structural information, and applying state-of-the-art single-molecule Förster resonance energy transfer (FRET) and molecular dynamics approaches.

As part of DNA replication and repair, Pol recognizes DNA substrates containing gaps of one or several nucleotides. Despite the multitude of crystal structures available for *E. coli* Pol and its homologs in complex with upstream DNA [3, 8], it has not been possible to capture the position of downstream DNA in the polymerase site of Pol [9]. As a result, the exact mechanism of strand displacement synthesis in Pol in unclear, and it is not known what the degree of unwinding is in downstream DNA, or where the non-template strand of the downstream DNA is channelled [9, 10]. Furthermore, this lack of structural information has left the binding mechanism of Pol unresolved. The latter is of particular interest because it is based on sequence-independent substrate recognition, and this type of protein-DNA recognition is significantly less well understood than the sequence-specific recognition [11]. In the first part of this thesis, we therefore use single-molecule FRET in combination with rigid-body docking to probe the structure of Pol bound to gapped DNA. We further test the structure and dynamics of the gapped-DNA substrate alone using coarse-grained simulations, to infer on the Pol binding mechanism. We also perform molecular dynamics simulations on the Pol-DNA complex, and investigate the degree of unwinding in downstream DNA, and the mechanism of strand separation.

During DNA synthesis itself, at each cycle of nucleotide incorporation, Pol undergoes a conformational change that involves the closing of its fingers subdomain towards the palm subdomain. This fingers-closing transition has been studied extensively using structural and fluorescence methods, and its sensitivity to the binding of the DNA and nucleotides is well understood [8, 12, 13]. However, it is not clear what residues in Pol contribute to

the structural integrity of the open and closed conformations of Pol, and how the delicate balance between their energetic stabilities is maintained. In the second part of the thesis, we therefore make use of biased molecular dynamics simulations and the energy decomposition method (EDM), to identify the stability-conferring residues in Pol. We test the effect of these residues on Pol conformational equilibrium by site-directed mutagenesis, and assay Pol variants using single-molecule FRET. We also explore several fluorescence labelling approaches, in order to achieve the labelling specificity required to accurately assess the conformational equilibria.

We note that *in silico* approaches are compromised by a number of assumptions and simplifications in their description of the system under study, highlighting the need for simulation results to be verified with *in vitro* data, which we aim to do throughout the thesis. However, it is just as important to understand that *in vitro* approaches are also based on a reductionist philosophy that assumes that the component under study will behave identically in isolation and in its native environment, such as the living cell. This assumption is particularly problematic because of the noted effects of the cellular environment on macromolecular structure and function, such as due to the high cytosol viscosity, the spatial organization of the cell and the regulatory interactions arising from other cellular components [14]. Hence, it becomes necessary to confirm any behaviours or mechanisms deduced from *in vitro* data in the relevant *in vivo* context. Equally, because experiments *in vivo* are technically difficult, and the accuracy of results is compromised by biological noise, *in vitro* data are needed for their correct interpretation.

In the last part of the thesis, we therefore translate our *in vitro* studies of Pol to live bacteria, noting that very few studies have previously attempted to do so [15]. We set out to develop methods for electroporation-based cell internalization of organically labelled Pol, and establish single-molecule tracking and single-molecule FRET capabilities in cells. These capabilities would open the door to a number of studies not previously possible, such as probing the controversial issue of domain arrangement in full-length Pol [16], measuring the rates of DNA binding and repair at long time scales, and following conformational dynamics of Pol during catalysis. As a proof of concept, we attempt to ap-

ply electroporation-based internalization and in-cell smFRET to probe the physiological structure of the polymerase domain of Pol. We further test for the existence of the Pol-DNA species observed *in vitro*, in order to establish if the suggested mechanism of DNA recognition by Pol is relevant in the context of the living cell.

The thesis is structured as follows: in the first introductory chapter (Chapter 2), we provide a primer on the biological roles, the molecular structure and the catalytic mechanisms of Pol. In the second introductory chapter (Chapter 3), we explain the principles behind the two main methods used in the thesis: single-molecule fluorescence microscopy and FRET, and molecular dynamics simulations. The three different projects are then presented in the five subsequent chapters, each of which includes its own introductory and discussion sections. Pol-DNA structure is probed in Chapters 4 and 5, Pol stability and conformational equilibrium are explored in Chapter 6, and the in-cell studies are presented in Chapters 7 and 8. Finally, we summarize the thesis and reflect on the synthesis of the different approaches used in studying the structure and function of Pol in Chapter 9.

2
DNA Polymerase I

In this chapter, we review the literature to present an up-to-date view of the structure and function of DNA polymerase I. We summarize the biological roles of Pol, its high-resolution structure and its mechanisms of polymerization and proofreading. We focus particularly on Pol-DNA interactions and Pol conformational changes during polymerization, as these form the basis for the work presented in Chapters 4, 5 and 6. We also highlight a key study of Pol diffusion and repair in live cells, which we build upon in our in-cell studies in Chapters 7 and 8.

2.1 DNA polymerases

In order for genetic information to be propagated from a mother cell to its daughter cells, chromosomal DNA needs to be replicated efficiently and accurately at each cell division. In addition, DNA is prone to various types of damage, which has to be repaired to ensure cell survival and the conservation of genetic information. DNA polymerases are a group of enzymes that are responsible for DNA synthesis and repair in both prokaryotes and eukaryotes, and are extremely diverse in terms of their size and complexity [17]. Based on their amino-acid sequence homologies, they are often grouped into seven families: A, B, C, D, X, Y and RT. Despite the size and sequence diversity, many polymerases share

very similar active site architectures and employ the same mechanism of catalysis, and different polymerase classes have been shown to complement each other *in vivo* [2]. A valuable implication of this similarity is that studies of the structure and function of the simpler polymerases are likely to be generally important for our understanding of all DNA polymerases.

DNA polymerase I (Pol) is a member of the A-family, and is the most abundant polymerase in *E. coli* [3]. It was the first polymerase to be discovered, and remains to be one of the best studied polymerases. It is not the main replicative DNA polymerase in bacteria (this role is carried out by DNA polymerase III), but it has important functions in the lagging-strand DNA synthesis and base-excision DNA repair. The enzyme is encoded by the *polA* gene in *E. coli*, comprises 928 amino acids in a single polypeptide chain, and is 103 kDa large. It consists of three domains: 5'-3' DNA polymerase, 3'-5' exonuclease and 5'-endonuclease, each of which hosts a different catalytic activity. We review the biological functions and the domain architecture of Pol in more detail in the following two sections.

2.2 Biological functions

The first biological function of Pol is in DNA replication. DNA replication proceeds through a semi-conservative mechanism, whereby each parental strand of the DNA duplex serves as a template for the synthesis of a new strand. Because DNA polymerases can only function in the 5'-3' direction, and the two strands of the DNA duplex have different directionality, only one strand (the leading strand) can be synthesised in a continuous fashion [18]. The opposite strand (the lagging strand) is instead synthesised discontinuously, in a series of fragments known as the Okazaki fragments, which are 1000-2000 nucleotides long (Figure 2.1). This process relies on a DNA primase, which produces a short (~10-nucleotide long) RNA primer that allows DNA polymerase III to assemble and initiate synthesis. Pol III dissociates as it encounters the 5' end of the previous Okazaki fragment, at which point Pol recognizes the gapped DNA substrate and elongates it. In this process, Pol displaces and finally also excises the previous RNA primer, thus replacing the RNA sequence with

its DNA counterpart. Finally, a DNA ligase reseals the resulting nick in the DNA.

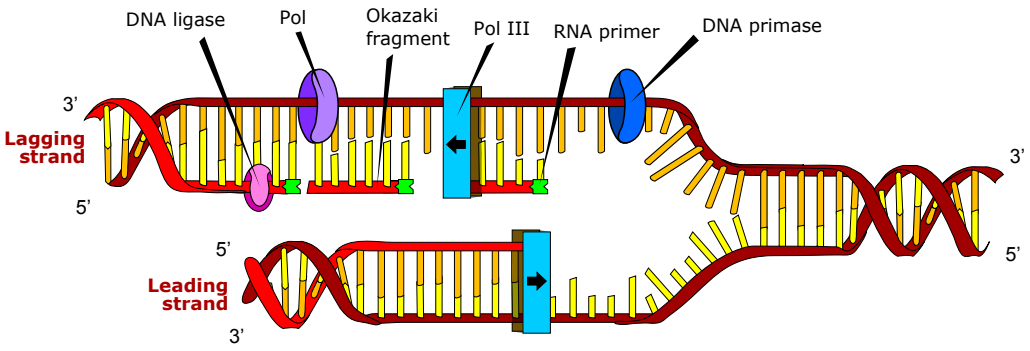


Figure 2.1: *Role of Pol in DNA replication. In each cycle of lagging strand synthesis (top), a DNA primase (blue torus) produces an RNA primer (green block), allowing Pol III (cyan box) to bind and replicate a short fragment of DNA. The new fragment is then processed by Pol I (purple torus) and a DNA ligase (pink oval). Leading strand synthesis is also shown, and the direction of movement of the two polymerases indicated with arrows. Other components of the replication fork are omitted for clarity. Adapted from Wikimedia commons.*

The second function of Pol is in the base-excision repair (BER) pathway of DNA repair. BER is a conserved pathway that processes lesions such as deoxidized, alkylated and deaminated bases, including uracil bases resulting from cytosine deamination [19]. A number of enzymes are involved in BER that recognize and repair the DNA in a series of sequential steps (Figure 2.2), whilst preventing accumulation of toxic intermediates. The damaged bases are removed by DNA glycosylases, which catalyse the hydrolysis of the N-glycosidic bonds linking the base to the sugar backbone. This results in sites without a base in the DNA, called apurinic / apyrimidinic (AP) sites, which are specifically recognized by AP endonucleases or AP lyases. The latter hydrolyse the phosphodiester bond 5' to the AP site, producing a nick in the DNA that is further processed to a gap by a 3'-phosphodiesterase. The gapped DNA is finally recognized by Pol I, which fills in the gap and displaces the 5' strand, leaving a DNA nick farther downstream. A DNA ligase then reseals the nick and restores the normal DNA structure.

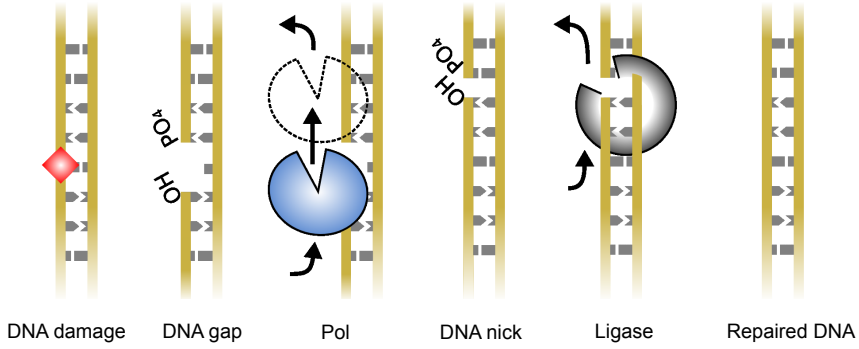


Figure 2.2: *Role of Pol in DNA repair. Following DNA-damage recognition and processing, Pol fills in the resulting DNA gap, leaving behind a DNA nick that is ligated by a DNA ligase. Adapted from reference [15].*

2.3 Domain architecture

E. coli Pol has three functional domains encoded by the same polypeptide: the C-terminal 5'-3' polymerase, the central 3'-5' exonuclease, and the N-terminal 5' endonuclease domain[1]. Limited proteolysis with the protease subtilisin yields two fragments, the larger of which contains the polymerase and 5'-3' exonuclease domains and is commonly known as the Klenow fragment in *E. coli*. The polymerase domain is the main functional domain and is responsible for strand-displacement synthesis during DNA replication and repair, as outlined above. The 5'-3' exonuclease domain has a proofreading function and serves to remove any errors made by the polymerase, whereas the 5' endonuclease domain functions to excise the RNA or DNA flap resulting from Okazaki fragment processing or base excision repair, respectively [2].

Crystal structures are available of both the *E. coli* Klenow fragment [21–24] and of other homologues, particularly of the thermophile *Bacillus stearothermophilus (Bst)* polymerase [8, 25, 26]. In all structures, the polymerase domain resembles the shape of a cupped human right hand, with the fingers, palm and thumb subdomains (Figure 2.3a). The fingers subdomain is responsible for binding the incoming nucleotide and the single-stranded DNA template, the palm subdomain harbours the catalytic residues, and the thumb sub-

[1]This domain is often referred to in the literature as the 5'-3' exonuclease, but the name is inaccurate because the domain is functionally an endonuclease [2, 20].

domain acts in positioning the duplex DNA and ensuring high processivity [27]. Whilst the sequences in distantly related Pol homologs vary significantly, the topologies and the three-dimensional structures of their polymerase domains are nearly identical [3]. The 3'-5' exonuclease domain extends from the palm domain and faces away from the hand structure of the polymerase domain. In some homologues such as the *Bst* and *Thermus aquaticus (Taq)* polymerases, parts of the 3'-5' exonuclease domain are missing, resulting in the loss of the 3'-5' exonuclease activity in these polymerases [25, 28, 29].

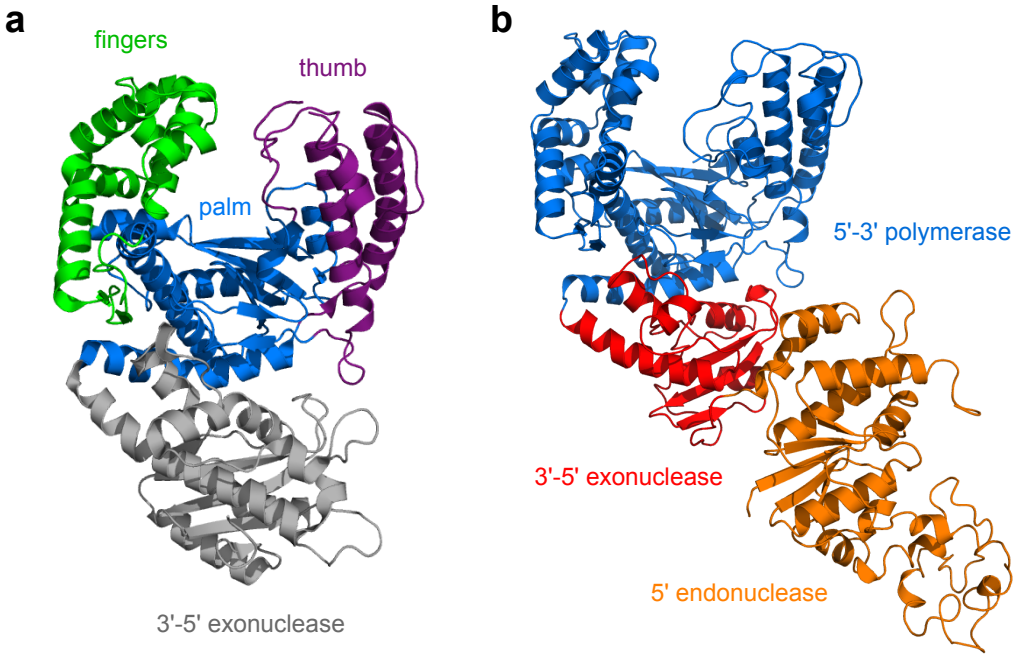


Figure 2.3: *Domain architecture of Pol.* ***(a)*** *Structure of the E. coli Klenow fragment (PDB code 1KLN, with DNA removed, reference [23]), comprising the 5'-3' polymerase (coloured) and 3'-5' exonuclease domains (grey). The fingers, palm and thumb subdomains of the polymerase domain are shown in green, blue and purple, respectively.* ***(b)*** *Structure of the Thermus aquaticus full-length polymerase. The 5'-3' polymerase, 3'-5' exonuclease and 5' endonuclease domains are coloured blue, red and orange respectively. The structure shown here is the one in the elongated conformation, with the 5' endonuclease domain extended out into solution (PDB code 1TAQ, reference [28]).*

Although no structure of the full-length enzyme is available from *E. coli*, several structures have been solved of the thermophilic homologue *Taq* polymerase. In one set of struc-

tures, the enzyme is seen to adopt an elongated conformation, with a very small surface area of contact between the 3'-5' exonuclease and 5' endonuclease domains (Figure 2.3b) [28, 30]. However, in a different structure, solved in the presence of an antibody fragment, the polymerase adopts a more compact conformation, with the 5' endonuclease folded up against the polymerase domain [31]. The active site of the 5'-endonuclease is located ~70 Å and ~38 Å away from the polymerase active site in the elongated and compact structures, respectively, and it is not obvious from either of the structures how the two active sites can work in concert. Analytical ultracentrifugation and small-angle X-ray scattering experiments suggest that the conformation of the full-length polymerase in solution is more consistent with the elongated structure [16].

2.4 DNA polymerization

In order to allow processive DNA polymerization, Pol must first associate stably with a primer-template DNA substrate. Catalysis then involves a number of cycles of deoxyribonucleoside triphosphate (dNTP) incorporation, each of which consists of dNTP binding, phosphodiester bond formation with the primer terminus, and pyrophosphate (PPi) release (Figure 2.4) [3]. A number of conformational changes occur in both the polymerase and the DNA during the steps of Pol-DNA binding, dNTP binding and PPi release. Finally, Pol must translocate towards the new primer terminus to initiate a new cycle of incorporation. We describe each of these steps in more detail in the following sections, focusing particularly on Pol-DNA interactions and the conformational dynamics of Pol.

2.4.1 DNA binding

It was initially thought that the primer-template DNA approached the polymerase active site from the rear side of the enzyme, and was channelled through the cleft formed by the fingers and thumb subdomains [21, 23]. However, structures of the *Bst* and *Taq* polymerases in complex with DNA showed that the DNA was instead bound at the cleft between the thumb and the 3'-5' exonuclease domains (Figure 2.5a) [26, 30]. Compar-

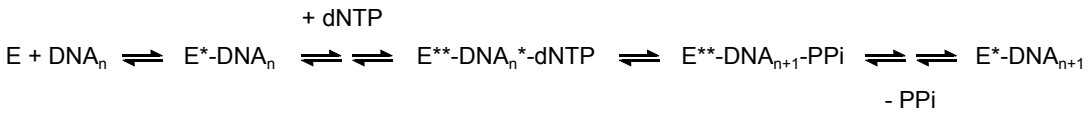


Figure 2.4: *A simplified reaction scheme for DNA polymerization by Pol. The enzyme (E) binds primer-template DNA and undergoes a conformational change in the process (E*). dNTP binding results in further conformational changes in both the enzyme (E**) and the DNA (DNA*), preparing the complex for catalysis. Phosphodiester bond formation results in DNA elongation by one nucleotide (n+1) and PPi formation. In the final steps, PPi is released, the enzyme reverts back to the initial DNA-bound conformation, and translocates to the new template position.*

ison of the apo and binary-complex structures shows that the conformation of the thumb subdomain changes to wrap around the DNA [12, 25, 26]. This conformational change involves a rigid-body rotation of the thumb to open up the DNA-binding cleft, along with the rotation of helices H1 and H2 at the tip of the thumb in the opposite direction, bringing them closer to the DNA. The loop connecting the H1 and H2 helices also changes its conformation and is more stably positioned in the DNA-bound complex.

The tip of the thumb inserts itself into the minor grove of the DNA duplex, forming sequence-unspecific contacts with the sugar-phosphate backbone of the DNA [26]. In this region, the DNA adopts a B-form, which extends from the protein-distal end until the fifth base-pair. In contrast, the first four base pairs at the 3' primer terminus adopt the A-form, causing the minor groove in this region of the DNA to be wider and more shallow, and allowing easy access of the protein side chains to the edges of the bases (Figure 2.5a). As a result, hydrogen bonds can be formed between conserved Pol residues R615 and Q797 (*Bst* numbering; R668 and Q849 in *E. coli*) and the O2 and N3 positions of pyrimidine and purine bases, respectively. The latter are the only groups that present the same pattern of hydrogen-bond acceptors in both G:C and A:T base-pairs, allowing sequence-independent binding [26]. Notably, no specific interactions are observed with the major groove of the DNA, which is generally involved in sequence-specific recognition.

At the active site, the first base-pair is stabilized by stacking interactions with a con-

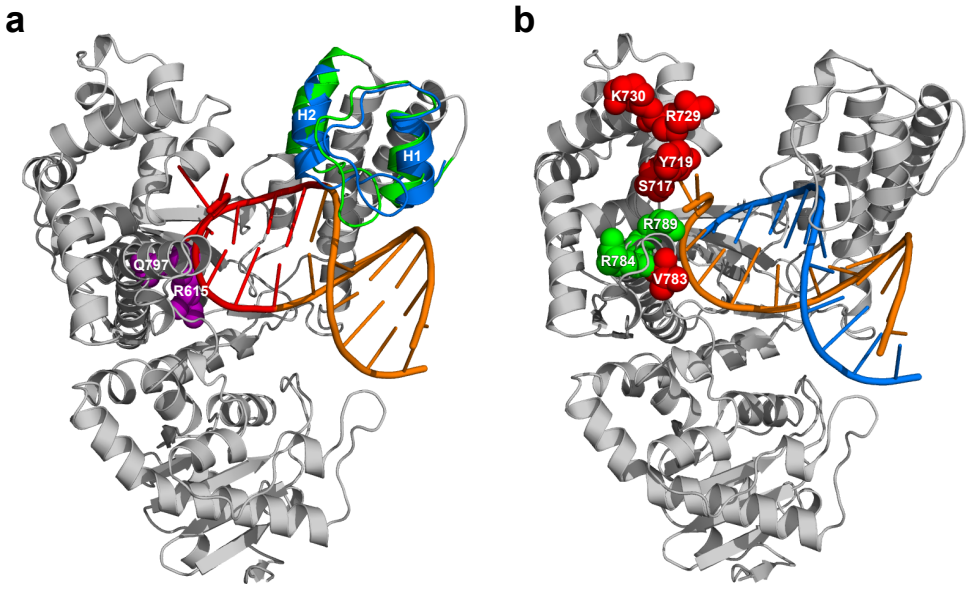


Figure 2.5: *DNA binding in polymerase site of Pol.* ***(a)*** *Structure of Bst Pol bound to primer-template DNA (PDB code 4BDP, reference [26]), with the B-form portion of the DNA in orange and the A-form portion in red. The H1-H2 helix motif is coloured green, and its conformation in the apo structure shown in blue for comparison (PDB code 1XWL, reference [25]). The conserved residues R615 and Q797 are shown as purple spheres.* ***(b)*** *Downstream DNA binding. Residues corresponding to the E. coli residues that were tested by mutagenesis are shown on the Bst structure. The two residues whose substitutions showed an effect on DNA binding (R836 and R841 in E. coli; R784 and R789 in Bst) are coloured green; residues whose substitutions had no effect are coloured red. Note that residue R835 is not conserved in Bst (V783). The template and primer strands of upstream DNA are shown in orange and blue, respectively.*

served tyrosine of the O-helix of the fingers (Y714 in *Bst* and Y766 in *E. coli* [2]), and the template base is flipped out from the helix axis and interacting with the loop between helices O and O1. The next downstream base (+1) is also ordered in some *Bst* structures, and is sometimes seen further rotated and stacking with Y719 (F771 in *E. coli*) [26]. The template strand is additionally stabilized by interactions with S717 and R789 (S769 and R841), although the exact contacts vary between the different structures [10]. The 3' OH group of the primer terminus, in contrast, forms a hydrogen bond to the conserved D830 (D882).

[2]The reader may find it helpful to note that, for almost all Pol residues considered in this thesis, the corresponding *E. coli* residue number can be obtained by adding 52 to the *Bst* residue number.

We describe the mechanistic importance of some of these interactions in the subsequent sections.

Whereas the binding of upstream DNA to the polymerase site is well established, only limited information is available on the binding of downstream DNA beyond the first two template residues (0 and +1). When additional downstream DNA was present in crystallization studies, its electron density was not observed, suggesting that it is flexible and dynamic [9]. Competition assays with radioactive oligonucleotides of different-length overhangs indicate that *E. coli* Pol contacts at least the first four residues of the downstream DNA template (0 to +3) [9]. Photo-crosslinking similarly showed contacts with up to residue +4 of the template strand, with residue +1 contacting the conserved Y766 and F771, and residue +2 contacting F771 [9, 32], consistent with predictions from the crystal structures. Finally, DNase protection assays were used to investigate the effect on primer-template DNA binding of substitutions of both the residues highlighted by the crystal structures (S769, R841 and F771), and selected conserved residues of the positively charged face of the fingers subdomain that were hypothesised to be important (R781, K782, R835 and R836) [9]. Appreciable effect was seen for only two substitutions: R841A resulted in a 9-fold decrease in DNA binding, whereas substitution R836A triggered a 3-fold increase in binding (Figure 2.5b). It was therefore suggested that the template strand of downstream DNA might follow a path across the face of the fingers, interacting weakly with R836 and neighbouring residues.

2.4.2 Nucleotide binding and fingers-closing

Following interaction with the DNA substrate, the polymerase binds a deoxyribonuclotide triphosphate (dNTP) that is complementary to the template base. Crystal structures of wild-type and variant *Bst* polymerases indicate that the fingers subdomain adopts an 'open' conformation in the presence of the DNA substrate only, and a 'closed' conformation when complexed with both the DNA and the complementary nucleotide [8]. Similarly, the *Taq* polymerase has been captured in the open and closed states, depending on the concentra-

tion of dNTP used [12]. Thus, Pol is thought to undergo a 'fingers-closing' transition upon dNTP binding, via a sophisticated induced-fit mechanism that turns the complex into a catalytically active form.

As described above, *Bst* and *Taq* structures in the fingers-open conformation show residue Y714 stacking against the first base-pair and blocking access to the template base, which is flipped out of the helix axis [8, 12]. *Taq* structures at low dNTP concentration further show electron density for the dNTP, which is partly solvent-exposed, with its triphosphate portion bound by conserved aspartate residues via metal coordination [12]. The conformational change from the open to the closed state involves a 40° rotation of the O-helix, which displaces Y714 and allows the template base to rotate back into the helix axis (Figure 2.6). Y714 is now stabilized by a glutamate residue (E658 in *Bst* and E710 in *E. coli*), hydrophobic residues in the O-helix provide a hydrophobic pocket that packs against the ribose and base portions of the dNTP, and positively charged residues additionally interact with its triphosphate portion. In this conformation, the incoming nucleotide forms a Watson-Crick base-pair with the template base, and is optimally positioned for nucleophilic attack from the 3' OH group of the primer terminus.

More recent crystal structures of the *Bst* polymerase in the presence of mismatched nucleotides revealed an additional conformation that appears to be intermediate between the open and closed states, and was termed the 'ajar' conformation [33]. In this conformation, most of the residues in helices O and O1 are nearer to the open state, but the C-terminal part of the O-helix is distorted at a conserved glycine residue, and Y714 is displaced from the insertion site (Figure 2.6). The template base is allowed to move into the helical axis, where it forms a 'wobble' base-pair with the incoming nucleotide, thus misaligning its α-phosphate for attack by the 3' OH of the primer terminus. The displacement also results in several water molecules filling the binding pocket, which are not present in the closed structure of Pol. Thus, the ajar state is proposed to act as a checkpoint, previewing the incoming dNTP for template-base complementarity, and preventing the transition to the fully closed state if a mismatch is detected.

Interpretations of the crystal structures have been corroborated by the results of single-

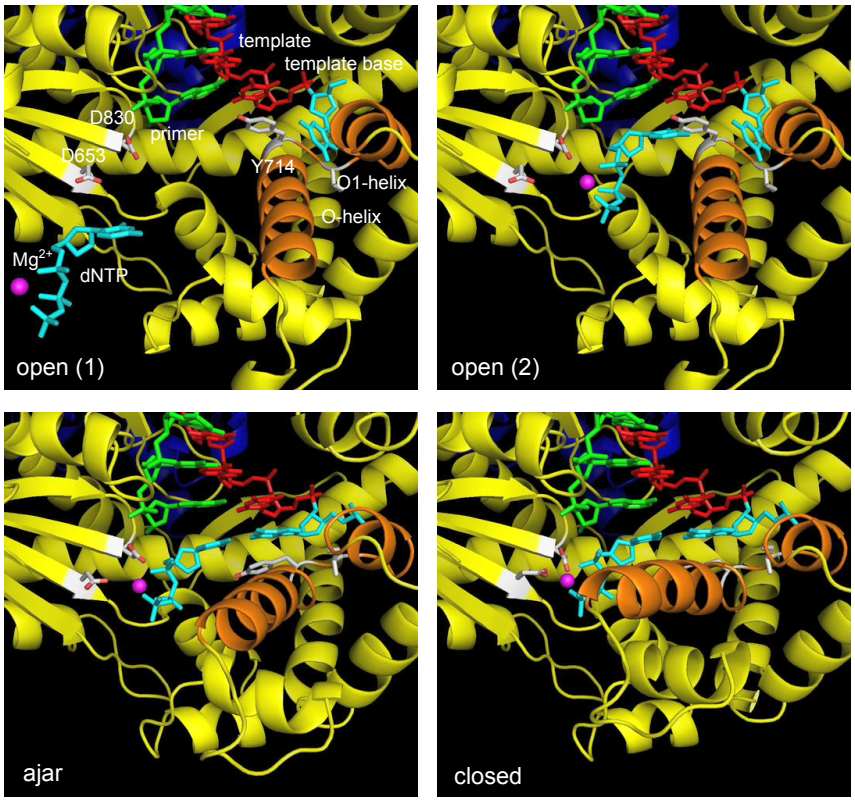


Figure 2.6: *Mechanism of fingers-closing, deduced from open (top), ajar (bottom left) and closed crystal structures (bottom right) of Bst polymerase. The open structure is shown both before (top left) and after dNTP binding (top right). Gradual rotation of the O-helix and displacement of Y714 can be observed, concomitant with template-dNTP base-pairing. Helices O and O1 are highlighted in orange, with Y714 and the catalytic aspartates (D653 and D830) in stick representation, in grey and white, respectively. The DNA and the incoming dNTP are shown in stick representation, with the primer in green, the template in red, and the template base and the dNTP both in cyan. The catalytic magnesium ion is depicted with a pink sphere. Adapted from the supplementary material of reference [33].*

molecule FRET studies on *E. coli* Pol. Fingers-closing was probed by attaching donor and acceptor fluorophores to the mobile segment of the fingers subdomain and to the base of the thumb subdomain, respectively, and measuring the intramolecular distance of diffusing Pol molecules using confocal microscopy (Figure 2.7a) [5, 13]. As expected, in the presence of a primer-template DNA substrate, the major population showed a low FRET value ($E^* = 0.45$), consistent with the open conformation, whereas in the presence of the DNA substrate and the complementary dNTP, the major population indicated a closed con-

formation ($E^* = 0.66$; Figure 2.7b). In addition, in the presence of a mismatched dNTP, a population with a FRET value in between those expected from the open and closed states was observed ($E^* = 0.48$). This conformation was termed the 'partially closed' state and likely corresponds to the ajar crystal structure. Interestingly, unliganded Pol exhibited a heterogeneous FRET distribution, suggestive of conformational dynamics in the absence of ligand. Burst variance analysis, which measures the standard deviation of fluorescence intensity in single-molecule bursts [34], was used to demonstrate that Pol was indeed interconverting between the open and closed conformations on a ~3 ms time scale [13].

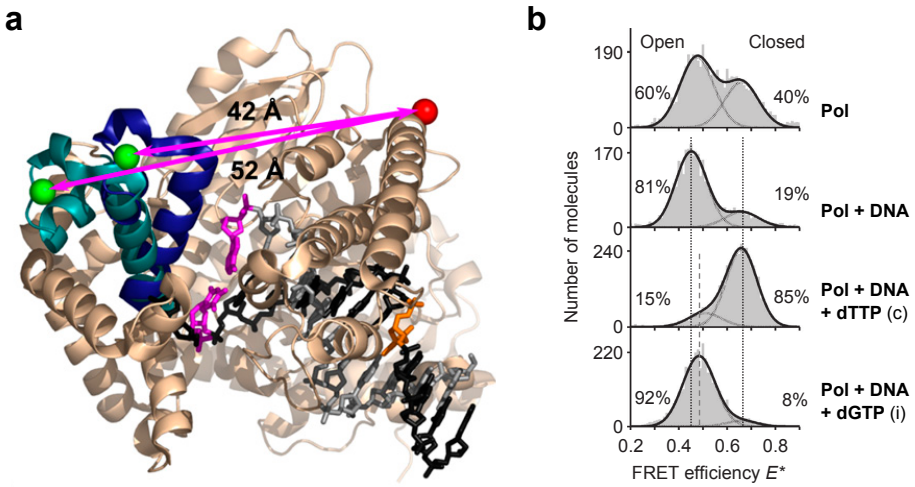


Figure 2.7: *Single-molecule FRET studies of fingers-closing.* ***(a)*** *Superimposition of open and closed structures of Bst polymerase, with the mobile portion of the fingers subdomain shown in teal and dark blue for the open and closed conformations, respectively. Fluorophore attachment positions are shown as green (donor) and red spheres (acceptor), with the arrows indicating the distance between them in the two conformations.* ***(b)*** *Histograms of FRET efficiency E^* for wild-type E. coli Pol, in the unliganded state (top), in the presence of DNA (centre top), DNA and the correct dNTP (centre bottom), or DNA and a mismatched dNTP (bottom). The relative proportions of the open and closed populations, estimated from Gausian fits to the histograms, are indicated. The dotted lines indicate mean E^* values of the low- and high-FRET species in the Pol-DNA binary complex, and the dashed line indicates the mean E^* value of the intermediate-FRET species as populated in the Pol ternary complex with the mismatched dNTP. Adapted from references [13] and [5].*

The observation of the ajar conformation begs the important question of whether this state arises only in the presence of a mismatch or if it is an intermediate in a general mechanism for nucleotide incorporation. Several lines of evidence speak to the latter case. *Bst* polymerase bearing substitution V713P, which sterically hinders the formation of the closed state, was shown to adopt the ajar conformation in the presence of the correct dNTP [33]. Similarly, a small population at the intermediate FRET value was recorded in the ternary complex with the correct dNTP (Figure 2.7b) [5]. Finally, targeted molecular dynamics (MD) simulations of the open-to-closed state transition showed an intermediate with a bent O-helix similar to the ajar conformation [35]. Nevertheless, more recent unbiased MD simulations of Pol fingers-opening discovered a new intermediate, stabilized by a salt bridge between residue R703 of the fingers and residue E562 of the thumb subdomain (*Bst* numbering), which is distinct from the ajar conformation [36]. The authors of the study argued that this intermediate was likely more important for general dNTP incorporation than the ajar conformation, although no dNTP was present in the simulations.

2.4.3 Synthesis fidelity

Pol is able to synthesise DNA with high fidelity, showing error rates as low as 10^{-5} and 10^{-6} per nucleotide, for single-base substitutions and single-base deletions, respectively [37]. Deletion errors are caused by primer-template slippage events, such as due to polymerase dissociation and reassociation. It is thought that the base-flipping mechanism of DNA polymerization plays an important part in preventing deletion errors, as it hinders dNTP base-pairing to templates further upstream [8]. Substitution errors, in contrast, result from the polymerase incorporating a mismatched nucleotide into the primer strand. It is generally stated the free-energy differences of correct and mismatched base-pairing alone are not sufficient to explain the high polymerase discrimination against mismatches, and hence other, polymerase-dependent mechanisms have been proposed [4, 38]. A number of crystal structures indicate that polymerases form a tight binding pocket around the nascent base-pair [3], allowing geometric selection for the correct versus the mismatched

base-pair. Mutagenesis studies in *E. coli* Pol have also shown that some of the residues forming the nucleotide binding pocket (E710, Y766, R668 and R682) are important for synthesis fidelity [39]. In addition, it has been suggested that polymerases employ mechanisms to exclude water from the active site, in order to amplify the enthalpic differences between the correct and mismatched base-pairs [40].

Perhaps the most important mechanism of ensuring fidelity in polymerases involves the induced fit mechanism of nucleotide biding [38]. As mentioned above, crystal structures and single-molecule data indicated that Pol adopts a partially-closed state when bound to a mismatch. Based on these results, it has been proposed that Pol proceeds to the catalytically-competent closed state only upon binding of the complementary but not a mismatched nucleotide, with the ajar conformation serving as the fidelity checkpoint [13, 33]. More recently, single-molecule FRET analysis of the conformational equilibrium of variants E710A and Y766A in the presence of the *complementary* dNTP showed major populations at the intermediate FRET state, suggesting that these Pol variants failed to transition from the partially closed to the closed conformation [5]. Thus, the sensing of nucleotide complementarity in the binding pocket by E710 and Y766 appears to be directly linked to the ability of the polymerase to transition to the fully closed state. Although FRET experiments could not distinguish between complexes with mismatched nucleotides and matched ribonucleotides, 2-aminopurine assays of DNA rearrangement show that ribonucleotides proceed further than mismatches [41]. Two separate checkpoints within the fingers-closing transition have therefore been suggested, whereby the incoming dNTP is first tested for its base and then for its sugar structure [5].

2.4.4 Nucleotidyl transfer

Phosphodiester bond formation between the incoming nucleotide and the 3' OH group of the primer terminus is catalysed by two metal (magnesium) ions [42]. In the open structure, metal A is absent and metal B is coordinated by D653 and D830; the latter residue additionally forms a hydrogen bond with the 3' OH of the primer [8]. Binding of the in-

coming nucleotide completes the assembly of the metal centre, with metal A now coordinated by D830, the 3' OH of the primer and the α-phosphate of the nucleotide, and metal B coordinated by all three phosphate groups (Figure 2.8). The configuration of metal A positions the 3' OH of the primer in line for nucleophilic attack on the α-phosphate, and also lowers the pKa of the 3' OH group, making it a better nucleophile. Both metals are thought to stabilize the resulting pentacoordinated intermediate, with metal B additionally stabilizing the leaving oxygen on the β-phosphate, promoting formation of the pyrophosphate product [43].

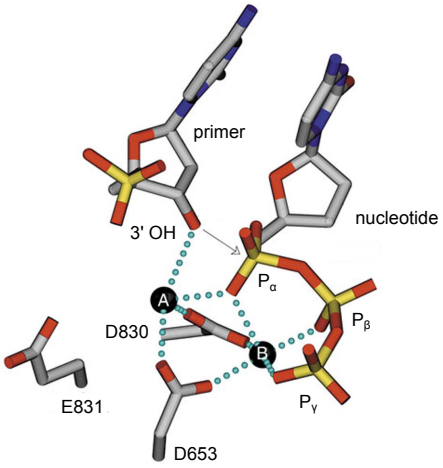


Figure 2.8: *Mechanism of nucleotidyl transfer in DNA polymerases, showing the final configuration of the primer, the incoming nucleotide and Pol catalytic residues. The two metals are shown as black spheres, with the dotted cyan lines indicating coordination. Adapted from reference [42].*

2.4.5 Translocation

Following nucleotide incorporation, the fingers subdomain opens, the PPi product is released, and Pol translocates to the next template position; however, the order of these events is unknown. A crystal structure of *E. coli* Pol complexed to PPi adopts an open conformation, suggesting that fingers-opening occurs prior to pyrophosphate release [24]. The PPi is seen to interact with conserved Lys and Arg residues, which could thus func-

tion to remove it from the active site. However, targeted molecular dynamics simulations of the *Bst* enzyme showed that PPi was released first and that its release facilitated the fingers-opening transition, which proceeded through an ajar-like intermediate [35]. Interestingly, Y714 was seen to couple fingers-opening to DNA translocation, moving back into its 'blocking' position and pushing against the template base of the nascent base-pair.

The non-specific nature of Pol-DNA interactions allows Pol to spiral along the DNA backbone. The strain created in the A-form region of the DNA substrate could be responsible for the 5'-3' directionality of translocation, and the base-flipping mechanism serve to ensure that translocation halts at the next template base [3]. The number of cycles that Pol performs before dissociating from the DNA has been determined to be on the order of ~7 nucleotides for the *E. coli* enzyme and ~110 nucleotides for the *Bst* enzyme [25]. A number of studies have also measured the rate of processive incorporation and thus translocation, and reported values at ~13-15 nucleotides per second for the *E. coli* Pol (KF) [10, 44, 45]. However, this rate is unlikely to be limited by the rate of DNA translocation itself, but could instead be due to the slow rate of the fingers-opening transition, as suggested by a recent single-molecule FRET study [6].

2.4.6 Strand displacement

The removal of RNA primers during Okazaki fragment synthesis relies on the strand-displacement activity of Pol, which resides in its polymerase domain. The kinetics of strand-displacement synthesis have been studied by single-molecule methods, and indicated similar rates to primer-extension synthesis [46]. However, since no structure is available of an A-family polymerase complexed with downstream DNA, the mechanism of strand separation has been elusive. A kinetic analysis of *E. coli* Pol variants S769A, F771A and R841A showed that, whilst the substitutions do not affect the polymerase activity of Pol, they significantly impair its strand-displacement function [10]. In addition, these residues are part of a conserved bundle of three helices (O, O1 and O2) that is seen to participate in the binding and strand separation of downstream DNA in the T7 RNA poly-

merase (RNAP) crystal structure [47]. It was therefore suggested that F771 may act as a wedge between the two strands of downstream DNA, with R841 and S769 helping to stabilize the template and non-template strands, respectively (Figure 2.9) [10].

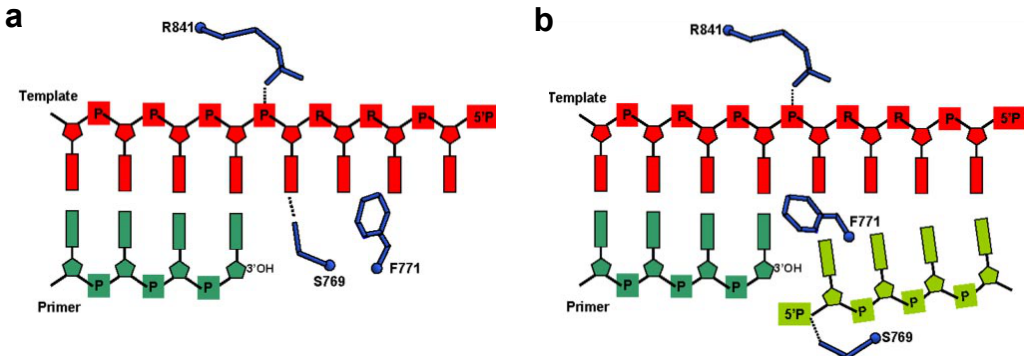


Figure 2.9: *Mechanism of strand displacement, based on mutagenesis studies.* ***(a)*** *Interactions of residues S769, F771 and R841 with the template strand, as observed in an open-state Bst structure (PDB code 1L3S, reference [8]).* ***(b)*** *Proposed rearrangement of the same residues during strand-displacement synthesis. The template strand is shown in red, the primer strand in dark green and the downstream non-template strand in light green. Reproduced from reference [10].*

2.5 Proofreading

Structures are also available of Pol in the editing complex, with primer-template DNA bound at the 5'-3' exonuclease site. The duplex region of the DNA is found in a similar position as the DNA approaching the polymerase site, but is rotated towards the exonuclease site and translated away from the protein along the helix axis (Figure 2.10) [23, 30]. The tip of the thumb subdomain moves to accommodate the different position of the minor groove, and forms interactions with the sugar-phosphate backbone of the DNA. The primer strand is channelled towards the exonuclease site and interacts with a number of conserved Pol residues, including the invariant K635. The binding pocket of the active site can accommodate single-stranded but not double-stranded DNA, and is formed of Asp and Glu residues that coordinate a pair of divalent metal ions, which in turn interact with the phosphodiester bond to be cleaved.

The polymerase and exonuclease catalytic sites are located ~30 Å apart and generally behave independently of each other; however, their DNA-binding functions are intertwined [2]. It is thought that editing is achieved as a result of the increased propensity of the mispaired terminus to fray, producing a single-stranded substrate that favours binding to the exonuclease site. In addition, any mismatches in the primer terminus have been observed to stall polymerization, likely due to the disruption of the minor-groove interactions, allowing sufficient time for the substrate to diffuse to the exonuclease site [48]. In contrast, the exonuclease-site geometry has been shown to prefer the mismatched termini over the correctly paired ones [48].

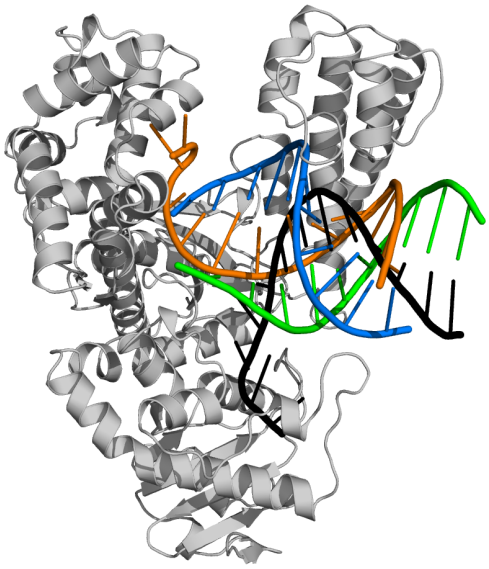


Figure 2.10: *Comparison of polymerase and 5'-3' exonuclease DNA-binding sites in Pol, based on Bst and E. coli structures (PDB codes 4BDP and 1KLN, references [23, 26]). The template and non-template strands are coloured orange and blue for the polymerase site DNA, and green and black for the exonuclease site DNA, respectively.*

2.6 DNA repair in cells

Until recently, Pol characterization in live cells was limited to mutagenesis and deletion studies. For example, it was found that *PolA*- deletion strain of *E.coli* was 10-fold slower in joining Okazaki fragments than the wild-type strain, and had a 7- to 10-fold higher

rate of spontaneous mutations [49, 50]. Similarly, the deletion strain was shown to be highly sensitive to DNA damage induced by UV radiation and alkylating agents [49, 51]. These early studies established the function of Pol and its role in DNA replication and base-excision repair, however, they could not address the kinetic and mechanistic details underlying these pathways.

Pol diffusion and DNA repair in live cells were recently studied using superresolution-based single-molecule tracking of a photoactivatable fluorescent-protein fusion of Pol (Figure 2.11a) [15]. The labelled Pol was seen to diffuse throughout the nucleoid, and was characterized by an apparent diffusion coefficient of ~0.8 $\mu m^2 s^{-1}$, corresponding to an accurate diffusion coefficient of 2.7 +/- 0.4 $\mu m^2 s^{-1}$ (Figure 2.11b, c). These results showed that Pol did not form stable oligomers in the cell, and that its diffusion largely followed Brownian motion. Although its association with the nucleoid indicated non-specific interactions with the DNA, it was concluded that these must have been limited to transient events beyond the time resolution of the experiment (~15 ms) [15].

The fraction of Pol molecules that were immobile in undamaged cells, corresponding to molecules active in lagging-strand replication and basal DNA repair, was only ~3 % (Figure 2.11b, c). Therefore, only a small percentage of the available wild-type Pol activity appears to be required for normal cell growth. However, upon treatment with a DNA-methylating agent methyl methanesulfonate (MMS) that generates gapped-DNA intermediates, the fraction increased to ~13 %, reflecting significantly higher repair rates. The repair sites were evenly distributed throughout the cell, indicative of Pol searching for damaged substrates on the chromosome, rather than these being directed to spatially organized 'repair factories'. Pol repair times were measured at ~2 seconds for both lagging-strand replication and DNA repair, and its search times were on the order of tens of seconds, with Pol molecules spending >80 % of time searching for damaged substrates. Regardless, the high copy number of Pol (~480 copies per cell) appears to ensure that almost all damaged sites are Pol-bound at any one time, preventing the build-up of the toxic DNA-repair intermediates [15].

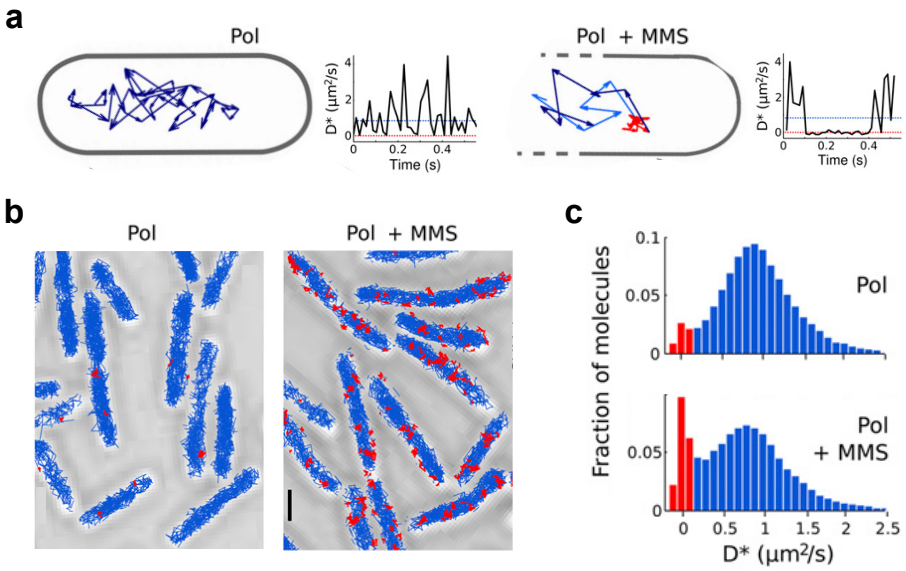


Figure 2.11: *Pol diffusion and DNA repair in live cells.* ***(a)*** *Example Pol tracks in an undamaged cell and an MMS-treated cell, with time traces of corresponding apparent diffusion coefficients shown on the right. In the damaged cell, the track consists of initial substrate search (light blue), a repair event (red) and continued search (dark blue).* ***(b)*** *Example fields of view of undamaged and MMS-treated cells, with single-molecule tracks overlaid on white-light images. Diffusing tracks are in blue and immobile tracks in red.* ***(c)*** *Distributions of apparent diffusion coefficient D* in undamaged and MMS-treated cells. Adapted from reference [15].*

3

Fluorescence and simulation approaches

In order to provide the reader with the necessary technical background, in this chapter we describe two key approaches that we employ in most of the subsequent chapters: fluorescence microscopy and molecular dynamics simulations. We give an introduction to the principles of fluorescence and FRET, and the concepts of single-molecule detection and single-molecule FRET. We use smFRET to probe the structure (Chapter 4) and conformational dynamics of Pol (Chapter 6), and we apply single-molecule detection and FRET to study Pol in live cells (Chapters 7 and 8). In the second part of this chapter, we review the key concepts underlying an MD simulation, including force fields, the treatment of long-range interactions and the solvent, and the different methods for controlling and biasing simulations. We use coarse-grained and all-atom MD simulations to investigate the dynamics of the gapped-DNA substrate (Chapters 4 and 5), of the Pol-DNA complex (Chapter 5), and as a basis for the energy decomposition method (Chapter 6). In addition to the general overview of fluorescence microscopy and MD simulations provided here, brief introductions to the methods employed in specific chapters are given at the start of each chapter.

3.1 Fluorescence microscopy

3.1.1 Introduction

Fluorescence-based methods are characterized by high sensitivity and specificity, allowing a strong, characteristic signal to be observed from a small amount of a complex sample. Many of them are straight-forward and non-invasive, requiring only minimal modification of the sample and causing little or no damage, and can be applied in a variety of *in vitro* and *in vivo* contexts [52]. Fluorescence has allowed the resolution limit of light microscopy to be broken, leading to breakthroughs in cellular imaging [53, 54]. Combined with single-molecule detection, it has enabled the study of intracellular dynamics via single-molecule tracking [55, 56], and has allowed molecular structure, function and interactions to be probed via single-molecule FRET [57]. Similarly, fluorescence has become the tool of choice in the applied sciences: it has superseded many of the traditional approaches and has become an indispensable tool in biotechnology, DNA sequencing, forensics, and medical diagnosis [52].

3.1.2 Principles of fluorescence

A molecule can absorb light when the energy difference between two of its electronic states matches the energy of incoming photons. The electronic states of a molecule can be illustrated with a Jablonski diagram (Figure 3.1a), and are termed the ground (S_0) and excited states (S_1, S_2, etc.); each state additionally consists of several vibrational energy levels [52]. Photon absorption causes an electron to transition from the ground state to one of the vibrational levels of an excited state (occurring at ~10^{-15} s), followed by vibrational relaxation to the lowest level of the excited state (~10^{-12} s). The excited electron can then return to the ground state through a number of mechanisms. One of these is *fluorescence*, which involves a direct relaxation from S_1 to S_0 with a concomitant emission of energy in the form of photons, and occurs at ~10^{-8} s. Because of the vibrational relaxation, the emitted light is lower in energy than the absorbed light, leading to a wavelength difference between the

two that is known as the Stokes shift (Figure 3.1b).

Alternatively, the excited state can switch its spin to match the spin of the ground state electron, forming the so-called triplet (T_1) state [52]. Since transitions from the triplet state to the ground state are 'forbidden', they occur at a very slow rate (10^{-3} s to minutes or hours), giving rise to a process of photon emission known as *phosphorescence*. The formation of triplet states is also one of the major phenomena underlying reversible transitions of fluorescent molecules to a dark state (fluorophore blinking). In addition to fluorescence and phosphorescence, relaxation from the excited to the ground state can also occur due to contact with another molecule in solution (collisional quenching), or through other non-radiative pathways.

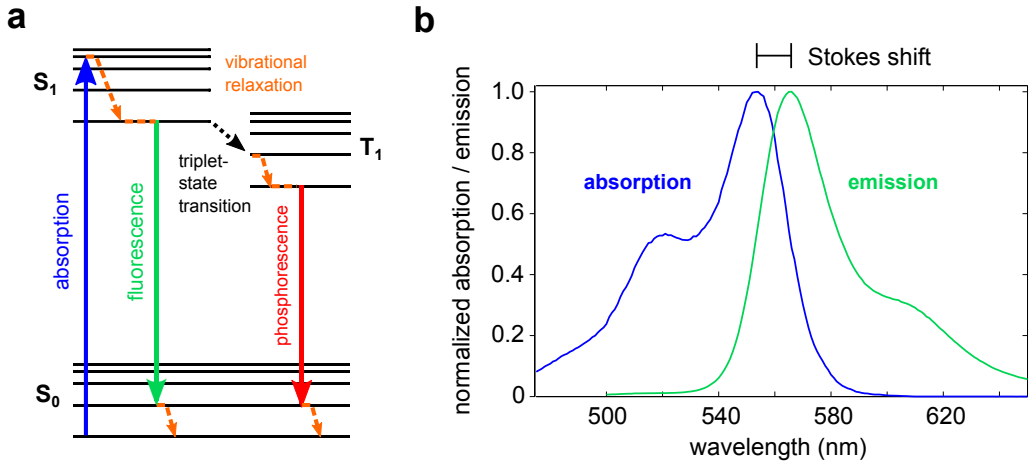


Figure 3.1: *Principles of fluorescence.* ***(a)*** *Jablonski diagram, showing the electronic ground state (S_0), an exited state (S_1) and a triplet state (T_1) of a molecule, as well as the different vibrational energy levels. Absorption, fluorescence and phosphorescence are indicated with blue, green and red arrows, respectively. Vibrational relaxation is denoted with dashed orange arrows, and triplet-state transition with a dotted black arrow.* ***(b)*** *Absorption (blue) and emission spectra (green) of a typical fluorophore, normalized to unity for direct comparison. The Stokes shift is indicated between the peaks of the two spectra. Adapted from Wikimedia and lab commons.*

Fluorophores are characterized by several photophysical properties, including quantum yield, fluorescence lifetime, and photostability [52]. Quantum yield (Q) is the ratio of the number of emitted photons and the number of absorbed photons. Along with the

extinction coefficient (ε), which describes how strongly a fluorophore can absorb light, it determines fluorophore brightness, often expressed simply as $B = \varepsilon Q$. Fluorescence lifetime (τ) is the average time a molecule spends in an excited state before relaxing to the ground state. In addition to the differences in emission wavelengths between fluorophores, differences in fluorescence lifetimes can be exploited for highly specific imaging. Finally, photostability refers to the probability of a dye to undergo irreversible photobleaching [58]. Photobleaching normally occurs from an excited triplet state, in which a fluorophore can interact with another molecule (often oxygen or water) to produce irreversible covalent modifications, resulting in the loss of the ability to fluoresce. Photobleaching can be controlled by reducing oxygen levels in solution and by use of triplet-state quenchers.

3.1.3 Förster resonance energy transfer

Förster resonance energy transfer is a process of non-radiative energy transfer between a 'donor' and an 'acceptor' molecule. The dependence of FRET efficiency on the distance between the molecules makes it ideally suited for measuring distances in the 2-10 nm range, which has earned it the title of a 'spectroscopic ruler' [59]. In the scenario described above, an excited-state electron can return to the ground state by transferring its energy via FRET to a ground-state electron of a different molecule (Figure 3.2a). No photons are transmitted between the donor and acceptor; instead, the two molecules are coupled through a dipole-dipole interaction [60]. In order for FRET to occur, the donor and acceptor have to be close to each other in space, and the emission spectrum of the donor needs to overlap with the absorption spectrum of the acceptor. The rate of energy transfer $k_T(r)$ can thus be expressed as:

$$k_T(r) = \frac{1}{\tau_D}\left(\frac{R_0}{r}\right)^6 \tag{3.1}$$

where τ_D is the lifetime of the donor in the absence of the acceptor, r is the distance between the donor and acceptor, and R_0 is the Förster radius, which is defined as the distance at which the donor-acceptor transfer efficiency is 50 %. R_0 depends on the properties

of the donor and acceptor, as follows:

$$R_0^6 = \frac{9000 \ln(10) Q_D \kappa^2 J}{128 \pi^5 N_A n^4} \tag{3.2}$$

where Q_D is the quantum yield of the donor, N_A is Avogadro's number, and n is the refractive index of the medium. The term κ^2 describes the relative orientation of the transition dipoles of the donor and acceptor. Its value lies in the range of 0-4, and it is often assumed to be equal to 2/3, which is the case when both fluorophores have unrestricted rotational freedom [59]. The overlap integral J is a measure of the degree of overlap between the donor emission and acceptor excitation [61]:

$$J(\lambda) = \int_0^\infty F_D(\lambda) \varepsilon_A(\lambda) \lambda^4 d\lambda \tag{3.3}$$

where F_D is the corrected donor fluorescence intensity at a particular wavelength λ, with the total intensity normalized to unity, and ε_A is the extinction coefficient of the acceptor at the same wavelength.

The efficiency of energy transfer E is equal to the ratio of the transfer rate $k_T(r)$ to the total decay rate of the donor, which recalling equation 3.1 can be expressed in terms of the distance r:

$$E = \frac{k_T(r)}{\tau_D^{-1} + k_T(r)} = \frac{1}{1 + (r/R_0)^6} \tag{3.4}$$

This equation indicates that the transfer efficiency is strongly dependent on inter-fluorophore distance in the region close to R_0 (Figure 3.2b). The value of R_0 thus determines the centre of the dynamic range of observable distances, with accurate measurements usually being limited to the linear region of the curve, between $E = 0.1$ and $E = 0.9$.

3.1.4 Single-molecule detection

Single-molecule experiments investigate the properties of individual molecules, in contrast to ensemble experiments, which are carried out on large collections on molecules. In

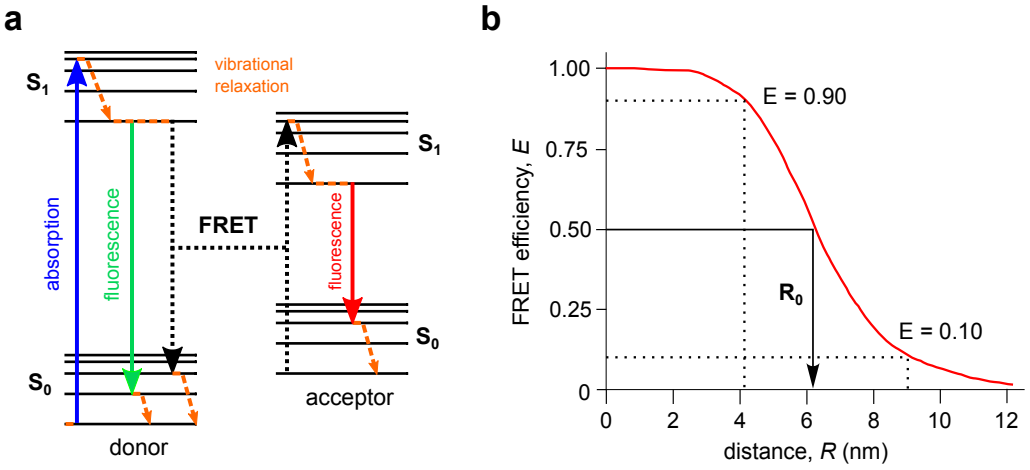


Figure 3.2: *Förster resonance energy transfer.* ***(a)*** *Jablonski diagram for FRET, showing the electronic ground and excited states of a donor and an acceptor molecule. The donor molecule with its electron in an excited state can either fluoresce (green arrow), or transfer its energy to the acceptor molecule (dotted black arrows), which can in turn fluoresce at a higher wavelength (red arrow).* ***(b)*** *Distance-dependence of FRET efficiency. The R_0 value, defined as the distance at $E = 0.5$, is indicated with an arrow. Distances corresponding to $E = 0.9$ and $E = 0.1$, which define the dynamic range of measurable distances, are also marked. Adapted from Wikimedia and lab commons.*

this way, single-molecule experiments can uncover the distributions of observables instead of relying on population-averaged values, thus revealing additional information about the system under study. Biological samples are often characterized by high heterogeneity, both static and dynamic, which arises from the biological complexity and the stochastic nature of chemical processes [62]. Static heterogeneity refers to the presence of different stable subpopulations, such as active and inactive molecules, or molecules in different conformations or functional states (Figure 3.3). Dynamic heterogeneity, on the other hand, refers to the case where molecules can interconvert between different states on the timescale of the experiment, such as an enzyme that can switch between two different catalytically competent states. With single-molecule detection, both the different molecular states and their interconversions can be observed, reaction pathways can be followed in real time, and transient intermediates can be captured without the need for molecular synchronization. The realization of single-molecule detection in many force-based and fluorescence-based approaches has revolutionized the field of biophysics over the last few decades [62–64].

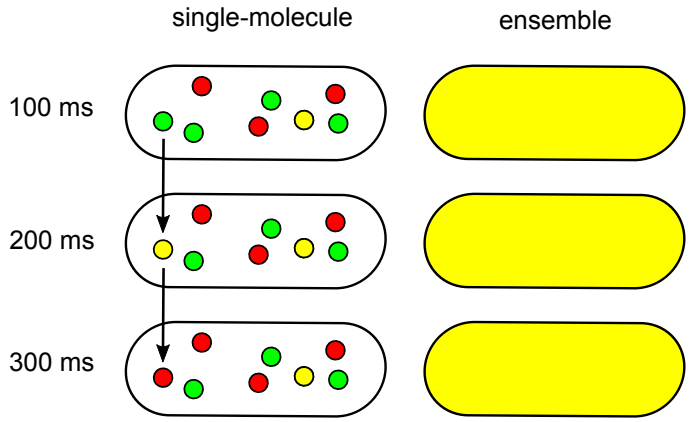


Figure 3.3: *Schematic of the ability of single-molecule (left) and ensemble detection (right column) to detect static and dynamic heterogeneity in a population of molecules. In this example, a molecule can exist in three different states (red, yellow, green; static heterogeneity) and can interconvert between them on a 100-ms time scale (dynamic heterogeneity). Single-molecule approaches can detect both the different states and the conversions between them, whereas ensemble approaches only observe a signal that is averaged across the population.*

3.1.5 Single-molecule fluorescence

Fluorescence is characterized by high sensitivity, arising from the fact that a single molecule can emit a large number of photons, and is thus an obvious choice for single-molecule detection [52]. However, single-molecule fluorescence is technically challenging. The main challenge arises from the background fluorescence, particularly auto-fluorescence of the sample and the optical components, and the Raman scattering from the solvent. The background signal can be controlled for by using reduced illumination volumes, such as by tightly focusing the laser beam, or by producing a shallow field of illumination, as we describe below. In addition, the wavelength properties of laser excitation, dichroic mirrors and filters should be carefully selected, and high-numerical aperture objectives and sensitive detectors can be used to maximize photon collection. Finally, the development of bright and photostable dyes has also made a significant contribution to the feasibility of single-molecule detection.

Single-molecule fluorescence is generally realized in one of two formats: confocal microscopy or total internal reflection fluorescence (TIRF) microscopy. In single-molecule

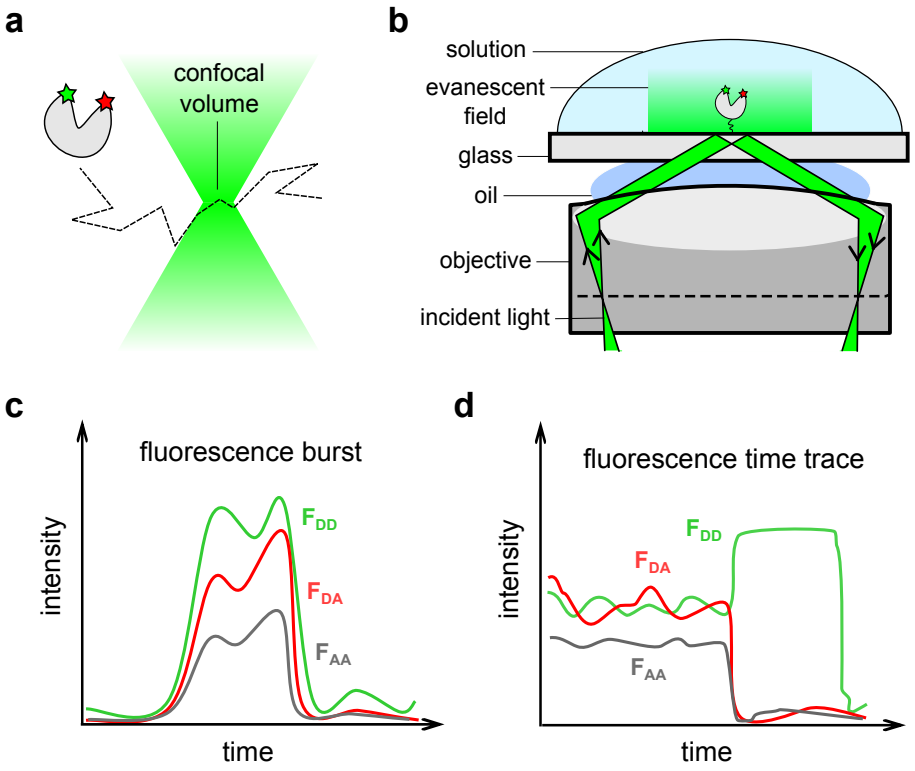


Figure 3.4: *Single-molecule microscopy, applied for smFRET measurements.* ***(a)*** *The principle of confocal microscopy. A labelled molecule diffuses through the confocal volume, during which time it absorbs light and emits a fluorescent burst.* ***(b)*** *The principle of TIRF. Total reflection of the laser beam from the glass/water interface creates the evanescent field, exciting a narrow layer of molecules at the surface.* ***(c)*** *Schematic of a fluorescence burst observed using confocal microscopy, which has been deconvolved into the different channels (F_{DD}, F_{DA} and F_{AA}).* ***(d)*** *Schematic of a fluorescence time trace observed using TIRF, with the intensities in the three channels correlated. In this example, a photo-bleaching event is evident from the drop in F_{DA} and the rise in F_{DD} intensities. Adapted from lab commons and reference [65].*

confocal microscopy (Figure 3.4a, c), a focused laser beam with the dimension of a diffraction-limited spot is used to produce an excitation volume of ~1 fl, significantly minimizing the background signal [66]. Fluorescent molecules, present at pM concentration, traverse the confocal volume due to Brownian motion and emit a burst of fluorescence each time. Fluorescence is collected using the same objective and passed through a pinhole that rejects any light below or above the focal plane of the objective, before being split into different channels and detected using avalanche photodiodes (APDs). The advantages of confocal microscopy include its high time resolution and the lack of need to immobilize the mo-

lecule of interest. However, the observation time is limited to the diffusion time of the molecule through the confocal volume (~1 ms) [67], obscuring information on molecular history. Scanning confocal microscopy overcomes this challenge, but can only image individual immobile molecules.

In total internal reflection fluorescence (TIRF) microscopy (Figure 3.4b, d), illumination is achieved by pointing the laser beam at a *critical angle* at the interface between the glass coverslip (with a higher refractive index) and water (with a lower refractive index) [68]. The beam is totally reflected from the interface, but generates an electromagnetic field, called the *evanescent field*, which decays exponentially from the interface and penetrates to a depth of ~100 nm into the sample. Only molecules within the evanescent field are illuminated, resulting in significant reduction of background. The resulting fluorescence is split into different channels and detected on an electron-multiplying charge-coupled device (EM-CCD) camera, where its intensity at any position can be tracked over time. Due to its widefield set-up, TIRF allows many immobilized or confined diffusing molecules to be imaged at the same time, with long observation times. However, immobilization procedures can be cumbersome and can introduce artefacts, and the time resolution is limited by the frame rate of the camera (~10 ms).

3.1.6 Single-molecule FRET and ALEX

Single-molecule detection reveals the true potential of FRET, enabling intra- and inter-molecular distance measurements in individual molecules [69]. This allows probing different structural states of proteins and molecular machines, their conformational dynamics and reaction kinetics (Figure 3.5). Unlabelled, singly-labelled or inactive species can be specifically detected and discarded from analysis, and only the species of interest examined [57]. Single-molecule FRET can be realized using either confocal or TIRF microscopy, and is generally measured from fluorescence lifetimes or fluorescence intensities. In the case of intensity-based smFRET, the sample is excited in the donor channel, and the fluor-

escence intensities resulting in the donor (F_{DD}) and acceptor channels (F_{DA}) [1] are recorded (Figure 3.4). The apparent FRET efficiency E^* is then calculated for each molecule as:

$$E^* = \frac{F_{DA}}{F_{DA} + F_{DD}} \tag{3.5}$$

The resulting FRET histograms allow the different species and states of molecules to be identified, but additional corrections are required for calculation of distances, as we describe below.

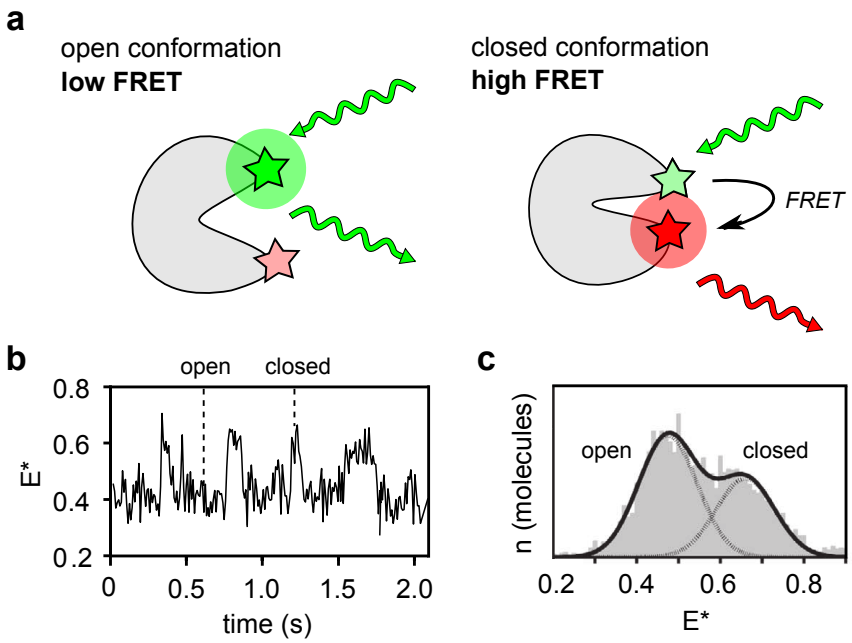


Figure 3.5: *Example of smFRET application.* ***(a)*** *An enzyme that interconverts between two conformations is doubly labelled for an smFRET experiment. In its open conformation, the distance between the labels is large, and the resulting FRET efficiency is low. In its closed conformation, the distance between the labels decreases, leading to a higher FRET efficiency.* ***(b)*** *Example raw single-molecule fluorescence intensity trace, measured using TIRF, showing interconversions between the open and closed states.* ***(c)*** *Example apparent FRET efficiency histogram, obtained from confocal data, showing the open and closed populations. Adapted from the DPhil thesis of Geraint Evans and reference [65].*

[1] We use this notation throughout the thesis, where channel XY describes the fluorescence detected in channel Y after excitation in channel X, and F_{XY} is the corresponding fluorescence intensity.

Occasionally, it is not possible to identify the different species in a FRET histogram even with single-molecule detection. For example, donor-only species, arising from incomplete labelling or dye photobleaching, can be difficult to distinguish from low-FRET species. In addition, complexes with different stoichiometries cannot be assigned simply from the measured efficiency. Alternating-laser excitation (ALEX) is an approach that addresses these limitations, by directly exciting the donor and acceptor fluorophores in an alternating fashion [70]. In the case of confocal microscopy, the alternation frequency is set to ensure that the donor and acceptor are excited several times during the passage of a labelled molecule through the confocal volume (~20 kHz). Using this scheme, one gains access to a third channel, AA, which describes the fluorescence detected in the acceptor channel after direct excitation of the acceptor. In turn, this information allows the calculation of the raw stoichiometry S^{raw}, which refers to the total fluorescence recorded after donor excitation, relative to the total fluorescence after donor and acceptor excitation:

$$S^{raw} = \frac{F_{DD} + F_{DA}}{F_{DD} + F_{DA} + F_{AA}} \tag{3.6}$$

Applying the equation, donor-only species will give S^{raw} values of ~1, and acceptor-only species will give S^{raw} values of ~0. The E and S parameters can be displayed and analysed using a two-dimensional histogram, allowing singly-labelled and FRET species to be deconvolved (Figure 3.6).

Having identified the various species, the extracted apparent efficiencies E^* need to be corrected for a number of factors, in order to obtain the accurate efficiencies E that can be related to inter-fluorophore distances. We describe this procedure for diffusion-based confocal microscopy, where single-molecule bursts are measured in the DD, DA and AA channels (Figure 3.4c) [65, 71].

- First, the three photon streams are corrected for background, which arises from impurities, Raman scattering from the solvent and dark counts in the detectors. For each burst, the corrected counts are calculated from the raw counts by subtracting the background count rate, multiplied by the length of the burst. From the

background-corrected intensities F^c, one can calculate the raw proximity ratio:

$$E_{PR}^{raw} = \frac{F_{DA}^c}{F_{DA}^c + F_{DD}^c} \tag{3.7}$$

- Second, the intensities are corrected for spectral cross-talk. Because of the limitations of spectral separation, the background-corrected F_{DA} is still not a true measure of transfer efficiency, but has a number of components:

$$F_{DA}^c = Lk + Dir + F_{FRET} \tag{3.8}$$

where *Lk* is the leakage of the donor emission into the acceptor channel, *Dir* is the direct excitation of the acceptor by the laser used for donor excitation, and F_{FRET} is the photon count due to energy transfer from the donor to the acceptor. The *Lk* and *Dir* contributions can be calculated from the donor-only efficiency and acceptor-only stoichiometry peaks in the E/S histogram, allowing the determination of F_{FRET}. The proximity ratio E_{PR} can then be defined as:

$$E_{PR} = \frac{F^{FRET}}{F_{DD}^c + F^{FRET}} \tag{3.9}$$

- Finally, the efficiencies are corrected using correction factor γ, which represents the ratio of quantum yields and detection efficiencies of the donor (in the absence of the acceptor) and of the acceptor, respectively. Factor γ can be obtained by plotting $1/S$ against E_{PR} for two or more FRET species, and extracting the intercept and the slope of the resulting linear fit. The accurate FRET E is then:

$$E = \frac{E_{PR}}{\gamma - (\gamma - 1)E_{PR}} \tag{3.10}$$

which can be used to calculate the true inter-fluorophore distance, using equation 3.4.

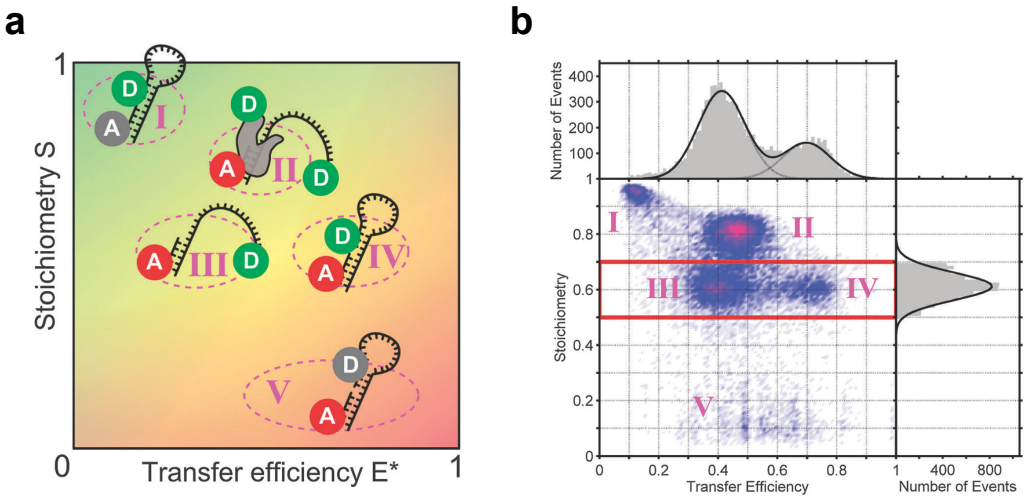


Figure 3.6: *Ability of ALEX to separate different singly-labelled and FRET species.* ***(a)*** *Schematic of an E/S histogram, showing the separation of five different DNA and DNA-protein species present in the sample. Green and red circles refer to active fluorophores, whereas grey labels indicate inactive fluorophores, e.g. due to photobleaching.* ***(b)*** *Simulated E/S histogram of the same sample, with the five species in (a) annotated. The one-dimensional histograms represent the data defined by the red box. Adapted from reference [65].*

3.2 Molecular dynamics simulations

3.2.1 Introduction

Biomolecules show significant dynamics, on levels ranging from bond rotations and vibrations to large-scale conformational transitions [72]. However, experimental methods for characterizing molecular dynamics with atomistic detail and at sub-microsecond time resolution are generally lacking. Molecular dynamics simulations use a potential energy function and Newton's laws of motion to calculate the position of every atom in the system after a short time increment, hence generating a trajectory of molecular motion. The trajectory contains all the information required to describe the transition of the system from one state to the other, and hence to understand the underlying molecular changes and interactions [73]. Conformational transitions, ligand binding, enzyme catalysis, membrane transport and protein oligomerization are just some of the processes that can be studied in this way [74–77]. In addition, simulations enable one to test conditions that would be difficult, expensive or even impossible to test experimentally (allowing 'thought experiments'), which can bring significant insight into the workings of the system under normal conditions [78].

3.2.2 Potential energy functions and force fields

An essential requirement for running a molecular dynamics simulation is defining the parameters governing the interactions and forces between atoms in the system [79]. One approach to do this, known as *ab initio* MD, is based on a quantum-mechanical treatment, and involves a first-principles calculation of the electronic structure. Although faithfully representing the behaviour of every electron in the system, this approach is extremely computationally expensive, and is currently only applicable to very short simulations of simple systems. For complex biological systems, a classical-mechanical treatment is employed instead, whereby one defines an (empirical) potential energy function that depends on all relative atomic positions and their interactions, such that the behaviour of the system is

reproduced as accurately as possible. To evaluate the potential energy, a set of parameters is required for each set of atoms, which is usually obtained by a combination of fitting to experimental (such as thermodynamical) data, and to values predicted from quantum-mechanical calculations. This procedure is known as *parametrisation*, and a means of describing a simulated system that includes both a potential energy function and its associated parameters is referred to as a *force field*.

Generally, potential energy functions describing biomolecular systems consists of bonded and non-bonded components [80]. The bonded terms include the contributions from bond stretching and bond bending, both of which are modelled as harmonic springs whose values deviate from the equilibrium value set in the force field. In addition, a dihedral-angle term is included, which refers to the orientation about the central bond in an arrangement of four atoms separated by three covalent bonds, and occasionally an 'improper' dihedral term, which relates to the planarity of the central atom relative to the three atoms attached to it. The non-bonded terms include van der Waal's interactions, approximated by Lennard-Jones potential, and electrostatic interactions, described by Coulomb's law. A typical potential energy function [81] has the following form:

$$U_{\text{total}} = U_{\text{bonded}} + U_{\text{non-bonded}} \tag{3.11}$$

$$U_{\text{bonded}} = \sum_{\text{bonds}} k_b(b - b_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{torsions}} \sum_n \frac{V_n}{2}[1 + cos(n\phi - \gamma)] \tag{3.12}$$

$$U_{\text{non-bonded}} = \sum_i \sum_{j>i} \varepsilon_{ij}\left[\left(\frac{r_{0ij}}{r_{ij}}\right)^{12}\right] - 2\left(\frac{r_{0ij}}{r_{ij}}\right)^{6}\Bigg] + \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{3.13}$$

where b and b_0 are the actual and equilibrium bond lengths, k_b is the bond-stretching force constant, θ and θ_0 are the actual and equilibrium bond angles, k_θ is the bond-bending force constant, ϕ and γ are the actual and reference dihedral angles, V_n is the energy barrier to rotation, and n is the number of rotational minima. In the non-bonded terms, ε refers to van der Waal's energy-well depth, r and r_0 are the actual and equilibrium distances between

atoms i and j, term q is the (partial) charge on the atom, and ε_0 is the dielectric constant.

A number of force fields have been developed for biomolecular simulations, each with its own advantages and disadvantages, depending on the simulated system and the type of simulation. The most commonly used classical force fields are AMBER [81], CHARMM [82], GROMOS [83], OPLS [84] and MMFF [85]. In addition, new *polarizable* force fields are being developed that account for polarization effects, whereby the molecular-charge distribution can vary depending on the dielectric environment [86]. The AMBER force field is a very versatile force field and can be used to simulate proteins, nucleic acids, carbohydrates, lipids and general organic molecules. It was developed by Peter Kollman's group in 1995 [81], and has since seen a number of developments. The most recent ones involved modifying backbone dihedral parameters to improve the balance in secondary-structure elements [87], and modifying torsion potentials of amino-acid side chains [88]. A software package has been developed specifically for the AMBER force fields [89], but the force fields can also be implemented in other software packages, such as GROMACS [90].

3.2.3 Long-range interactions and periodic boundary conditions

Whilst the bonded terms of the potential energy function can easily be evaluated, exact computation of long-range non-bonded interactions is computationally not feasible [91]. As a result, a cut-off distance is usually applied, above which the interaction is assumed to be zero. In the case of van der Waal's interactions, applying a cut-off of e.g. 10 Å is reasonable, since the potential decreases with the negative sixth power of interatomic distance [92]. In order to avoid artefacts, a switching function can be used, which ensures that the energy of the interaction smoothly transitions to zero as it approaches the cut-off distance. However, this approach is rarely successful with electrostatic interactions, whose potential decreases only with the inverse of interatomic distance, preventing long-range interactions from being neglected. The most popular procedure to deal with long-range interactions such as electrostatic interactions is known as Ewald summation, and is based on the idea

that the term $(1/r)$ in the last part of equation 3.13 can be split into two terms, as follows:

$$\frac{1}{r} = \frac{erf\,(\alpha r)}{r} + \frac{erfc\,(\alpha r)}{r} \tag{3.14}$$

where *erf* and *erfc* are the error and the complementary error functions, respectively [79]. The first term of the equation accounts for long-range interactions, and can easily be evaluated in reciprocal space, whereas the second term accounts for short-range interactions and can be calculated in real space. The parameter α is chosen such that a computationally optimal split is achieved between the real-space sum and the reciprocal-space sum. A fast and efficient way of computing the Ewald sum is known as the particle-mesh Ewald method, which uses fast Fourier transformation (FFT), and whose computational cost is proportional only to $N\log(N)$, with N being the number of charges in the system [93].

Ewald summation implicitly assumes that the system is periodic, and hence periodic symmetry needs to be imposed in the simulated system. This can be done by applying periodic boundary conditions (PBCs), which can be thought of as replicating the simulation box to infinity by periodic translations in all three dimensions (Figure 3.7) [79]. The use of PBCs is also important to avoid surface effects at the boundary of the simulated box, which would otherwise inevitably arise due to the small size of the simulated system. In this way, a particle that leaves the simulation box on one side is replaced by its copy entering the box on the opposite side, from its neighbouring box image. In order for PBCs to be used, the simulation box needs to be large enough to prevent any molecule from interacting with itself, and the electrostatic charge of the system has to be zero to prevent summing to an infinite charge.

3.2.4 Solvent models

The solvent can be modelled either by considering individual solvent molecules (i.e. *explicitly*) or by treating it as a continuous medium (i.e. *implicitly*). A number of explicit water models exist that differ in the number of interaction points being modelled, in whether they are rigid or flexible, and in whether they account for polarization effects [94]. Some of

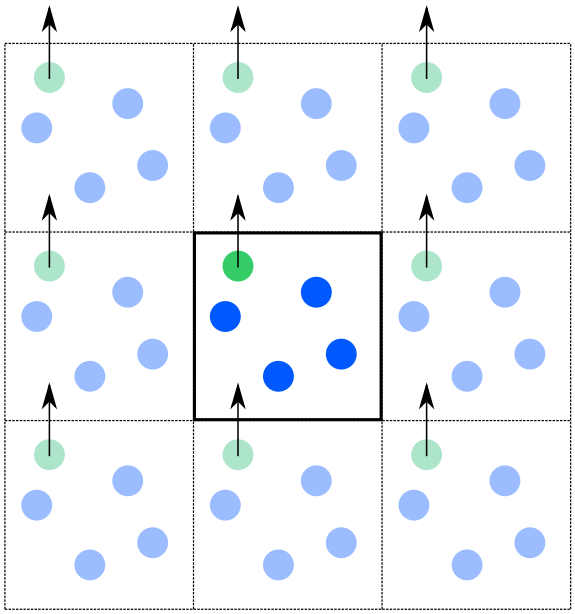


Figure 3.7: *Periodic boundary conditions, illustrated in two dimensions. Simulation box is highlighted in the centre, and is surrounded by periodic images of itself. As the green particle leaves the simulation box, it is replaced by its image from a neighbouring box.*

the popular models, in order of increasing sophistication and computational cost, include SPC [95], SPC/E [96], TIP3P [97] and TIP4P [98]. The TIP3P model has three interaction sites, corresponding to the three atoms in the water molecule, each of which has a point charge; the oxygen atoms also have the Lennard-Jones parameters. In all of these models, polarization effects are only accounted for as corrections to the molecular dipoles, but more sophisticated models exist that can be used with polarizable force fields [99]. The simplest implicit solvent models treat water as a dielectric continuum, but the majority distinguish between high dielectric regions (i.e. in the solvent) and low dielectric regions (inside the macromolecule) [100]. A theoretically well justified model that correctly computes the forces between the solvent and the solute uses the Poisson-Boltzmann equation, but it is computationally too expensive to be used for large-scale simulations; approximate models such as the generalized Born model are frequently used instead [101].

3.2.5 Coarse-grained models

In addition to atomistic descriptions, biomolecular systems can be modelled using less-detailed, coarse-grained models. Generally, a certain number of atoms are mapped into a single bead, resulting in a decrease in the number of particles and a simplification of the system [102]. In addition, because fast vibrations between atoms are removed, the time step can be increased (Section 3.2.6), which further lowers the computational cost of the simulation. In this way, larger systems can be simulated, and longer simulation time scales become feasible. Although atomic-level accuracy is compromised, coarse-grained parameters can be tuned to reproduce the behaviour of atomistic simulations [103]. In addition, different parts of the simulated system and different time sections of the simulation can be modelled as atomistic or coarse-grained, depending on the level of detail required [104].

3.2.6 Computing trajectories

Provided that the total potential energy of the system U_{total} has been evaluated with respect to all atomic positions r, the force F experienced by each atom in the system can be calculated as a derivative of the potential energy [78]:

$$F_i = -\nabla_{r_i} U_{\text{total}}(r_1, r_2, ..., r_N) \tag{3.15}$$

If the mass of each atom is known, then its acceleration can be calculated simply by applying Newton's first law of motion $F = ma$. The starting positions of the atoms are taken from the structure to be simulated, and the starting velocities are usually randomly assigned from the Maxwell-Boltzmann distribution at the desired simulation temperature. Finally, from the acceleration and the starting position and velocity, the position and velocity after a time step Δt are obtained for each atom.

The exact method by which atom positions and velocities are calculated depends on the algorithm used, a number of which exist that differ in accuracy, computational efficiency

and the quantities that are computed. One of the most widely used is the *leapfrog* algorithm [105], whereby velocities are first calculated at time $t+\Delta t/2$, and are then used to calculate positions at time $t + \Delta t$, as follows:

$$v(t + \Delta t/2) = v(t - \Delta t/2) + a(t)\Delta t \tag{3.16}$$

$$r(t + \Delta t) = r(t) + v(t + \Delta t/2)\Delta t \tag{3.17}$$

In this way, the positions and velocities 'leap' over each other, giving the algorithm its name. The time step has to be chosen appropriately; a short time step will give a smaller error in position estimation, but will also increase the computational cost of the simulation. For biomolecular systems, a time step of a few femtoseconds is appropriate, as it is short enough to account for bond vibrations (>10 fs) [92]. Multiple-time stepping can also be employed, whereby slower-varying forces are calculated less often than the faster-varying ones, reducing the overall computational cost [91].

3.2.7 Energy minimization and equilibration

Prior to the actual simulation run, a number of steps are commonly performed to prepare the system for the simulation. First, in order to correct for any errors in the coordinates of the starting structure, the structure is *energy-minimized*. This involves calculating the potential energy of the system at the current geometry, and adjusting the geometry in a stepwise fashion, with the derivative of the potential energy used to determine the direction and magnitude of each step [106]. The process is repeated until a (local) minimum on the potential energy surface has been found, as indicated by the derivative of the potential energy converging to zero. The system then needs to be *equilibrated*, which involves relaxing the solvent and ions around the biomolecule, and bringing the system to the correct temperature and pressure [107]. One protocol for doing so involves equilibration first in the NVT ensemble (constant number of particles, volume and temperature), whereby the temperature is brought to the desired value, followed by the NPT ensemble (constant number of particles, pressure and temperature), whereby the pressure is equilibrated. The

actual simulation run, known as the *production run*, is often also carried out in the NPT ensemble.

In order to keep the temperature constant during the simulation, a variety of methods can be used. These can be stochastic, such as the Andersen method [108], whereby a fraction of atoms are chosen at each time step and their velocity changed to a value selected from the Maxwell-Boltzmann distribution. Alternatively, deterministic methods (such as the Berendsen method [109]) can be used, whereby atom velocities are scaled by a factor equal to the ratio of the the desired and actual temperatures. To avoid instant corrections that would alter the dynamics of the system, the system is coupled to an external 'heat bath', and a time constant is defined to control the rate of heat transfer. Similarly, pressure control can be achieved by rescaling the size of the simulation box during the simulation, such as applied in the Berendsen method [109]. In order to preserve correct fluctuations of the temperature and pressure of the system at equilibrium, the Nosé-Hoover [110] and Parrinello-Rahman methods [111] can be used, respectively.

3.2.8 Biased molecular dynamics simulations

Due to the complexity of atomistic simulations, the simulation time is often restricted to the nanosecond time scale, preventing adequate sampling of some portions of the energy landscape. This is particularly problematic in the case of biomolecules, which are often characterized by multiple local minima of potential energy wells, separated by high free-energy barriers. In order to sample the rare transitions between the energy wells within a short amount of time, a number of 'biased' molecular dynamics approaches have been developed [112].

One of the simplest approaches is *targeted MD*, which uses steering forces to guide the molecule towards the desired conformation, based on the root-mean-square deviation (RMSD) between the current and target structure coordinates [113]. The resulting trajectories should be interpreted with caution, since the transition states achieved in this way may not be physically or biologically relevant. An alternative approach is *steered MD*,

whereby a constant force is applied to a set of atoms in the form of a harmonic restraint, in order to move them in a desired direction [91]. Steered MD can be useful to understand the mechanical properties of molecules, but due to the additional forces applied, the trajectories have to be analysed from a strictly non-equilibrium point of view. Another popular approach is *umbrella sampling*, where a weighting function is added to the true potential energy function in order to compensate for the energy barriers, and bias the sampling to higher-energy conformations.

A disadvantage of all of the above methods is that they require prior knowledge of the free-energy profile and the conformations of interest. One of the alternative approaches that do not rely on this knowledge is *accelerated MD* [114]. In this case, a threshold is selected, called the boost energy, which defines how the simulation is performed. When the true potential is below the chosen threshold, a biased boost potential is added to it, and the simulation run on the modified potential. However, when the true potential is above the threshold, the simulation is run on the true potential itself. In this way, the wells of the potential energy surface are effectively raised, whereas the barriers are left unaffected, leading to an enhanced rate of transitions. The statistics sampled on the biased potential are afterwards corrected to remove the effect of the bias.

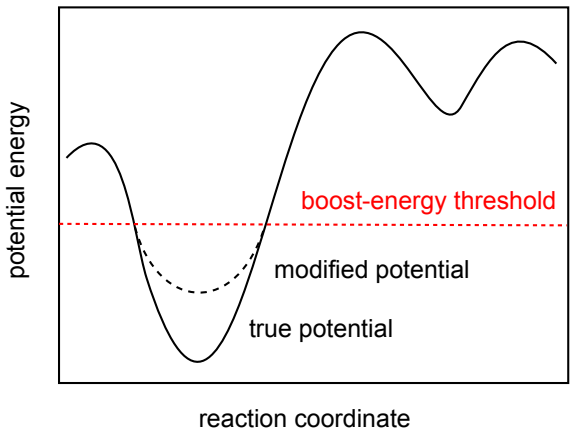


Figure 3.8: *The principle of accelerated MD. Whenever the true potential (solid black line) lies below the boost energy threshold (dashed red line), it is modified by a boost potential, and the simulation performed on the modified potential (dashed black line). Adapted from reference [114].*

4

Structure of Pol bound to gapped DNA

4.1 Introduction

4.1.1 Project rationale

Despite the abundance of crystal structures, our understanding of Pol-DNA interactions in the various catalytic modes of Pol is limited. In particular, structures of Pol in the polymerization mode have established the position of upstream DNA [3], but no structural information is available on the position of downstream DNA. The latter is chemically a DNA duplex in the context of base excision repair, or a DNA-RNA duplex in the context of Okazaki fragment synthesis. In both cases, the strand-displacement activity of Pol is known to promote downstream-DNA unwinding, but the identity of Pol-DNA contacts involved in this process has thus far only been probed using biochemical assays [9, 10]. Similarly, the extent of downstream-DNA unwinding and the exact molecular mechanism of strand separation in Pol are unclear. Although the position of downstream DNA is known for the mammalian DNA polymerase β [115, 116], this information is of limited relevance for the understanding of Pol, due to the lack of strand-displacement activity in Pol β.

It is also not currently understood whether or how the structure of the DNA substrate

is modulated in the Pol-DNA complex, or indeed what the structure and dynamics of the DNA are in the absence of the protein. This information is crucial for understanding the binding mechanism of Pol, and in turn the specificity determinants that allow Pol to recognize its DNA substrates over any other DNA. In the case of base excision repair, the substrate is a DNA duplex containing a gap of one or several nucleotides in one of its two strands, from here onward referred to as *gapped DNA*. Since essentially all of the genome is prone to DNA damage, Pol has to be able to recognize a gapped-DNA substrate of any sequence, unlike many other DNA-binding proteins (such as those involved in transcription), which are sequence-specific [11]. Hence, the determinants of Pol substrate recognition are likely to be encoded in the structure and dynamics of the gapped-DNA substrate.

In order to address these questions, in this chapter we set out to determine the structure of Pol (Klenow fragment)[1] bound to its gapped-DNA substrate. Due to the challenges involved in structure determination of protein-DNA complexes (Section 4.1.2), we used a combination of single-molecule FRET and distance-restrained rigid-body docking. Further, we explored the structure and dynamics of the gapped-DNA substrate alone, using rigid-body docking and coarse-grained simulations. In the following introductory sections, we give a brief overview of FRET-based structure determination, as well as of the OxDNA model that we use for the coarse-grained simulations.

4.1.2 FRET for structure determination

Conventional structural biology techniques, including X-ray crystallography, NMR and cryo-electron microscopy, have been instrumental in understanding protein structure and function. However, X-ray crystallography relies on the ability to crystallize the species of interest, which can be challenging for large and dynamic species, or those whose non-compact shape prevents tight crystal packing [117]. Similarly, NMR is generally limited to small and medium-sized proteins. X-ray crystallography and cryo-electron microscopy

[1]From this point onward, we use the abbreviation 'Pol' to refer to the large fragment of DNA polymerase I (Klenow fragment in the case of *E. coli*, or Bacillus fragment in the case of the *Bst* protein). The full-length polymerase that includes the 5'-endonuclease domain is referred to as 'full-length Pol' or 'flPol'.

also rely on observing the sample under non-physiological conditions (i.e. in the crystal or at very low temperatures), which may affect the structure of the macromolecule. Finally, due to the ensemble nature of all of these approaches, structural heterogeneity is often obscured, and can also hinder the process of structure determination.

FRET offers a promising alternative to macromolecular structure determination in these challenging situations. It can be performed in solution, under close-to-physiological conditions and at the single-molecule level, thus overcoming all of the above limitations. Both intra- and inter-molecular structure can be probed, provided that the molecule(s) of interest can be labelled at the desired positions, and that a sufficient number of distances are measured. Importantly, using FRET for structure determination relies on measuring *absolute* distances, in contrast to most other FRET applications, where only *relative* distances and distance changes are probed [118]. For this reason, the measured raw FRET efficiencies need to be corrected for background, the spectral cross-talk and the γ-factor (Section 3.1.6) [119]. In addition, dye orientation effects and variations in quantum yields can complicate the process of FRET-to-distance conversion, such that it may be necessary to determine the Förster radii experimentally for the labelling positions used.

smFRET measurements are normally carried out using organic fluorophores, which can be attached to any protein or DNA position through sulphide or amine groups. Most organic dyes have long, flexible linkers, which cause their average positions to be significantly displaced from their attachment points on macromolecules [119]. Hence, dye dynamics have to be simulated in some way, such as using MD simulations or with simpler models that probe the accessible volume of the dyes. Once the positions of dyes relative to the macromolecules have been determined, the distances can be integrated into a model using simple triangulation, rigid-body docking, MD simulations, or a combination of these [119]. Some of the more sophisticated methods include the nano-positioning system (NPS) [118], which uses a probabilistic data analysis approach to find the most likely positions of the dyes and macromolecules, and the FRET-restrained positioning and screening (FPS) [120], which we use in this thesis, and is hence described in more detail in the next section.

Examples of smFRET applications have thus far included structures of DNA, such as forked DNA structures [121], of protein-DNA complexes, including the complex of HIV-1 reverse transcriptase binding to primer-template DNA [120], and of protein-RNA complexes, such as the complex of yeast RNA polymerase bound to the nascent RNA [118]. Protein-protein complexes have also been studied, with the complex of synaptotagmin 1 and the SNARE protein being one example [122].

4.1.3 FRET-restrained positioning and screening

The FPS approach is a toolkit for structural determination of macromolecular complexes by smFRET [120]. It is distinguished from other approaches by a number of features, which we summarize below.

- smFRET data are collected using multiparameter fluorescence detection (MFD) [123] and analysed using probability distribution analysis (PDA) [124]. Distance errors are estimated by taking into account both the uncertainties in FRET, obtained from photon statistics, and the uncertainties in the orientation of the dyes, obtained from anisotropy decays. [2]

- FRET-derived distances are converted into true distances between mean dye positions. Because of the different averaging of FRET efficiencies and the donor-acceptor distance R_{DA}, the average FRET measured experimentally is not directly related to the distance between the mean dye positions R_{mp}. Instead, the FRET-averaged distance $\langle R_{DA} \rangle_E$ has to be converted to the R_{mp} distance, which can be done using a polynomial function derived from fitting simulated $\langle R_{DA} \rangle_E$ and R_{mp} values (Figure 4.1a). The $\langle R_{DA} \rangle_E$-to-R_{mp} conversion is particularly important for short distances, where errors of up to 10 Å can result otherwise.

- Dye behaviour is modelled using a geometric accessible volume (AV) algorithm

[2] In the application of FPS in the thesis, this step is not followed. Instead, smFRET measurements are performed as described in Section 3.1.6.

[125], which is computationally much more feasible than full MD simulations, and provides good agreement with experimental data. The dyes are modelled as spheres with a certain radius, connected to the macromolecule by a flexible linker of a defined length and width. The accessible volume of the dye is calculated from all possible dye positions within the linker length from the attachment point, which do not result in steric clashing with the macromolecule (Figure 4.1b). In order to account for the flat shape of organic dyes, the simulation can be repeated using each of the three dimensions of the dye as the radius, and the position distributions superimposed. The centre of the resulting AV cloud is taken as the mean dye position.

- The model is generated by a rigid-body docking approach that takes into account the FRET restraints (Figure 4.1c), and attempts to minimize the data-model deviation (χ^2_E) and any steric clashing (χ^2_{clash}). Each distance is implemented as a spring connecting the mean positions of the dyes, characterized by an equilibrium length of R_{mp} and a strength derived from the distance error. Starting from random configurations of the binding partners, a large number of models of the complex are generated, allowing a certain degree of steric clashing. The models are then refined by recalculating the AVs and the mean positions of the dyes, to account for their new steric environments, and the clash tolerance is progressively decreased. Alternatively, an ensemble of structures can be generated from an MD trajectory, and the agreement with the FRET data calculated for each structure.

- Model quality and uniqueness are assessed by comparing the χ^2_r values of the models, calculated from both χ^2_E and χ^2_{clash}, and the relative RMSD values of the structures. Model precision is estimated using a bootstrapping-type approach, whereby all distances of the best model are modified by random numbers, normally distributed within the range of the experimental errors. The procedure is repeated several times, and the resulting ensemble of the perturbed structures is taken to indicate the distribution of atom positions consistent with the experimental data.

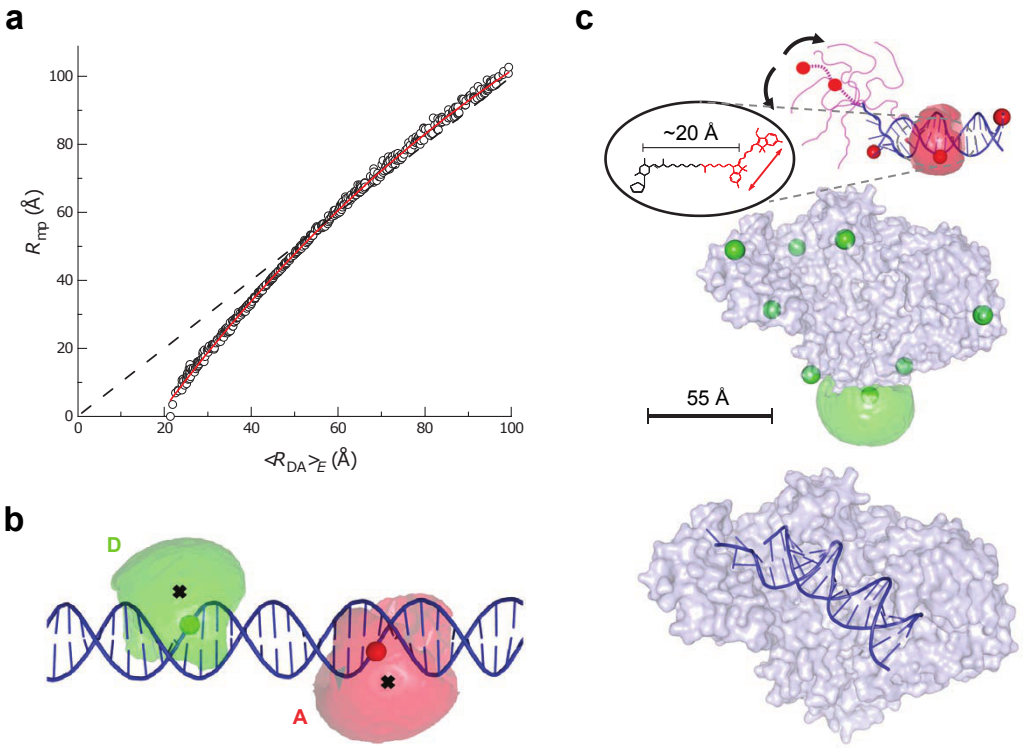


Figure 4.1: *FPS method.* ***(a)*** *Relationship between $\langle R_{DA} \rangle_E$ and R_{mp} distances. Simulated data, obtained for a set of randomly oriented accessible volumes of Alexa488 and Cy5 dyes on double-stranded DNA (shown in circles), are fitted using a third-order polynomial function (solid red line). The dashed black line has a slope of 1 and is shown as a reference.* ***(b)*** *Accessible-volume representation of the donor (D, in green) and acceptor dyes (A, in red) on dsDNA. The attachment points of the dyes are shown with coloured spheres, and their mean positions with black crosses.* ***(c)*** *Top, the protein and DNA structures being docked in the benchmark study, with the donor and acceptor positions shown with spheres, and the flexible part of the DNA indicated with arrows. The molecular structure of the acceptor dye is shown in the inset. Bottom, the resulting docked model of the protein-DNA complex. Adapted from reference [120].*

4.1.4 OxDNA model of DNA

The OxDNA model is a coarse-grained model of DNA that allows simulations of DNA dynamics equivalent to the experimental time scale of microseconds [126]. It is referred to as a 'top-down' model, in that it does not focus on the atomistic details of DNA structure and chemistry, but instead aims to reproduce their net effect on the observed properties of DNA. Parameterization is performed by hand in order to fit the thermodynamic and mechanical behavior of DNA observed experimentally or using atomistic DNA simulations. The

model has been used to study DNA hybridization, strand exchange, hairpin formation, the response to mechanical stress and the formation of nanostructures, and it has reproduced many of the physical phenomena that were not used in its parameterization [127].

In the model, DNA is treated as a string of rigid nucleotides, with each nucleotide represented with a backbone moiety and a base moiety (Figure 4.2) [126]. The nucleotides interact through five types of interactions: (i) sugar-phosphate backbone interactions, (ii) coplanar stacking between adjacent bases, (iii) cross-stacking between a base and the nearest-neighbour bases on the opposite strand, (iv) hydrogen bonding that leads to base-pairing, and (v) excluded-volume interactions. The right-handed structure of double helices with anti-parallel arrangement of complementary strands is achieved by using appropriate parameters for the coplanar stacking and H-bonding. The model can be used either in Newtonian molecular dynamics simulations or with Monte Carlo sampling. The solvent is modelled implicitly, and diffusional dynamics are normally achieved either using Langevin dynamics or with Andersen-like thermostats [127].

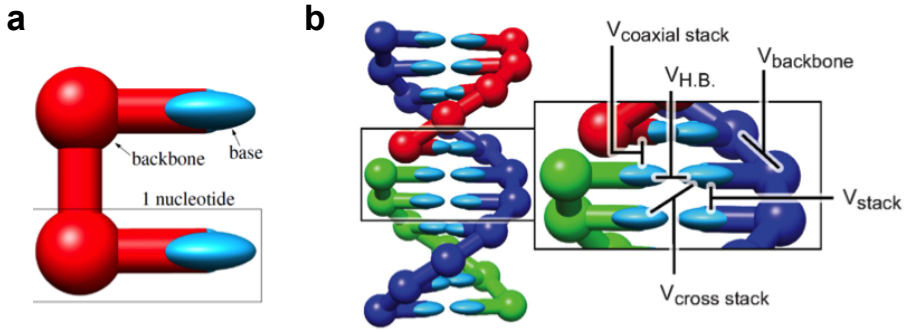


Figure 4.2: *OxDNA model.* ***(a)*** *Representation of DNA as a string of rigid nucleotides, consisting of the backbone and base moieties.* ***(b)*** *The different interactions modelled in OxDNA: backbone interaction ($V_{backbone}$), coplanar stacking (V_{stack}), cross-stacking ($V_{cross\ stack}$) and H-bonds (V_{HB}). Coaxial stacking between two non-bonded bases at a DNA nick is also shown. Adapted from reference [127].*

As any coarse-grained model, the OxDNA model is based on a number of simplifications [127]. The sequence dependence is limited to Watson-Crick base pairing, with some implementations also accounting for the different interaction strengths of the stacking and

H-bonding interactions [128], but not for other effects such as the size differences between the bases. In addition, no explicit electrostatic interactions are included, and for this reason parametrisation is done using experimental data measured at 500 mM NaCl concentration, which ensures that electrostatic properties are well screened. Finally, the double helix in OxDNA is symmetrical, with equal sizes for the major and minor grooves. Electrostatic interactions and the correct grooving have recently been implemented in a newer version of OxDNA [129].

4.2 Pol-DNA complex structure

4.2.1 Labelling scheme

To determine the structure of Pol in complex with a one-nucleotide gapped-DNA substrate, we aimed to measure a number of distances both within the DNA substrate, and between the DNA and Pol. For DNA-DNA distance measurements, we chose a number of positions in the upstream half of the DNA for labelling with the donor (Cy3b), and in the downstream half for labelling with the acceptor (Atto647N). In our initial experiments, 7 DNA labelling positions were used (3 upstream and 4 downstream), which provided a unique but still relatively low-precision structure. Based on this structure, new positions were designed and the corresponding hypothetical distances added to the existing distance set, in order to find the positions that would be predicted to improve the structure precision. This analysis led us to choose an additional 6 labelling positions, giving a total of 13 positions (6 upstream and 7 downstream, Figure 4.3a), which were used for the final distance measurements in Pol-DNA complex structure determination.

For Pol-DNA distance measurements, we labelled both upstream and downstream DNA positions with the acceptor, and chose three Pol positions for labelling with the donor. The rationale was to use a high number of DNA positions, due to the ease of DNA labelling, and use the minimum number of Pol positions required in order to uniquely determine the position of Pol relative to the DNA in 3-dimensional space. The Pol positions

selected were K550 and L744, which were substituted to cysteines to allow maleimide-based labelling, as well as C907, a native cysteine that could be labelled directly (Figure 4.3b). These Pol constructs have previously been used in single-molecule studies, and it has been shown that neither the substitutions nor the fluorescent labelling significantly affect their polymerization activity [13, 130].

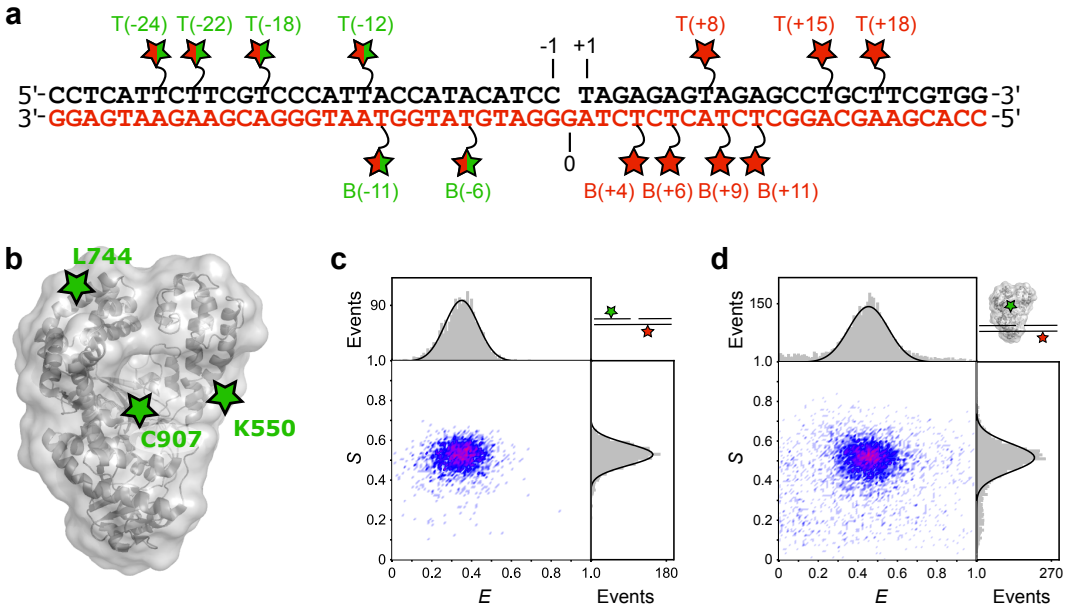


Figure 4.3: *Labelling scheme and example smFRET measurements for distance determination.* ***(a)*** *Labelling positions in one-nucleotide gapped DNA. Donor positions are shown with green and acceptor positions with red stars; mixed red-green stars refer to positions that were labelled with either donor or acceptor, depending on the distance being measured. The positions are annotated according to whether they are found in the non-template (top, T) or the template strand (bottom, B), and according to their position relative to the gap, with upstream positions given negative numbers.* ***(b)*** *The three donor positions in Pol, with the corresponding native residues indicated.* ***(c)*** *Example corrected E/S histogram from an smFRET experiment measuring a DNA-DNA distance.* ***(d)*** *As in (c) but for a DNA-Pol distance.*

4.2.2 Distance measurements

We performed single-molecule FRET measurements on diffusing molecules in solution, using confocal microscopy and alternating-laser excitation. We measured FRET efficiencies within the DNA and between the DNA and Pol, using constructs with one donor- and one acceptor-labelled position in each experiment (Figure 4.3c, d). Based on preliminary experiments, a Pol concentration of 3 nM appeared optimal for populating the mid-FRET species corresponding to the binary complex, and was used for DNA-DNA distance measurements. In contrast, Pol-DNA distance measurements had to be conducted at 100 pM Pol concentration, due to the limit on the concentration of fluorescent species that can be present in a single-molecule confocal experiment. At this concentration, a small but detectable amount of the binary complex was present, which could be distinguished from the remaining species by the single-molecule sorting capabilities of our ALEX setup. We measured a total of 34 DNA-DNA and 39 DNA-Pol FRET efficiencies, which we corrected for background fluorescence, donor fluorescence leakage into the acceptor channel, direct excitation of the acceptor, and the different detection efficiencies and quantum yields of the two dyes (Section 3.1.6). The resulting accurate FRET efficiencies were converted to FRET-averaged distances $\langle R_{DA} \rangle_E$, using experimentally determined R_0 values (see below), and then to mean dye positions R_{mp}, which we used for rigid-body docking calculations.

4.2.3 Förster radius determination

As described in Section 3.1.3, the Förster radius R_0 determines the relationship between the interdye distance and the observed FRET efficiency, and is dependent on the dye pair used, as well as on their molecular environment. Therefore, in order to account for the photophysical behaviour of our FRET dyes in the context of their attachment to DNA or Pol, we measured the isotropic Förster radii experimentally (Table 4.1). Whilst the overlap integrals were unaffected by the dye environment, the quantum yield of the donor dye (Cy3b) was significantly higher when the attachment point was on the DNA (0.79 to 0.81) as opposed to the protein (0.42 to 0.49). However, the quantum yields of the dyes attached

to the DNA were unaffected by Pol binding, and the exact Pol attachment site also had little effect on the observed quantum yield. Notably, steady-state anisotropies of the DNA- and Pol-attached dyes were sufficiently low (0.18 to 0.26) that dynamic averaging of dye reorientation could be assumed, allowing isotropic Förster radii to be calculated.

Two different isotropic R_0 values were thus used to account for the observed photophysics: 64.5 Å for DNA-DNA dye pairs, and 59 Å for DNA-Pol dye pairs. Docking with these R_0 values resulted in a very good agreement between the docked and X-ray structures (Figure 4.8), a significant improvement compared to when using a single R_0 value measured for the Cy3b and Atto647N free dyes (62 Å). We also estimated the anisotropic R_0 values, using a Monte Carlo simulation of the orientation factor [118]. Rigid-body docking with anisotropic R_0 values (63 Å for DNA-DNA and 57 Å for DNA-Pol dye pairs) returned the same structure (RMSD = 0.04 Å), but with a minor decrease in the goodness of fit to the data, compared to when using the isotropic R_0 values ($_{\triangle}\chi_r^2 = 0.15$).

Cy3b	Atto647N	QY (donor)	J (M^{-1} cm^{-1} nm^3)	iso R_0 (Å)	SSA (donor)	SSA (acceptor)	aniso R_0 (Å)
free dye	free dye	0.644	4.88×10^{15}	**62.3**	0.056	0.048	**62.0**
DNA	DNA	0.811	4.90×10^{15}	**64.8**	0.185	0.159	**63.7**
DNA (+Pol)	DNA (+Pol)	0.794	4.81×10^{15}	**64.4**	0.206	0.201	**62.9**
KF-550	DNA	0.436	4.98×10^{15}	**58.6**	0.214	0.201	**57.2**
KF-744	DNA	0.421	4.91×10^{15}	**58.1**	0.224	0.201	**56.7**
KF-907	DNA	0.486	4.91×10^{15}	**59.5**	0.265	0.201	**57.9**

Table 4.1: *R_0 measurements. The first two columns indicate the attachment points of the donor and acceptor dyes. QY, quantum yield; J, overlap intergral; iso, isotropic; aniso, anisotropic; SSA, steady-state anisotropy.*

4.2.4 Component structures

We generated a B-DNA model of the upstream and downstream fragments of the DNA sequence used in single-molecule experiments. In order to avoid any steric clashes with Pol during rigid-body docking, the DNA fragments were shortened by 3 nucleotides at their gap-proximal ends. Docking attempts with full-length DNA fragments resulted in

structures that had a very poor goodness of fit, and displayed unreasonable arrangements of the DNA relative to Pol (data not shown).

We used the *Bst* structure of Pol for rigid-body docking, even though our experiments were carried out using the *E. coli* protein. The reasoning was two-fold: first, crystal structures exist of the *Bst* protein in complex with the upstream DNA, allowing us to compare our structure to the existing data. Second, *Bst* structures have been solved at high resolution (1.80 Å, PDB codes 1L3U and 4BDP [8, 26]), whereas the best *E. coli* structure available (PDB code 1KLN, [23]) was solved at 3.20 Å resolution. The high resolution was particularly important for subsequent molecular dynamics simulations (Chapter 5), where molecular details were investigated. We justify using the *Bst* polymerase by noting that the *E. coli* and *Bst* proteins are highly homologous. The sequence present in the respective crystal structures (1L3U and 1KLN) aligns with 43.5 % identity and 71 % sequence similarity. Moreover, the proteins are structurally highly similar, and align with an RMS of 1.8 Å (Figure 4.4).

Figure 4.4: *Sequence- and structure-based alignment of E. coli (PBD code 1KLN [23], blue) and Bst (PDB code 1L3U [8], grey) homologues of Pol.*

4.2.5 Complex structure overview

We performed three-body rigid-body docking using the FPS software, starting from 1000 starting configurations of Pol, upstream and downstream DNA components. A number of complex-structure solutions were found using an initial search at high clash tolerance (Figure 4.5a). However, following refinement at 1 Å, a unique solution was obtained, with superior goodness of fit to any of the remaining solutions (χ_r^2 = 3.15, vs. >3.62). In this model (Figure 4.5b), the upstream DNA is positioned in the cleft between the thumb and the exonuclease domain of Pol, and is channelled into the active site, as observed in the crystal structures. The DNA then exhibits a stark bend of 120°, resulting in the downstream DNA fragment docking at the face of the fingers subdomain. Interestingly, if the downstream DNA is extended to its full length (i.e., if the 3 nucleotides that were removed for the docking purposes are added back), it causes a two-base-pair clash with the fingers of Pol, indicating that this part of the downstream DNA is unlikely to be double-stranded in the complex. The end of the downstream DNA is also found close to Pol residues previously implicated in strand separation, including Y719 and S717 [9, 10].

4.2.6 Model accuracy and precision

The agreement of the best model with the experimental data was generally high. We noted that the short and long distances showed poorer agreement with the data than the intermediate distances (Figure 4.6a), as expected due to the $1/r^6$-dependence of FRET efficiency, but the corresponding FRET efficiencies showed good agreement with the data for all distances (Figure 4.6b). Interestingly, the distribution of deviations between the model and experimental distances shows a slight skew towards the positive end (Figure 4.6c), indicating that the model distances tend to be longer than the experimental distances. This effect could not be due to steric clashing, as further shortening of the DNA fragments did not affect the docked structure or the distance skew. The skew was less pronounced with isotropic R_0 values, compared to anisotropic R_0 values, and was one of the reasons for using the isotropic values in preference (Section 4.2.3). The remaining skew could be accounted

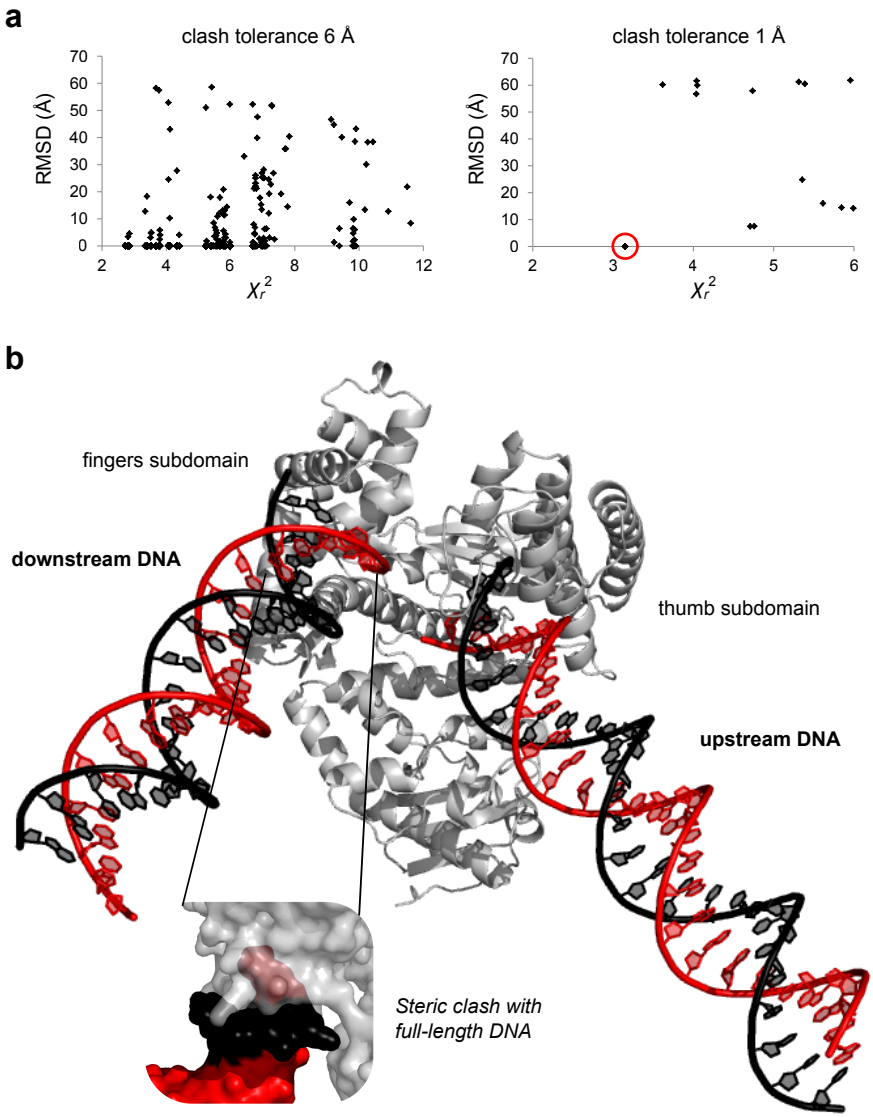


Figure 4.5: *Pol-DNA complex structure.* ***(a)*** *Distribution of structure solutions, in terms of their χ_r^2 values and their RMSD values compared to the best (lowest-χ_r^2) solution, from the initial search with 6-Å clash tolerance (left), and after refinement with 1-Å clash tolerance. The best solution is marked with a red circle.* ***(b)*** *Docked structure of the complex. The template and non-template strands of the DNA are shown in red and black, respectively. The inset shows the steric clash between the DNA and Pol resulting from the extension of downstream DNA to its full length. The protein is shown transparent to reveal the clash.*

for by minor systematic errors in the R_0 or FRET-efficiency determination, although its definite cause is unclear. Notably, no outlier labelling positions were observed that would consistently show high deviations between the model and the experiment in all distances.

The precision of the complex structure was estimated by a bootstrapping approach,

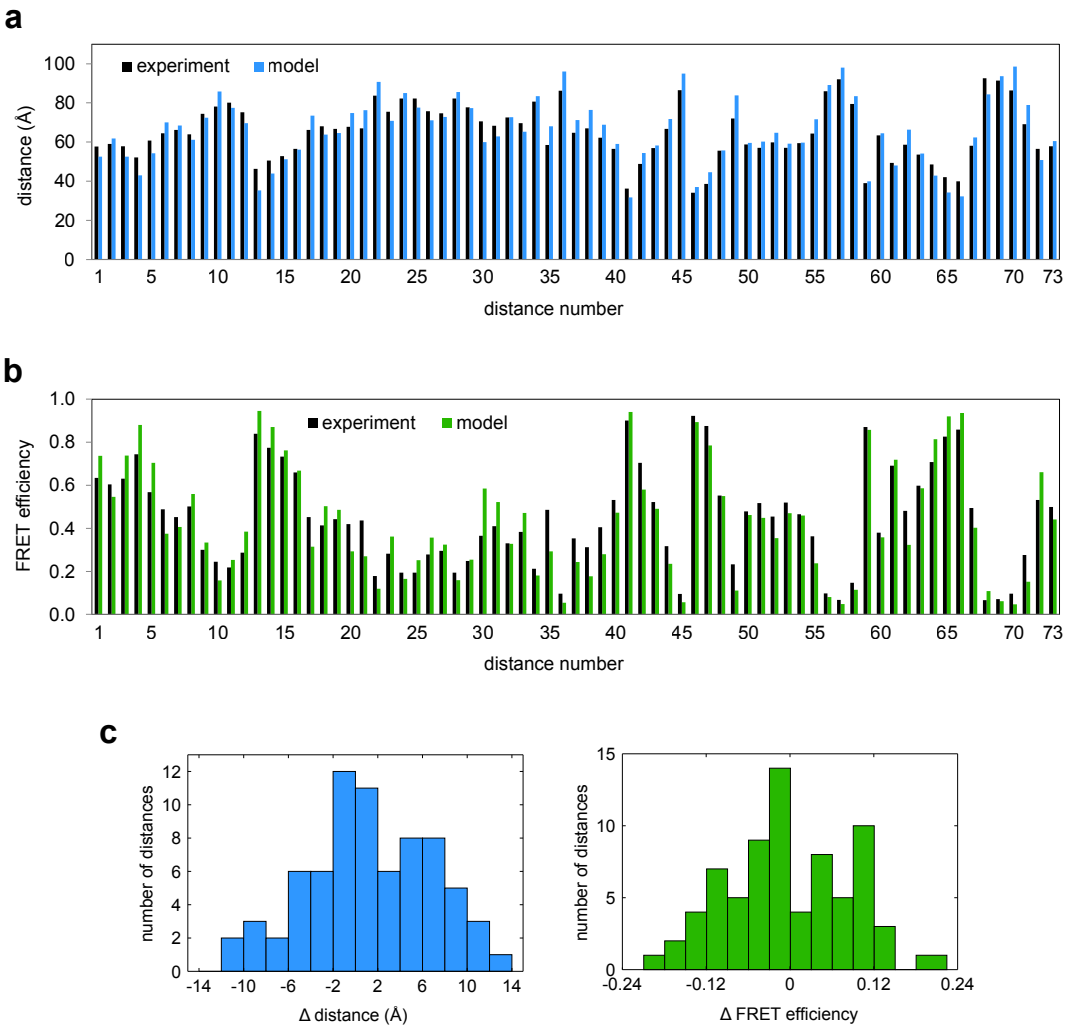


Figure 4.6: *Agreement between the model and the experiment, in terms of* ***(a)*** R_{mp} *distances, and* ***(b)*** *corresponding FRET efficiencies, across all distances used in rigid-body docking.* ***(c)*** *Histograms of distance deviations (left) and FRET deviations (right) between the model and the experiment.*

whereby 100 structures were generated by modifying the best-model distances by random errors (Section 4.1.3). The resulting ensemble of structures were seen to deviate only slightly from the original model (Figure 4.7), indicating high model precision. To quantify the model precision, we calculated the RMSD of each phosphorous atom of the DNA across all 100 structures. The RMSD was found to vary between approximately 3 and 6 Å, and generally increased with increasing distance from Pol, due to the lower degree of steric restraint in this region of the structure (Figure 4.7, right). Similarly, the deviation was higher for the downstream DNA than for the upstream DNA, again because the protein-proximal

end of the latter is sterically more restrained.

Varying the estimate for the experimental error in FRET and R_0 determination had a negligible effect on the final structure, but did significantly change the spread of the bootstrapped structures. We noted that the RMSD plots consistently showed a helical pattern, which corresponded to one face of the DNA always showing higher deviations than its opposite face. However, the orientation of the high-deviation face of the DNA relative to Pol differed depending on the input values of the experimental error, and could not be explained visually in terms of the steric hindrance effects.

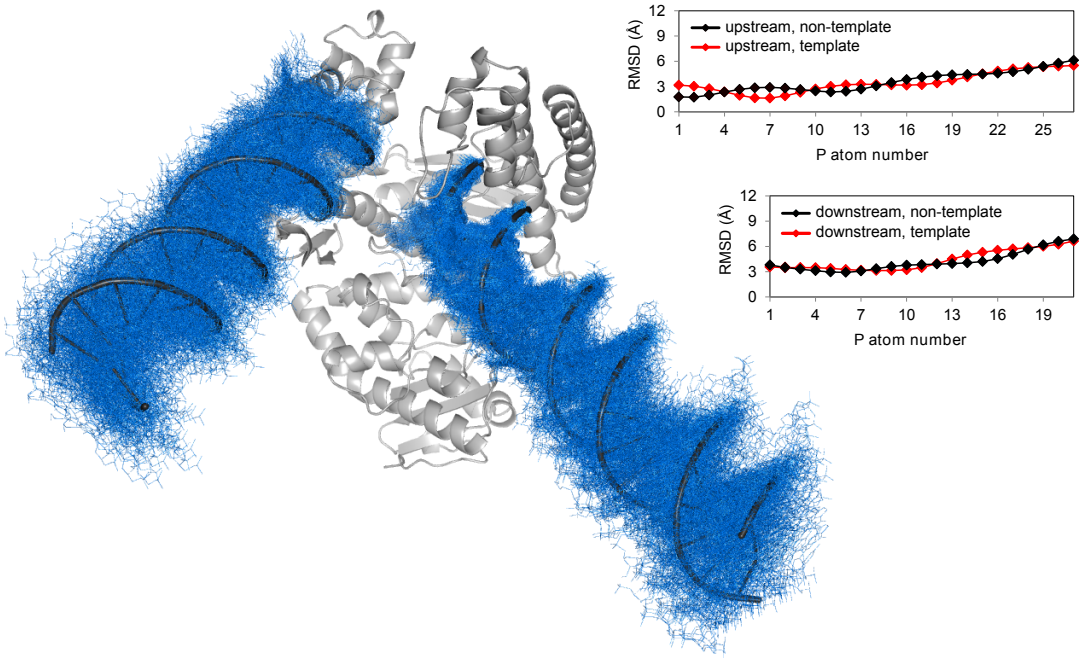


Figure 4.7: *Model precision. Left, an overlay of the best model (black) and 100 bootstrapped structures (blue), indicating the range of structures consistent with experimental data. Right, RMSD of each phosphorous atom of the DNA across all the bootstrapped structures, shown separately for each of the four DNA strands.*

4.2.7 Comparison with X-ray structures

In order to compare our FRET-restrained structure to the existing X-ray structures, we aligned the polymerase component of our structure to the *Bst* X-ray structure that was

crystallized with a fragment of upstream DNA bound at the active site (PDB code 1L3U, Figure 4.8a). The fragment overlays very well with the corresponding portion of the upstream DNA present in our structure, with an RMSD of 2.9 Å. Notably, the agreement is limited by the fact that our upstream DNA is modelled as B-DNA, whereas the X-ray DNA adopts a mixture of B-DNA and A-DNA conformations (Section 2.4.1), and hence an even better agreement would be expected if the X-ray DNA was used as one of the components in rigid-body docking.

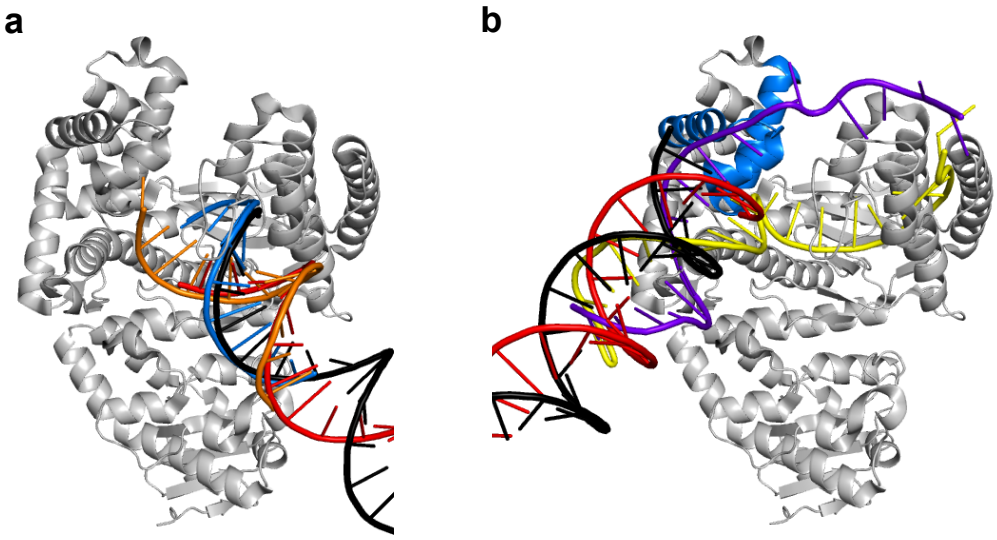


Figure 4.8: *Comparison of FRET-restrained complex structure with X-ray structures.* ***(a)*** *Comparison of positions of upstream DNA in the docked structure and the DNA fragment in the Bst X-ray structure (PDB code 1L3U, reference [8]), after alignment with respect to the protein component. The docked DNA is shown in red (template) and black (non-template strand), and the X-ray DNA in orange (template) and blue (non-template strand).* ***(b)*** *Comparison of positions of downstream DNA in the docked structure and the T7 RNAP structure (PDB code 1S67, reference [47]), after alignment with respect to the conserved 3-helix bundle (blue). The protein component of the RNAP structure is removed for clarity; the DNA is shown in yellow (template) and purple (non-template strand).*

Notably, Pol shares a structural motif with another strand-displacing polymerase, T7 RNA polymerase [131]. The motif is a three-helix bundle, and is seen to bind downstream DNA in the T7 RNAP crystal structure (Section 2.4.6) [47]. If the two proteins are aligned

with respect to the three-helix bundle, the resulting position of the downstream DNA in our docked structure is very close to the DNA in the RNAP structure (Figure 4.8b). Considering that the two proteins are substantially different in terms of their structure and function, the similar positioning of the downstream DNA at the three-helix bundle may indicate a conserved mechanism of strand separation.

4.3 DNA structure in complex with Pol dimer

Our Pol-DNA titration experiments indicated that an additional high-FRET peak was populated at high Pol concentration (~10 nM) for many of the DNA-DNA distances measured, corresponding to the Pol_2-DNA ternary complex. We therefore attempted to determine the structure of the gapped-DNA substrate in this complex, using the same approach as above. The high-FRET population could in fact be extracted even at 3 nM Pol concentration, and thus additional experiments at higher Pol concentrations were not necessary. For 12 pairs of positions, no high-FRET peak was observed, suggesting that these distances were too similar in the binary and ternary complexes to be resolvable. For the remaining 22 pairs, high-FRET peaks could be resolved, and were used to obtain 22 distances between the upstream and downstream DNA in the ternary complex. This allowed two-body docking of the DNA fragments, each shortened by 3 base-pairs, in the absence of Pol.

We obtained two good-quality solutions, with χ_r^2 values of 1.02 and 1.28 (Figure 4.9a, b). Both structures showed a high bend angle between the upstream and downstream DNA (140° and 141°, respectively), with the DNAs being farther apart from each other in the lower-χ_r^2 structure. When docking was repeated with additional distances from the dye pairs for which the two FRET peaks could not be resolved (i.e. the distances that remain the same in the monomer and dimer complexes), two solutions were returned again. The best solution was essentially the same as the best solution obtained when only the novel distances were used, whereas the second-best solution was different from the previous second-best solution. This led us to conclude that the lower-χ_r^2 structure was likely the correct structure of the DNA in complex with the Pol dimer. The agreement of this model

with experimental data (Figure 4.9c) was comparable to the binary complex structure.

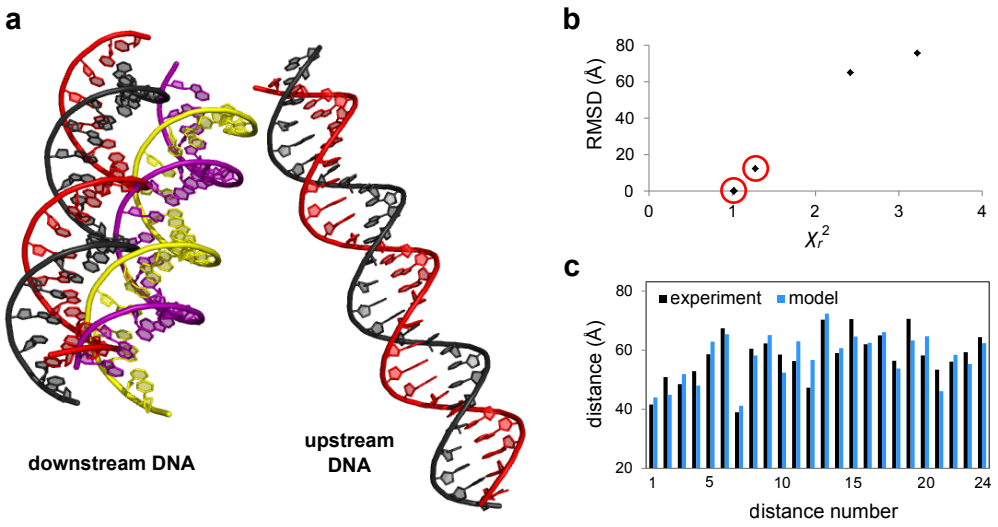


Figure 4.9: *Structure of DNA in Pol_2-DNA complex.* ***(a)*** *Best two solutions obtained when using only the novel distance set, aligned relative to upstream DNA. The best model is shown in red (template) and black (non-template strand); the 2nd best in yellow (template) and purple (non-template strand).* ***(b)*** *Distribution of structure solutions, in terms of their χ_r^2 values and the RMSD values compared to the best solution, after refinement with 1-Å clash tolerance. The two solutions shown in (a) are marked with red circles.* ***(c)*** *Agreement between the model and the experiment, in terms of R_{mp} distances, across all distances used in the docking.*

The precision of the structure could then be evaluated as previously, by calculating the phosphate-atom RMSDs across the bootstrapped structures (Figure 4.10a). Based on this analysis, the precision of the dimer structure was found to lie in the range of 8 to 16 Å, with an average RMSD of 12 Å. This is significantly lower than the precision of the binary Pol-DNA complex structure, as expected due to the lower number of distance restraints, and the lack of steric restraints that results from the absence of Pol during docking. The spread of the bootstrapped structures shows that, whilst the bend angle is quite well defined, the position of the downstream DNA in the perpendicular dimension (in/out of the page when viewed as in Figure 4.10) is more uncertain.

Assuming that the upstream DNA is bound by one of the two Pol molecules in the dimer in a similar fashion as in the binary Pol-DNA complex, the two structures can be

aligned with respect to the upstream DNA (Figure 4.10b). This allows comparison of the position of the downstream DNA in the two structures: the DNA is found at the face of the fingers subdomain in both cases, but its gap-proximal end is slightly more removed from Pol in the dimer structure. This shift accounts for the observed 21° difference in the bend angle between the two structures.

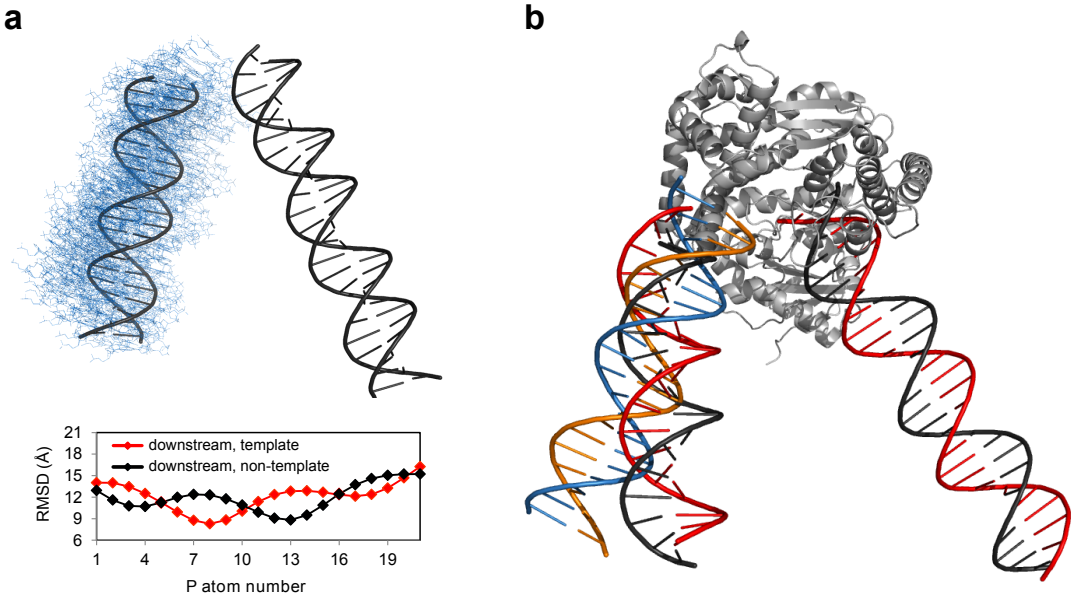


Figure 4.10: *Further analysis of DNA structure in Pol_2-DNA complex.* ***(a)*** *Model precision. Top, overlay of the best model (black) and 20 bootstrapped structures (blue). Bottom, RMSD of each phosphorous atom of downstream DNA across all bootstrapped structures, shown separately for each of the two strands.* ***(b)*** *Comparison of positions of downstream DNA in Pol dimer and monomer structures, after alignment with respect to upstream DNA. The downstream DNA of the dimer complex is shown in red and black; the monomer in orange and blue.*

4.4 DNA substrate structure

In order to understand the mechanism of substrate recognition and binding by Pol, we aimed to determine the structure of the gapped-DNA substrate itself. The same labelling positions were used as previously (Figure 4.3a), and 34 DNA-DNA distances were measured between the upstream and downstream DNA, now in the absence of Pol. Two-body docking was performed, with the DNA fragments at their full length, and with an additional distance constraint imposed between their gap-proximal ends, to account for the fact that the template strand is continuous.

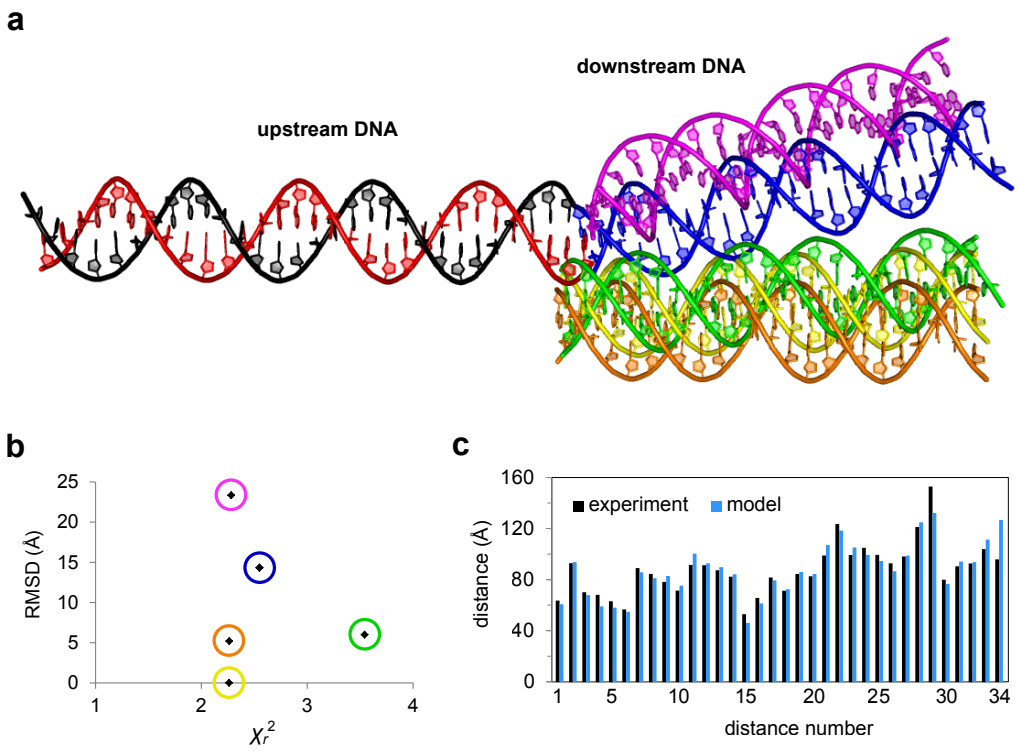


Figure 4.11: *DNA substrate structure.* ***(a)*** *The five solutions obtained from docking, aligned relative to upstream DNA (shown in red and black). Downstream DNA is shown in yellow (best), orange (2nd best), purple (3rd best), blue (4th best) and green (5th best solution).* ***(b)*** *Distribution of structure solutions, in terms of their χ_r^2 values and their RMSD values compared to the best solution, after refinement with 1-Å clash tolerance. The solutions are marked with circles according to the colouring in (a).* ***(c)*** *Agreement between the best model and the experiment, in terms of R_{mp} distances, across all distances used in the docking.*

We obtained five solutions, four of which showed similar goodness of fit to the data (χ_r^2 between 2.26 and 2.55; see Figure 4.11a, b). In all structures, the downstream DNA exhibited a slight twist relative to the upstream DNA, and was bent by 8-25° relative to the duplex conformation. The low structure uniqueness is in part due to its elognated shape, and the presence of many long distances (> 90 Å) that are not significantly affected by minor bends or twists. The agreement between the model and the data (shown for the best model in Figure 4.11c) was similar across the five models, with the same distances appearing as outliers in all models.

4.5 DNA substrate simulations

Whilst the docked model of the gapped-DNA substrate is instructive, it likely corresponds to an ensemble of many dynamic, interconverting structures, averaged over the time course of the experiment. In order to understand the binding mechanism of Pol, a more detailed picture of the conformational dynamics of the DNA substrate is needed. Coarse-grained models can provide sufficient detail and allow simulations on a microsecond time scale, ensuring efficient sampling of the conformational space available to the DNA. Therefore, we modelled our gapped DNA using the OxDNA coarse-grained representation (introduced in Section 4.1.4), and performed 100 simulations of 10^8 steps each, corresponding to ~1.5 μs each [3].

In the simulations, the DNA substrate was found to adopt a variety of straight and bent conformations, switching between them on a nanosecond timescale. Interestingly, we found that the stacking interactions between the three nucleotides opposite the gap were a key determinant of the gapped-DNA dynamics. When both interactions were formed, the DNA substrate maintained a largely straight conformation, which accounted for ~80 % of the simulation time. In contrast, when either of the two interactions was broken (~20 % of the simulation time), the DNA could adopt higher bend angles (Figure 4.12a).

[3]The physical time scales can only be approximated, due to the incorrect scaling of diffusion rates and the smoothing of free-energy landscapes inherent in OxDNA [127].

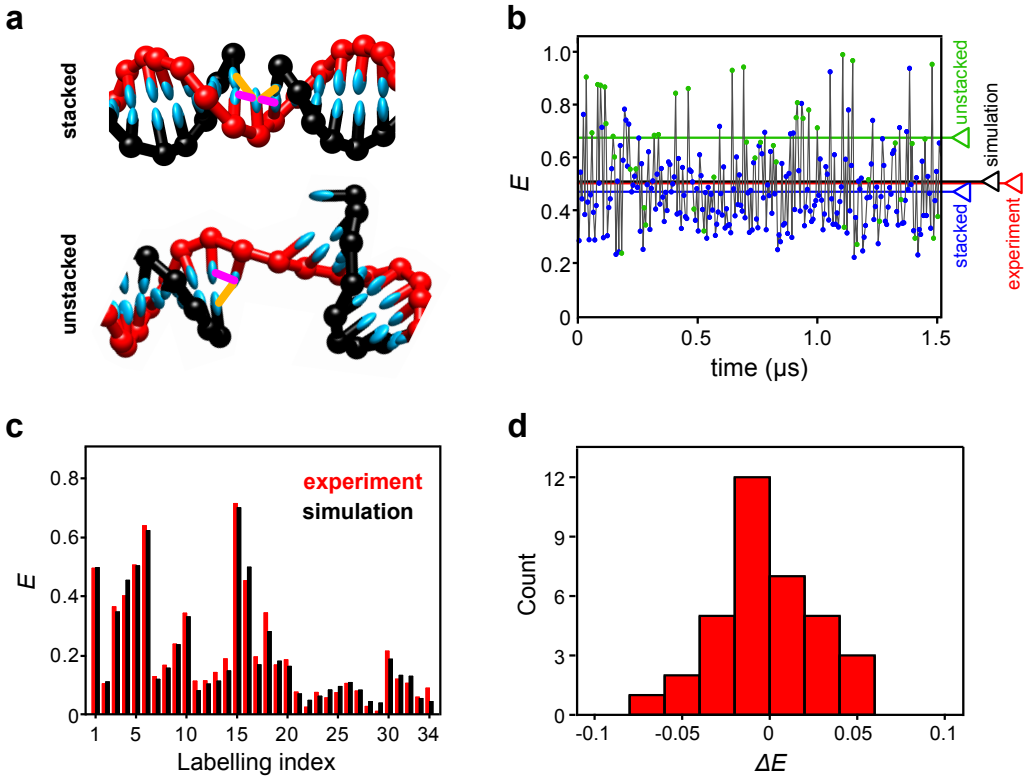


Figure 4.12: *Coarse-grained simulations of gapped-DNA substrate.* ***(a)*** *Example snapshots of OxDNA-modelled gapped DNA in stacked and unstacked configurations. The template and non-template strands are shown in red and black; the base moieties in cyan, and the stacking and cross-stacking interactions in pink and yellow, respectively.* ***(b)*** *Expected FRET efficiency as a function of time, for construct T8/B-11. Each data point is labelled blue or green, depending on whether the DNA is in the stacked or the unstacked configuration. The average FRET efficiency observed in the simulation and the experimentally measured efficiency are both indicated.* ***(c)*** *Comparison of FRET efficiencies between the simulation and the experiment, across all 34 distances measured.* ***(d)*** *Histogram of deviations between the experimental and modelled FRET efficiencies, derived from data in (c).*

We modified the accessible volume approach to model our FRET dyes on the DNA, and calculated the FRET efficiency expected from each dye pair as a function of time during the simulation (Figure 4.12b). Again, significant FRET efficiency variations occurred on a timescale much faster than the temporal resolution of our single-molecule confocal experiments (nanosecond, versus the experimental ~1 ms). However, when the FRET efficiencies observed in the simulation were averaged over time, we obtained an excellent agreement between the modelled and experimental FRET efficiencies, with $_{\Delta}E$ of -0.0024

across the 34 distances (Figure 4.12c, d). Notably, the average FRET efficiencies of unstacked configurations were generally higher than the average efficiencies of the stacked configurations, due to the difference in flexibility of the two. However, if only the average FRET efficiencies arising from the stacked conformations are compared to the experiment, the agreement is considerably worse ($_{\Delta}E$ of -0.0249), suggesting that the unstacked states are an important component of the experimental structural ensemble of gapped DNA.

In addition, to understand how gapped-DNA dynamics differed from other DNA constructs, we ran our coarse-grained simulations on the nicked and (intact) duplex DNA. As before, 100 simulations of ~1.5 μs each were performed. In each simulation, we measured the frequency of any bend angle being observed during the simulation, and calculated the corresponding relative free energies using Boltzmann distribution (Figure 4.13). The results suggest that the gapped DNA can access higher bend angles much more easily than the nicked or duplex DNA. In order to access the bend angle observed in the Pol-DNA complex structure (120°), the free-energy cost is less than 4 kT for the gapped substrate, compared to ~7 kT for the nicked DNA and > 15 kT for the duplex DNA. Interestingly, we found the free-energy landscape to be much flatter for the unstacked configurations of gapped DNA, compared to its stacked configurations, with almost no free-energy cost involved in the bending.

Finally, we measured the degree of base-pairing in the downstream portion of the gapped DNA, to see if the unpairing observed in the Pol-DNA complex can occur in the absence of the polymerase. We found that the A-T base pair immediately adjacent to the gap was melted in 28 % of the configurations in our simulations. If only the unstacked configurations were taken for analysis, the frequency of melted configurations increased to 35 %, which can be accounted for by the loss of cross-stacking interactions present in the stacked state. The melting extended to two base pairs in 4 % of the configurations, which increased to 5 % when only the unstacked configurations were considered.

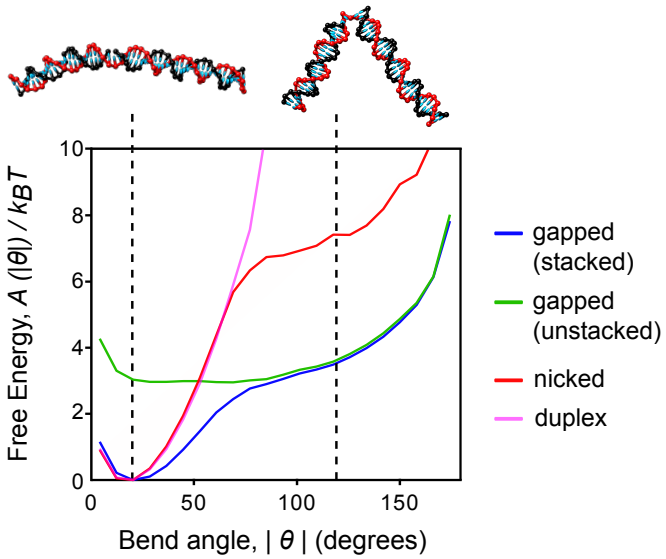


Figure 4.13: *Relative free-energy landscape for gapped, nicked and duplex DNA bending, calculated from coarse-grained simulations. The gapped DNA is shown in blue and green for the stacked and unstacked configurations, respectively, the nicked DNA is in red and the duplex DNA in pink. Snapshots indicate the most energetically stable DNA conformation (top left) and a conformation adopting the same bend angle as in the Pol complex structure (top right). The corresponding bend angles are indicated with vertical dashed lines.*

4.6 Discussion

4.6.1 Pol-DNA complex structure

We obtained a unique and high-precision structure of the Pol-DNA complex. The location of upstream DNA in the docked structure agrees very well with the existing X-ray structures, which gives weight to the rigid-body docking approach, and suggests that the position of downstream DNA is likely to be very accurate. The position of downstream DNA on the face of the fingers subdomain conclusively rejects the early propositions that the DNA might be channelled through the cleft formed by the fingers and thumb subdomains [21, 23]. The observed position is also consistent with the mutagenesis studies, which highlighted residues R784 and R789 of the fingers as being important for downstream DNA binding (Section 2.4.1 and Figure 2.5b) [9]. Finally, the DNA is in close proximity to residues S717 and Y719, consistent with the involvement of *E. coli* residues

S769 and F771 in strand-displacement synthesis (Section 2.4.6) [10].

In the docked structure, the DNA substrate exhibits a stark, 120° bend. DNA bending has also been observed in the crystal structure of the mammalian gap-filling DNA polymerase β, where the 90° bend was suggested to be important for the mechanisms of polymerisation and fidelity [115]. In particular, the role of the bending would be to expose the templating base for interrogation by the incoming dNTP and increase the surface area over which the polymerase could test the base pair for complementarity. Consistent with this interpretation, a substitution of residue T79, which appears to assist in the stabilization of the bent conformation of DNA, has been shown to significantly compromise the fidelity of polymerase β [132]. In the case of Pol, the conformation of the templating base appears to be controlled primarily by the 'base-flipping' mechanism (Figure 2.6), although DNA bending could enhance the conformational flexibility of the templating base. In addition, DNA bending in Pol is likely also important for the mechanism of substrate recognition, as we discuss below.

It is important to also review the docked Pol-DNA structure from a technical standpoint. Notably, rigid-body docking is guided by two types of restraints: (i) distance restraints, in this case derived from FRET data, which provide information on the relative positioning of the macromolecules, with some degree of error, and (ii) steric restraints, which exclude certain arrangements of the complex that result in steric clashes [120]. In our case, steric restraints are low, due to the non-compact shape of the structure and the low surface area of Pol-DNA contacts, and hence a large number of distances are required to define a unique structure. Whilst removing 20 out of the 73 distances from our dataset still returned the 'correct' complex structure as one of the solutions, the solution was no longer unique, and it was difficult to distinguish between the different similar-quality models.

In addition, the precision of the docked structure also increases with the number of input distances, until eventually plateauing out, as was previously predicted for small-molecule ligand docking to macromolecular complexes [133]. The high number of distances used in this study implies that the plateau has likely been reached, and that any

further increase in precision would require additional steric restraints, or higher-precision FRET data. Interestingly, when rigid-body docking was performed using distances that were banded into three classes (short, 42 +/- 12 Å; medium, 62 +/- 8 Å; and long, 82 +/- 12 Å), we again found that the 'correct' structure no longer stood out as superior to the alternative models, which stresses the importance of using precise distances in determining a unique, high-precision solution of the complex structure.

4.6.2 Pol dimerization

The structure of the DNA substrate at high Pol concentration, corresponding to the ternary complex of the DNA and two Pol molecules, shows an even greater bend angle between the upstream and downstream DNA. The functional role of the greater bend angle, or indeed of the Pol dimer, is currently unclear. Pol dimerization on primer-template DNA has previously been observed using DNA gel-shift assays and sedimentation-equilibrium experiments, and it has been suggested that the second Pol molecule would bind at the upstream end of the DNA [134]. However, given the tight grip of the thumb domain of Pol around the upstream DNA, it is difficult to imagine how the binding of a second Pol molecule in that region could result in additional bending of the DNA substrate. Hence, we suggest that the second Pol molecule is more likely to bind at the downstream end, where Pol-DNA contacts are looser and additional modulation of the substrate structure may be possible.

Further, using DNA gel shift assays, it was shown that the dimeric form of Pol was the dominant species in complex with the primer-template DNA and the correct dNTP, but only the monomeric form could be detected in complex with the primer-template DNA and a mismatched dNTP [134]. Based on these observations, it was suggested that the 3'-5' exonuclease domain of Pol would bind a second Pol molecule whilst in the polymerization mode, but would release it during the editing mode. The role of the second polymerase could be to enable DNA proofreading simultaneously with polymerization, resulting in a higher catalytic efficiency. Dimerization has also been observed with a variety of other

DNA polymerases and has been particularly well characterized for Pol β, with similar proposed functional roles [135].

Although concentration-dependent aggregation has been observed for Pol [16, 134], the concentration of Pol used in our experiments is 3 nM, which is considerably lower than the estimated concentration of full-length Pol in cells. The latter is expected to be on the order of 400 nM, assuming a previously measured value of ~400 Pol copies per cell [15], and the cell volume of 1 fl. Hence, the observed dimer formation *in vitro* is unlikely to be an effect of an unreasonably high Pol concentration. However, it is possible that dimerization is specific to Klenow fragment, and indeed no study has yet reported dimerization for the full-length Pol, or indeed for any Pol construct *in vivo*. We probe the physiological relevance of Pol dimerization in Chapter 8, using smFRET measurements in live *E. coli*.

4.6.3 DNA substrate structure and simulations

The coarse-grained simulations show high dynamics of the gapped-DNA substrate on a nanosecond time scale, confirming the speculation that our rigid-body docked structure of the substrate represents an average of all of its interconverting conformations. However, we see an excellent agreement in FRET efficiencies between the coarse-grained simulations and the single-molecule experiments across all 34 distances measured, suggesting that the simulations sample a representative range of experimentally observed DNA conformations. Both the stacked and unstacked states need to be taken into account to observe good agreement, indicating that although the DNA spends the majority of the time in a stacked state, the bent conformations arising from the unstacked states are an important component of the experimental structural ensemble of gapped DNA.

Only limited structural and dynamic information is available for gapped DNA in the literature. NMR data on a 13-mer DNA containing a 3'-phosphoglycolate / 5'-phosphate gapped lesion indicated a canonical B-form structure, with the base-pairs adjacent to the gap remaining stacked in the DNA duplex [136]. Similarly, an NMR study of a 14-mer DNA with a full one-nucleotide gap observed a mainly B-form structure [137]. However,

MD simulations of the same DNA construct indicated two families of structures, one of which was significantly kinked at the gap, with bend angles of up to 40°. It should be noted that these simulations were only 500 ps in length, and thus were very limited in terms of conformational sampling. It is possible that the bent states of the gapped DNA could not be observed using NMR due to their short lifetime, in the same way that they were obscured in our FRET-restrained structure. In addition, the DNA constructs used in these studies were relatively short in length compared to our gapped-DNA substrate (13-14 vs. 55 nucleotides), and are thus likely to be less conformationally flexible.

To our knowledge, gapped DNA has not been simulated at nano- or microsecond time scales before, but MD simulations using umbrella sampling have been used to investigate the dynamics of 15-mer DNA duplexes in the presence and absence of base mismatches [138]. The free-energy plots of duplex DNAs in the absence of mismatches agree with the behaviour observed here, showing energy minima at bend angles of ~20°, and a steady increase in free energy for the higher bend angles. However, the free-energy values observed in all-atom MD simulations are ~2-times greater than the values we observe, most likely reflecting the effects of explicit solvent and electrostatic interactions absent in the OxDNA model. We probe the gapped-DNA dynamics further in the next chapter, using all-atom MD simulations, albeit only at short time scales.

The results of the coarse-grained simulations have several important implications for the binding mechanism of Pol. Since the breaking of the stacking interactions opposite the gap increases DNA bendability, unstacking will likely occur as a step on route towards Pol binding. In addition, the high flexibility of the unstacked DNA suggests that the substrate can adopt a close-to-final conformation even prior to Pol complex formation. Finally, the simulations provide an explanation for Pol substrate specificity, and its increasing binding preference for the duplex, nicked and gapped DNA, previously observed in gel shift assays [139]. The preference appears to arise from the differential flexibility of these DNA structures, and the different energy cost required for their bending. Therefore, the substrate specificity is encoded in the structure and dynamics of the DNA substrate itself, allowing sequence-unspecific recognition of damaged DNA by Pol.

4.6.4 Binding mechanism

From the above observations, and the structure of the Pol-DNA complex, it is possible to suggest a stepwise mechanism for Pol binding to gapped DNA (Figure 4.14). In its stacked configuration, the substrate DNA is found mainly in a straight, duplex-like conformation, displaying only limited flexibility. Occasional unstacking of the nucleotide opposite the gap increases the flexibility of the DNA and allows downstream DNA to be displaced sufficiently to allow upstream DNA to dock to the polymerase. Single-molecule FRET titrations of Pol binding to the upstream DNA have indicated a K_d of 3 nM, whereas the K_d for full-length DNA binding is on the order of 0.3 nM (T. Craggs et al., unpublished). Therefore, we suggest that the tight interaction between the upstream DNA and Pol would precede downstream-DNA binding. At this stage, the downstream DNA may still exhibit high flexibility, until finally docking to the polymerase, which would provide some additional stabilization.

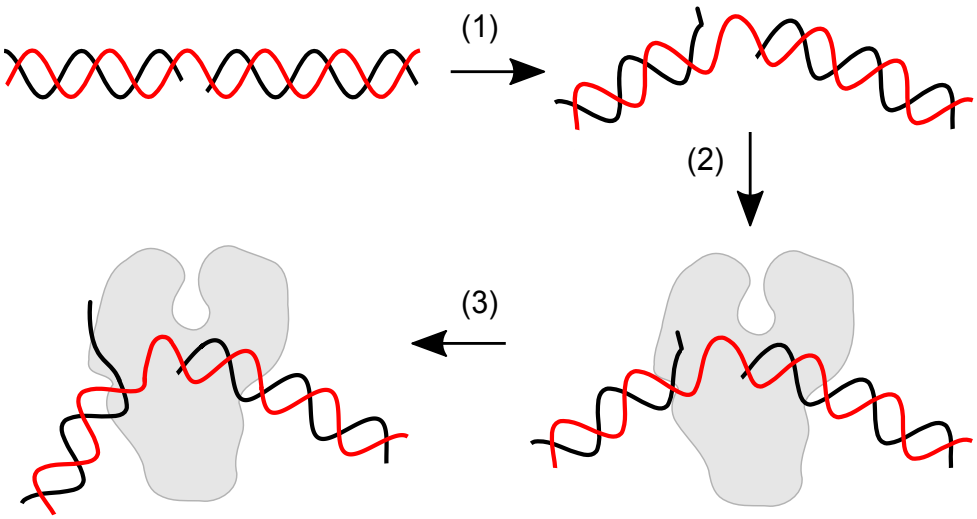


Figure 4.14: *Proposed mechanism of Pol binding to gapped DNA. Initially, DNA is found largely in a duplex-like conformation.* ***(1)*** *Nucleotide unstacking allows increased flexibility in the DNA, which can now adopt a variety of bent states. DNA fraying also occurs at the gap.* ***(2)*** *With downstream DNA displaced, upstream DNA can bind to Pol. Downstream DNA may still be dynamic at this point, and may continue to probe its conformational space.* ***(3)*** *Downstream DNA also finally docks to Pol, resulting in the conformation observed in the complex structure.*

4.7 Conclusions and future work

In this chapter, we have successfully combined experimental and computational approaches to investigate the structure of Pol bound to its gapped-DNA substrate, as well as the underlying binding mechanism. The complex structure contains novel information on the position of the downstream DNA relative to the polymerase, which was inaccessible to conventional structural approaches. The DNA is highly bent, with the downstream DNA positioned close to the residues known to be involved in strand displacement. In addition, we have simulated the gapped-DNA substrate at a microsecond time scale using coarse-grained OxDNA modelling, and have shown that it can access highly bent states, unlike the nicked or duplex DNA. Taken together, the complex structure and the results of the substrate simulations allow us to propose a two-step model for gapped-DNA binding by Pol, and explain the sequence-independent substrate specificity of Pol.

Notably, the rigid-body docked structure of the Pol-DNA complex leaves some questions unanswered, such as the degree of downstream-DNA unwinding and the atomistic details of Pol-DNA interactions, which we address in the next chapter. In addition, it should be noted that the structure was determined with the shortened construct (Klenow fragment) of Pol, and thus it would be of interest to probe the Pol-DNA structure with the full-length protein. With this capability, Pol-DNA complexes formed in other modes of catalysis (proofreading and flap-excision) could also be investigated. In addition, intramolecular distances could be measured to explore the domain arrangement in full-length Pol, thus resolving the ambiguity of full-length Pol structure arising from the X-ray models. Finally, the biological relevance of the ternary Pol_2-DNA complex observed here is unclear, and should be confirmed by repeating the smFRET experiments with full-length Pol. We also aimed to probe the Pol-DNA and Pol_2-DNA complex formation in live cells, using internalization by electroporation and single-molecule FRET, as we report in Chapter 8.

4.8 Materials and methods

4.8.1 Protein and DNA labelling

Pol variants were previously expressed from an N-His_6, D424A construct and purified, essentially as described in Section 6.7.6. The variants were singly labelled with maleimide derivatives of Cy3b (GE Healthcare) or Atto647N (Atto-tec), as described in [13]. Variant C907 was labelled on the native cysteine, whereas the other variants contained additional substitutions to allow specific labelling (C907S / K550C and C907S / L744C). DNAs were prepared by automated synthesis (IBA GmbH), and were labelled with NHS-ester derivatives of Cy3b (GE Healthcare) and Atto647N (Atto-tec) via dT-C6 linkers at selected positions. DNA strands were annealed in a buffer consisting of 20 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM ethylenediaminetetraacetic acid (EDTA). Samples were heated to 94 °C and subsequently cooled to 4 °C, in steps of 10 °C over 45 minutes.

4.8.2 smFRET measurements

For DNA-DNA measurements, labelled DNA was present at < 100 pM and unlabelled Pol (when present) at 3 nM concentration. For Pol-DNA measurements, both Pol and DNA were present at 100 pM concentration. All dilutions were done into 'Pol buffer', consisting of 40 mM Hepes-NaOH pH 7.3, 10 mM $MgCl_2$, 1 mM dithiothreitol (DTT), 100 µg/ml bovine serum albumin (BSA), 5 % glycerol, 1 mM mercaptoethylamine. Single-molecule FRET measurements were performed at room temperature using a home-built confocal microscope with 20 kHz alternating-laser excitation between a 532-nm (Samba, Cobolt, operated at 240 µW) and a 638-nm laser (Cube, Coherent, operated at 60 µW), coupled to a 60x, 1.35 numerical aperture (NA), UPLSAPO 60XO objective (Olympus) [140, 141]. 3-6 datasets of 10 min were recorded for each distance measurement and combined for the analysis.

Photon streams in DD, DA and AA channels were recorded and processed using custom-written software, and burst search performed by means of a published algorithm [71].

The bursts were filtered for the correct labelling stoichiometry, and accurate FRET calculated as described in Section 3.1.6. Corrected FRET histograms were fitted to single, double or triple Gaussian functions, constrained by a maximum width of $\sigma \leq 0.07$.

4.8.3 Distance calculations

Accurate FRET efficiencies were converted to their corresponding $\langle R_{DA} \rangle_E$ distances by assuming a FRET error of +/- 0.025, and using experimentally determined R_0 values. The R_0 values used were 64.5 Å for DNA-DNA and for 59.0 Å for DNA-Pol distances, and the error in R_0 was assumed to be the error that was propagated from the uncertainty in quantum yield determination of +/- 0.10. The resulting mean, maximum and minimum $\langle R_{DA} \rangle_E$ values were converted to the distances between mean dye positions R_{mp}, using a third-order polynomial function that was established by calculating R_{mp} and $\langle R_{DA} \rangle_E$ values for pairs of dyes at different positions along a double-stranded DNA.

4.8.4 Förster radius determination

Quantum yields were measured according to established methods [52, 142], for the following donor samples: free Cy3b-maleimide dye, Cy3b attached to gapped DNA (in the presence and absence of unlabelled Pol in a 1:1 molar ratio), and Cy3b attached to different positions of Pol (K550, L744, C907). Each sample was diluted from a glycerol stock to 5 μM final concentration in Pol buffer. Free Cy3b-maleimide dye was reduced with 10 mM DTT for 10 min prior to dilution. Absorbance at 490 nm was recorded for each sample, using a UV-visible spectrophotometer (Cary 50 Bio, Varian). An emission scan was taken of the same sample using a steady-state fluorometer (PTI), exciting at 490 nm and recording at 510-700 nm. Samples were diluted and recordings repeated 5 times, to populate absorbance in the 0 to 0.1 region, where absorbance and emission are linearly related. The same procedure was applied to the reference dye, rhodamine 6G, dissolved in

ethanol. The quantum yield of the donor dye was then calculated according to equation:

$$Q_D = \frac{Q_R E_D A_R n_D^2}{A_D E_R n_R^2} \tag{4.1}$$

where Q is quantum yield, E is integrated emission across the whole spectrum, A is absorption at 490 nm, n is the refractive index of the medium, and D and R refer to the donor and the reference dyes, respectively. Established values were taken for the quantum yield of rhodamine 6G in ethanol (0.95; reference [143]), and for the refractive indices of water and ethanol (1.333 and 1.361, respectively).

To calculate the overlap integrals [60], absorption spectra of the following acceptor samples were also measured: Atto647 free dye, Atto647 attached to DNA (in the presence and absence of Pol in a 1:1 ratio), and Atto647N attached to Pol. Samples were diluted in Pol buffer to 2 μM concentration, and absorption recorded at 400-710 nm. Both the absorption spectra of the acceptor, and the emission spectra of the donor (see above) were corrected for background, normalized, and the overlap integral calculated as in equation 3.3. The extinction coefficient of Atto647N at A_{max} was taken as 150,000 $M^{-1}cm^2$ [144]. This allowed isotropic R_0 values to be calculated, using equation 3.2 and assuming orientational averaging ($\kappa^2 = 2/3$) and the refractive index of water (n=1.333).

In order to test if rotational averaging is justified, steady-state anisotropies were measured [52]. Samples were diluted to 100 nM in Pol buffer and excited with vertically polarized light at 532 nm (donor) or 639 nm (acceptor samples) in a steady-state fluorometer (PTI). Fluorescence was measured through horizontally and perpendicularly oriented emission filters at 570 nm (donor) or 669 nm (acceptor samples), over 1 minute. Anisotropy was calculated from the difference of vertically and horizontally polarized emission intensities, corrected for background and for the different sensitivities of the emission channel for vertically and horizontally polarized light. Finally, measured anisotropies were used to calculate anisotropic R_0 values, using a Monte Carlo simulation-based method implemented in the nano-positioning system software [118].

4.8.5 Structure preparation and AV modelling

The polymerase structure was obtained from the *Bst* X-ray crystal structure (PDB code 1L3U, reference [8]). DNA was removed and Cys substitutions were introduced at positions K498, V692 and A855 (corresponding to *E. coli* residues K550, L744 and C907), using PyMol (Schrödinger, LLC). B-DNA models of the upstream and downstream DNA were made using 3D-DART modelling server [145], and were truncated at the gap-proximal ends by 3 base-pairs each for the purposes of rigid-body docking. The radii, linker lengths and linker widths of Cy3b and Atto647N dyes were estimated from their structures *in silico* using ChemDraw (Perkin Elmer). The parameters used are summarized in Table 4.2.

Dye	Linker length	Linker width	Radius 1	Radius 2	Radius 3
Cy3b (DNA)	14.2	4.5	8.2	3.3	2.2
Cy3b (protein)	9.1	4.5	7.7	2.5	1.3
Atto647N (DNA)	17.8	4.5	7.4	4.8	2.6

Table 4.2: *Dye dimensions. The attachment chemistry of dyes (DNA vs. protein) is indicated. All dimensions are given in Å.*

We used the accessible-volume algorithm of the FPS software [120] to model the mean positions of the dyes for each Pol and DNA attachment site. The attachment points were taken to be the Sβ atoms of Cys residues, and the C7 atoms of thymine residues.

4.8.6 Rigid-body docking

Three-body rigid-body docking with Pol, upstream and downstream DNA structures was performed in the FPS software [120], using the calculated R_{mp} distances. Docking was repeated 1000 times from different starting configurations of the binding partners, using a clash tolerance of 6 Å. This generated several clusters of structures, which were distinguished by the different RMSD values relative to each other, and the different goodness of fit to experimental data (χ_r^2). Structures with χ_r^2 values above 6 were rejected, and one

structure from each of the remaining clusters was further refined using a clash tolerance of 2 Å, and then again using a tolerance of 1 Å, during which steps the AV clouds were recalculated. The structure with the lowest χ_r^2 was taken, and the R_{mp} distances from the model back-converted to $\langle R_{DA} \rangle_E$ and FRET efficiency values, to compare with the experimental FRET data. For precision estimation, 100 bootstrapped structures were generated from the best model, using a clash tolerance of 1 Å. The coordinates of each phosphorous atom were extracted using PDB editor [146], and its RMSD calculated across the 100 bootstrapped structures.

To compare the position of the upstream DNA in the docked structure with the crystal structure, the protein components of the FRET-restrained and crystal structures were aligned in PyMol. The RMSD of the upstream DNA fragment between the two structures was calculated as for the bootstrapped structures, but across all phosphorous atoms.

The DNA structure in complex with the Pol dimer was obtained using the same procedure as the Pol-DNA structure, but with no polymerase present in the docking. In the case of the DNA structure in the absence of Pol, the DNA fragments were at their full length, and with an additional distance restraint of 5 +/- 2.5 Å imposed between the C_α atoms in the template strand opposite the gap, to account for the covalent link between the two.

4.8.7 DNA substrate simulations

The OxDNA model used was an averaged model with sequence-independent parameters for the hydrogen-bonding and stacking interactions [126]. 100 simulations of 10^8 steps each were performed for each of the gapped, nicked and duplex DNAs, with interaction energies and configurations sampled every 10^3 steps. The time step was 0.005 simulation units, where one simulation unit implies a time of 3.03×10^{-12} s. The temperature was set to 295 K, and an Andersen-like thermostat was used. Particle velocities were refreshed every 10^3 steps from the Maxwell distribution corresponding to the simulation temperature, with fixed probabilities of 0.02 and 0.0067 for the linear and angular velocities, respectively.

The bend angle was calculated from the vectors placed along the midlines of the two

helix segments, by adapting a previously described approach [147]. The relative free energies were calculated from the MD trajectories, as follows:

$$A(|\theta|)/k_B T = -\log\left(\frac{p(|\theta|)}{p(|\theta_0|)}\right) \tag{4.2}$$

where $A(|\theta|)$ is the free energy, k_B is the Boltzmann constant, $p(|\theta|)$ is the observed probability density for the DNA adopting a bend angle $|\theta|$, and $|\theta_0|$ is the reference bend angle, for which $A(|\theta_0|) = 0$.

4.9 Contributions

- Pol variants were previously purified and labelled by members of Cathy Joyce's group.
- Labelling of DNA constructs, collection of single-molecule confocal data, accurate FRET corrections, construction of the R_{DA}-to-R_{mp} polynomial equation and *in silico* measurements of dye dimensions were carried out by Tim Craggs, who also prepared Figures 4.3 and 4.14.
- Coarse-grained DNA simulations were carried out and data analysed by Majid Mosayebi, Hendrik Kaju and Jonathan Doye, who along with Tim Craggs also provided elements for Figures 4.12 and 4.13.

5

Molecular dynamics simulations of Pol-DNA complex

5.1 Introduction

5.1.1 Project rationale

In the previous chapter, we used single-molecule FRET and rigid-body docking approaches to produce a structure of Pol bound to its gapped-DNA substrate. Despite the significant insight that the structure provides for the understanding of the Pol binding mechanism, structure determination via rigid-body docking faces a number of limitations. Most importantly, biomolecules are treated as *rigid bodies*, which they are generally not [72], and any dynamics or binding-induced conformational changes are ignored. Component structures may also need to be broken down into fragments to allow rigid-body docking, resulting in missing information in the docked structure. In our case, the conformational change in the DNA is accounted for by breaking it into the upstream and downstream components, which not only results in the loss of intermediate DNA sequence, but also relies on the assumption that the bending involves a single 'bend point', which may not be accurate. Finally, the resolution of docked structures does not normally allow molecular details, such as inter-residue interactions, to be inferred with certainty.

Due to these limitations, the FRET-restrained structure of the Pol-DNA complex leaves the following questions unanswered: (i) How dynamic is the complex? (ii) Where is the template strand of downstream DNA channelled? (iii) What is the position of the non-template strand of downstream DNA, and how many base-pairs are unwound? (iv) What are the molecular interactions between Pol and downstream DNA? In order to address these questions and gain further insight into the mechanisms of DNA binding and repair by Pol, we used our docked structure to generate a model of the Pol-DNA complex with the full DNA sequence, and subjected it to all-atom molecular dynamics simulations. Since these simulations rely heavily on accurate modelling of the DNA substrate, we provide a brief technical overview of atomistic DNA simulations in the following introductory section, highlighting some of the challenges involved. We also review the field of simulations of DNA-protein complexes, particularly in terms of the range of biological problems that have been addressed to date, both generally and in the case of DNA polymerases.

5.1.2 All-atom DNA simulations

Molecular dynamics simulations of nucleic acids have traditionally lagged behind protein simulations, and a number of technical developments have been required to make them routine. The key step was the development of second-generation AMBER and CHARMM force fields, which showed good agreement with X-ray and NMR structures in short simulations [148]. With increasing size and complexity of the simulations, the original force fields had to be frequently updated to ensure stability of the simulated structures and agreement with experimental data. In particular, both force fields have been updated for their bonded parameters, to account for the complexity of the backbone torsional motions in DNA (which involves six torsion angles, compared to two in peptide bonds) [149], resulting in force-field variants AMBER ff99bsc0 [150] and CHARMM36 [151]. AMBER force field has been more widely used of the two, for instance by the 'ABC consortium', which has produced simulations on the order of 100 μs in total [152]. However, CHARMM could be preferred for simulations of single-stranded DNA, as AMBER parameters have been found

to bias ssDNA towards a helical conformation observed in duplexes [153].

DNA simulations in explicit solvent can be computationally expensive, due to the non-globular shape of DNA that requires large numbers of particles for full solvation. Implicit solvent models have been used as an alternative, with promising results. Simulations of short DNA duplexes have reproduced many of the results of explicit-solvent simulations, and studies of longer duplexes not amenable to explicit systems are in good agreement with experimental data [154]. In addition, since DNA is a highly charged molecule, a more careful consideration of electrostatics is required than with small globular proteins. Early simulations used group-based truncation methods for long-range electrostatic interactions, which were later shown unable to produce accurate and stable trajectories [155]. Atom-based truncation approaches with a force-shifting function and sophisticated smoothing of electrostatic forces have produced better results, but the real breakthrough arrived with the development of the particle-mesh Ewald method. The PME approach has produced stable conformations of canonical DNA with good computational efficiency [156], and has allowed high parallelization of MD simulations.

The high charge density of DNA also attracts a number of counterions in solution, sometimes referred to as the *ion atmosphere*, which need to be included in the simulation and correctly parametrized. Artificial salt aggregates have been observed with default ion parameters [157], and have prompted the development of new parameters for both the AMBER and CHARMM force fields [158, 159]. Of particular importance for the structure and function of DNA are divalent cations, which in addition to forming the ion atmosphere also participate in catalysis, such as in DNA-processing enzymes. Unfortunately, developing a set of parameters that would accurately describe both roles of divalent cations has been challenging [149]. The solvation energetics and kinetics of divalent cations are also subject to strong polarization effects, which are difficult to reproduce with standard force fields [160, 161], although this has been achieved to a limited extent using parameter optimization [162].

5.1.3 DNA-protein complex simulations

Simulations of DNA-protein complexes are further complicated by the fact that the chosen force field needs to account for the properties of both DNA and protein in a balanced way. AMBER ff99bsc0 and CHARMM36 are most commonly used, and the simulation conditions are adjusted to suit both biomolecules. Due to the importance of atomic detail and solvent effects in protein-DNA interactions, all-atom explicit solvent models are often the best choice for DNA-protein simulations. However, implicit-solvent simulations of DNA-protein complexes are possible, and generalized Born-based approaches have yielded stable trajectories that agree with experimental data and explicit-solvent simulations [163]. Similarly, a multiscale approach has been presented that combines a coarse-grained model of DNA with an atomistic model of the protein, and its feasibility demonstrated in simulations of the *lac* repressor binding its 76-base-pair long looped DNA substrate [164].

MD simulations have often been used to investigate protein-induced DNA distortions and bending. Indeed, the first DNA-protein simulation observed DNA binding and bending by glucocorticoid receptor DNA-binding domain [165]. In another example, binding of transcription factor p53 to its consensus sequence was found to induce bending of two different DNA sequences by 20° and 35°, in agreement with experimental data [166]. MD simulations, along with Monte Carlo approaches and free-energy calculations, have also been instrumental in explaining the sequence specificity of DNA-binding proteins. DNA recognition by restriction endonuclease BamHI was investigated by comparing the structure and energetics of water at the protein-DNA interface of the specific and non-specific complexes, and the water distribution found to serve as a fingerprint that mediates specificity [167]. Similarly, molecular dynamics simulations combined with the molecular mechanics / generalized Born surface area approach (MM/GBSA) were used to investigate sequence-specific and non-specific DNA binding of lac repressor [168]. This approach has also been used in studying cooperativity, for example to explain why DNA binding by transcription factor RUNT is enhanced by a second protein, CBFβ [169].

MD simulation studies of polymerases have focused largely on X-family polymerases,

such as DNA polymerase β and other repair polymerases. Two studies have investigated the conformational states of pol β and polymerase X, and simulated the open-to-closed transition upon DNA and nucleotide binding, confirming predictions from X-ray structures [170, 171]. Simulations of DNA polymerase μ in complex with different mispaired nucleotides identified several 'gate-keeper' residues that distort the active site when non-cognate nucleotides are bound, and help to discriminate against them [172]. Similarly, simulations of pol μ variants helped to identify the residues involved in stabilizing the template nucleotide, and suggested a mechanism for how frameshift errors are prevented [173]. Free-energy calculations were also carried out to compare the specificity of nucleotide binding in pol β and T7 DNA polymerase, and it was found that the latter discriminated more effectively against mis-matches [174, 175]. Finally, two studies have investigated the details of the active site architecture and the catalytic mechanism of pol β, including the properties of the catalytic cations and the protonation states of the coordinating ligands [176, 177].

The *Bst* homologue of DNA polymerase I has also been studied, but only in terms of the fingers-opening transition. In one example, targeted MD simulations were used to investigate the mechanism of DNA translocation following nucleotide insertion [35], a poorly understood step of Pol reaction cycle. It was found that pyrophosphate release from Pol-DNA-PPi complex facilitates the opening of the fingers subdomain, which involves the O-helix bending at two conserved glycine residues. Fingers-opening also affects the positioning of the conserved Y714 to stack against the terminal base pair, thus coupling the conformational changes to DNA translocation. In the other example, Pol-DNA complexes in the closed and ajar states were simulated using unbiased simulations on a microsecond time scale, and Pol shown to transition to the open state when the nucleotide was removed from the structure [36]. The opening mechanism was described in terms of the interactions and dihedral switches involved, and a previously unknown intermediate structure was observed. Notably, no molecular dynamics study has focused on the dynamics of DNA in Pol-DNA complexes, or on Pol-DNA interactions beyond those observed in the active site.

5.2 Control simulations

We chose to perform the simulations with the Amber ff99sb force field and parmbsc0 nucleic acid parameters, which are the state-of-the-art parameters for protein-nucleic acid complex simulations. Conditions were kept as close to the experimental ones as possible, including the temperaturre of 298 K and the presence of 10 mM $MgCl_2$. In order to test the suitability of the force field and the simulation conditions for studying Pol dynamics, we performed control simulations on crystal structures of Pol. We used a *Bst* structure (PDB code 4BDP) that is almost identical to the structure used for rigid-body docking (PDB code 1L3U), for reasons that we describe in the next section. First, the protein component of the crystal structure was extracted, and Pol simulated in the apo state for 100 ns. The RMSD of C_α protein atoms relative to themselves stayed at a low and constant value during the simulation, indicating that the structure was stable under our conditions (Figure 5.1). Second, the crystal structure was simulated for 100 ns with the short fragment of upstream DNA present. The protein structure was again stable, and the DNA remained stably bound to the polymerase, as indicated by the RMSD of DNA atoms relative to C_α protein atoms. This result suggested that the chosen simulation conditions were also optimal for sustaining Pol-DNA interactions observed in the crystal structure, and allowed us to proceed with the full complex simulations.

5.3 Model preparation

We constructed a starting model of the Pol-DNA complex for molecular dynamics simulations by combining the DNA components present in the FRET-restrained and X-ray structures, using basic molecular sculpting (see Section 5.9 for details). Whilst the docked structure contained most of the upstream and downstream DNA, the X-ray structure provided information on the position of DNA at the active site, hence allowing the gap between the upstream and downstream DNA to be built in. We chose a *Bst* structure (PDB code 4BDP) that is virtually identical to the structure used for rigid-body docking (PDB code 1L3U)

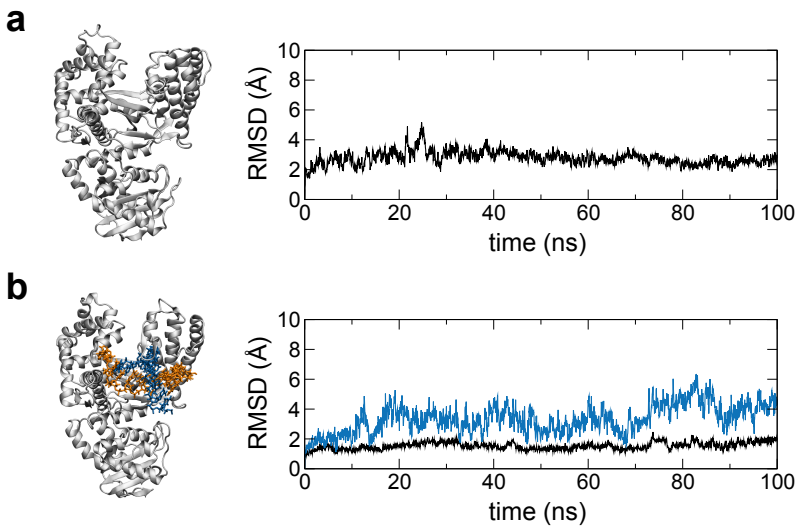


Figure 5.1: *Control simulations of Bst crystal structure (PDB code 4BDP), in the apo state (top) and with the X-ray DNA present (bottom). The snapshots indicate representative conformations of the protein in the two simulations. The plots show the RMSD of C_α protein atoms relative to C_α atoms (black), and the RMSD of DNA atoms relative to C_α atoms (blue).*

in terms of the polymerase component, but contains electron density for an additional nucleotide downstream (+1), which aided in the generation of the starting model.

5.4 High-temperature simulations

The resulting model represents an 'educated guess' of the structure of the Pol-DNA complex; however, it may not correspond to its energetically most stable state. In order to allow the DNA component of the model to explore the range of conformations accessible to it, and hence to probe its energy landscape, we subjected it to high-temperature (400 K) simulations. We performed five simulations of 2 ns each, with all atoms except for the 6 base-pairs of the protein-proximal portion of downstream DNA position restrained. A variety of conformations were observed, from which we selected a small number of extreme (and different) conformations, in order to test the whole range of possible starting positions for the simulations. We added the polymerase structure back to the DNA, discarded any conformations that showed a steric clash, and selected five of the resulting Pol-DNA structures as starting models for complex simulations (Figure 5.2).

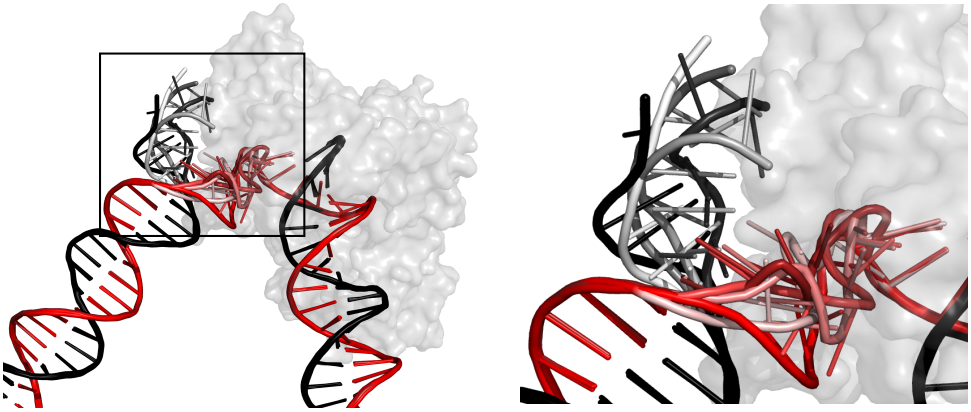


Figure 5.2: *Overlay of the five conformations of DNA that emerged from high-temperature simulations and were in turn used as starting models for complex production runs. The region that was allowed to move and that differs between the models (shown zoomed-in on the right) is represented with different intensities of red and black (highest, model 1; lowest, model 5). The protein is shown transparent for clarity.*

5.5 Complex simulations

We performed two 100-ns all-atom simulations on each of the five starting models, which produced a total of ten trajectories and 1 μs of simulation time. In all of the simulations, the polymerase and the DNA components remained stable and bound to each other, and showed no obvious artefactual behaviour. Although variation was observed between the simulations, in terms of the detailed dynamics and interactions, several patterns also emerged, as we discuss below. The trajectories obtained from the same starting model appeared as different from each other as those obtained from different models, suggesting that (i) the starting model does not 'lock' the complex in any particular conformation, and (ii) conformational sampling is limited on the time scale of our simulations.

5.5.1 Structure dynamics

In all of the simulations, the fragment of upstream DNA present in the X-ray structure remained stably bound to Pol. For this part of DNA, we calculated an RMSD of 2.8 +/- 0.8 Å relative to the polymerase, which indicates only minor side-chain rearrangements. In contrast, we observed the upstream and downstream DNA to be very mobile on the whole,

giving RMSD values of 11.1 +/- 4.4 Å and 19.2 +/- 8.8 Å, respectively. Figure 5.3 shows the variation in RMSD over time for one example simulation that displayed representative dynamics.

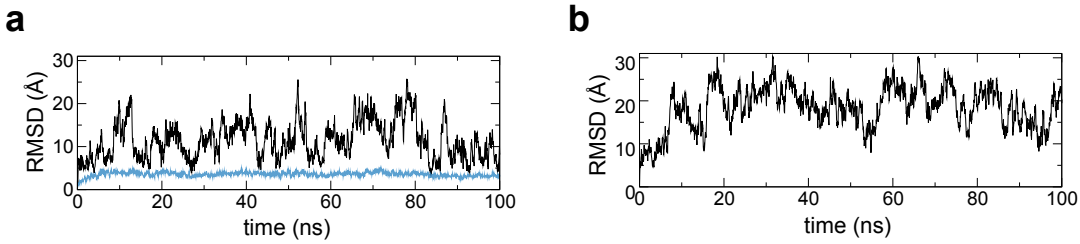


Figure 5.3: *Dynamics of DNA in one example simulation.* ***(a)*** *RMSD of upstream DNA atoms (black), and of the part of upstream DNA present in the X-ray structure (blue), both relative to C_α protein atoms.* ***(b)*** *RMSD of downstream DNA atoms, relative to C_α atoms.*

We further characterized DNA dynamics by measuring the distance between the terminal non-hydrogen atoms of the template strand, referred to as the *end-to-end distance* (Figure 5.4). We found the distance to vary between 24 and 144 Å across all simulations, indicating a range of DNA conformations and bend angles being adopted. In one outlier simulation, the upstream and downstream DNA were seen to bind to each other via their phosphate groups and a bridging magnesium ion present in solution. Such interactions can be physiological, but their relevance for the Pol-DNA complex is unclear, both because they were not observed in any of the other trajectories and because of the known difficulties with divalent-cation parameterization.

5.5.2 Non-template flap

The 6-nucleotide gap-proximal end of the non-template strand of downstream DNA, from here onwards referred to as the *non-template flap*, showed appreciable dynamics, with an RMSD of 8.7 +/- 3.7 Å across all simulations (Figure 5.5a, b). The flap did not dock to any one stable position in our simulations, but instead formed transient interactions with a variety of Pol residues. The most consistent of these involved residues R729 and K730,

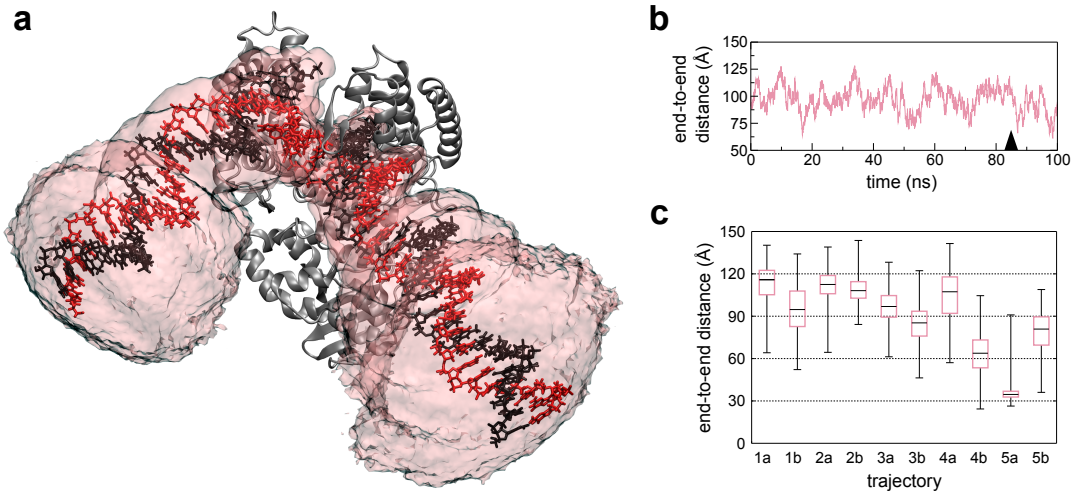


Figure 5.4: *(**a**) Representative snapshot of Pol-DNA complex during a 100-ns trajectory. The volume accessible to DNA during the time course of the simulation is shown in transparent pink. (**b**) End-to-end distance as a function of time during the same simulation. The time point corresponding to the snapshot in (a) is indicated with an arrowhead. (**c**) Range of end-to-end distances accessible to the complex in each of the ten trajectories, with the numbers (1-5) referring to the different starting models, and the letters (a and b) indicating repeats. Black lines are median values, boxes denote 25th and 75th percentiles, and whiskers show the minimum and maximum values.*

which contacted the phosphate groups in the DNA, and less frequently also its sugar and base moieties. The flap was also occasionally seen to fold onto itself, forming hydrogen-bond interactions with the rest of the downstream DNA.

Most importantly, whereas all simulations were started from a model in which 6 base-pairs of the downstream DNA were unpaired, only some of the base-pairs remained unpaired in the simulations (Figure 5.5c). For more than 40 % of the simulation time, the non-template flap formed 5 hydrogen bonds with the template strand (Figure 5.5d), corresponding to an A-T and a G-C base-pair, with the remaining 4 nucleotides unpaired. During ~20 % of the time, the flap formed 2 base-pairs with the template strand, equivalent to an A-T base-pair, with 5 nucleotides unpaired. Occasionally, transient formation and breakage of H-bonds were observed; however, the number of unpaired nucleotides almost always remained between 3 and 5.

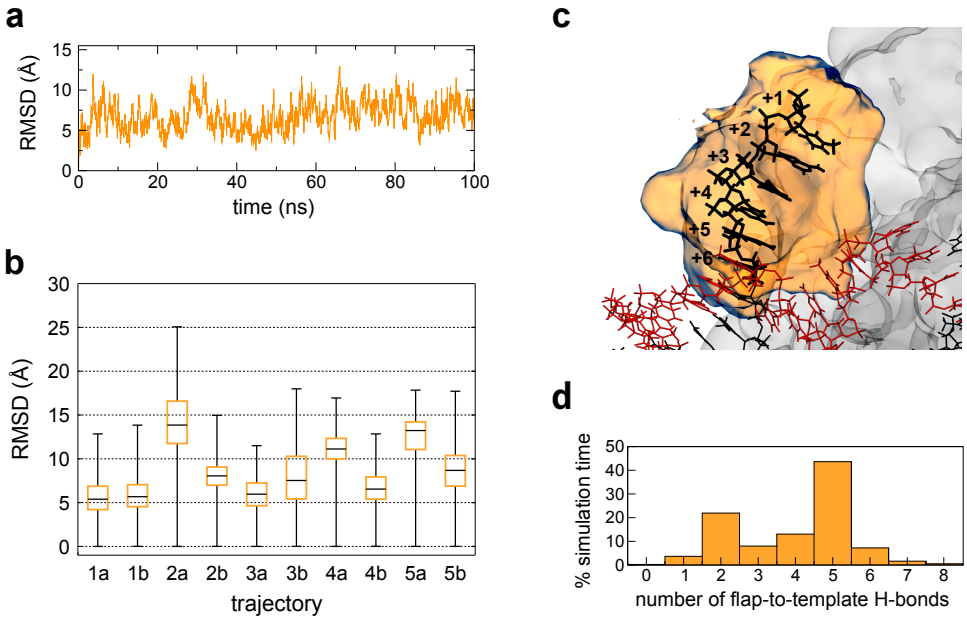


Figure 5.5: *Non-template flap.* ***(a)*** *RMSD of the flap, relative to the rest of downstream DNA, during an example simulation.* ***(b)*** *Range of RMSD values for the flap, relative to the rest of downstream DNA, in each of the ten trajectories. The representation is as in Figure 5.4.* ***(c)*** *Representative snapshot of the conformation of the non-template flap, with its volumetric map during a 100-ns simulation (transparent orange).* ***(d)*** *Frequency of the different number of hydrogen bonds that are formed between the flap and the template strand of downstream DNA, during the entire 1-µs simulation time.*

5.5.3 Strand separation by Y719

Protein-DNA interactions in the active site were similar to what is observed in the X-ray structure [26] and hence the starting model, with Y714, S717, Y719 and R789 contacting the template strand. Here, we focus on residue Y719 (Figure 5.6), equivalent to F771 in *E. coli*, which is contained in the conserved three-helix bundle and is known to be important for the strand-displacement activity of Pol. We observed Y719 to be consistently positioned between the template and non-template strands, as shown in a representative snapshot (Figure 5.7a) and in volumetric maps (Figure 5.7b). However, the exact position and orientation of Y719 varied. During 73 % of the simulation time, Y719 was positioned at residues +1 or +2 of the template strand, and occasionally stacking against them; for the remainder of the time, Y719 was found closer to residues 0 or +3 (Figure 5.7c). As

discussed above, the corresponding residues in the non-template strand (residues +1 to +3) are seen to be unpaired in all of the simulations, and hence the positioning of Y719 appears to provide a wedge to separate the non-template strand from its template counterpart. Whilst we do not see Y719 acting directly on nucleotides +4 or +5, their unpairing is likely a direct consequence of the position of the Y719 wedge farther upstream. Despite the intrinsic dynamics of the non-template strand, therefore, the stable positioning of Y719 at the template strand appears to prevent any re-pairing from taking place during catalysis.

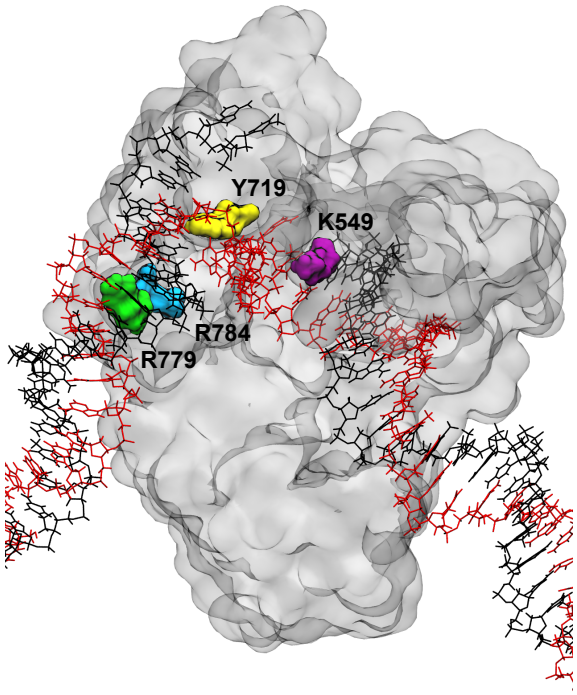


Figure 5.6: *Overview of important Pol residues, involved in strand separation and interactions with downstream DNA, and discussed in Sections 5.5.3 and 5.5.4.*

5.5.4 Interactions with downstream DNA

We observed novel interactions between downstream DNA and Pol, which consistently involved positively charged residues on the Pol surface, and negatively charged phosphate groups of the DNA backbone. These interactions occurred in two regions: the first involved the base of the fingers domain, with residues R779 (S831 in *E. coli*) and R784 (R836

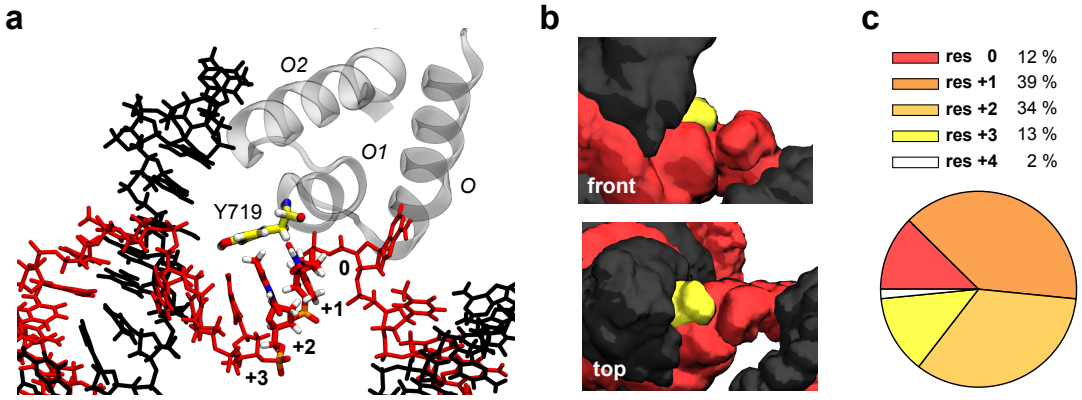
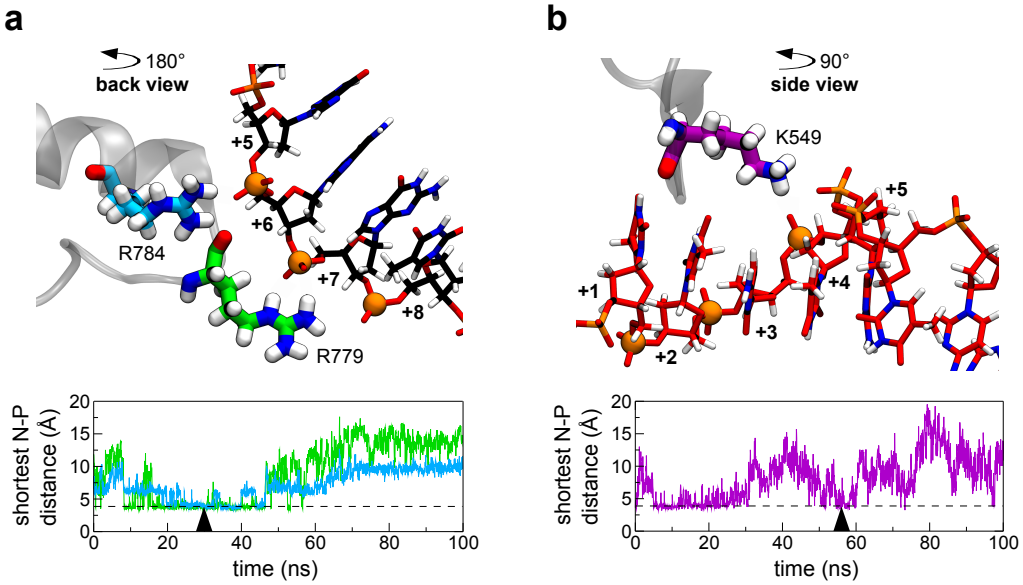


Figure 5.7: *Strand separation by Y719.* ***(a)*** *Representative snapshot of the position of Y719 relative to template strand of downstream DNA. The two DNA residues that are positioned closest to Y719 during the time course of the simulation (+1 and +2) are highlighted in CPK colouring. The position of the three-helix bundle is shown for reference; the rest of the protein is omitted for clarity.* ***(b)*** *Two different views of volumetric maps of Y719 (yellow), template (red) and non-template DNA strands (black) during a 100-ns simulation.* ***(c)*** *Frequency of each of the bases of downstream residues of the template strand (0 to +4) being the closest residue to the side chain of Pol residue Y719, during the entire 1-μs simulation time.*

in *E. coli*) contacting the duplex region of downstream DNA (Figures 5.6 and 5.8a). The interactions involved both the side-chain and the backbone N-H groups of these residues, forming hydrogen bonds with the oxygen atoms of the phosphates. Whilst both R779 and R784 were found to interact with the non-template strand, R779 also additionally made contacts with the template strand. Normally, 1-3 phosphate groups of the DNA were involved, although in some simulations up to 6 groups were seen to interact. The identity of the interacting phosphate groups varied with time and between the trajectories, but generally located to the same region of the DNA, which was residues +5 to +9 of the non-template strand and +11 to +15 of the template strand (Figure 5.8c). Whilst any individual contact was transient, the large number of possible contacts resulted in binding being observed for up to 50 % of the simulation time.

The second interacting region located to the H1-H2 loop in the thumb domain of Pol, which is also involved in contacting upstream DNA. The key residue in this region was K549 (K601 in *E. coli*), which interacted with the unpaired template strand, again via both



c

		trajectory									
		1a	1b	2a	2b	3a	3b	4a	4b	5a	5b
Pol residue	**DNA strand**	**phosphate number**									
K549	template	3-4	1-3	2-4	2-4	2-4	2	2-4	1-3	1-2	3, 5
R779	template	12-15	12-13	13-18	12-13	11-14	12-15	12-15	11	11-12	13-14
R779	non-template	7-8	7-8	7-9	7-8	6-8	6	/	6-7	5-7	/
R784	non-template	/	6-7	/	6-7	6	6	/	5-6	6	/

Figure 5.8: *Interactions with downstream DNA.* ***(a)*** *Representative snapshot of the interactions of R779 (green) and R784 (cyan) with phosphate groups of the non-template strand of downstream DNA. The phosphorous atoms involved in interactions in this simulation are shown as orange spheres. The view is rotated clockwise by 180° relative to the view in Figure 5.6. The plot shows the minimum distance between the side-chain nitrogen atoms of R779 (green) or R784 (cyan) to any phosphorous atom of the non-template strand of downstream DNA, during an example simulation. The distance corresponding to an interaction is shown with a dashed line; the time point corresponding to the snapshot is indicated with an arrowhead.* ***(b)*** *Representative snapshot of the interaction of residue K549 (purple) with phosphate groups of the template strand of downstream DNA. The view is rotated clockwise by 90° relative to the view in Figure 5.6. The representation is as in (a).* ***(c)*** *Detailed list of interactions observed between specific Pol residues and phosphate groups of the template and non-template strands of downstream DNA, for each of the ten trajectories. The 'phosphate number' indicates DNA residue numbers (with the '+' sign omitted) whose phosphate groups are interacting.*

its side-chain and backbone N-H groups (Figures 5.6 and 5.8b, c). These interactions occurred close to the position of the strand-separating Y719, and normally involved the same range of DNA residues that were held unpaired by Y719. K549 occasionally interacted with the base moieties of the template strand as well, and even with the phosphate groups of the non-template strand, although the significance of these contacts is unclear as they were not observed consistently across all the trajectories. Interestingly, despite their close proximity, the neighbouring residues K548 and K551 were not seen to interact with downstream DNA in our simulations.

5.6 DNA substrate simulations

Finally, to test the validity of the coarse-grained simulations presented in Chapter 4, we performed all-atom simulations on the gapped-DNA substrate alone, as well as on a duplex DNA bearing the same sequence. Due to the large simulation box required to accommodate the flexibility of the DNA, and hence the high computational cost, we limited the simulations to 20 ns each. The duplex and gapped DNA showed similar dynamics in terms of the RMSD of the downstream DNA relative to the upstream DNA, with average values of 26.7 Å +/- 8.8 Å and 22.4 Å +/- 8.4 Å, respectively. The average end-to-end distances were also found to be similar, with 173.4 Å +/- 5.1 Å for the duplex DNA and 168.8 Å +/- 9.4 Å for the gapped DNA. However, the gapped DNA could access much lower end-to-end distances than the duplex DNA (128.1 Å versus 158.4 Å), due to a significantly bent state adopted on one occasion (Figure 5.9). Notably, no fraying was seen at the ends of either the duplex or gapped DNA, except for the occasional breakage and formation of H-bonds that occurred throughout the length of the DNAs.

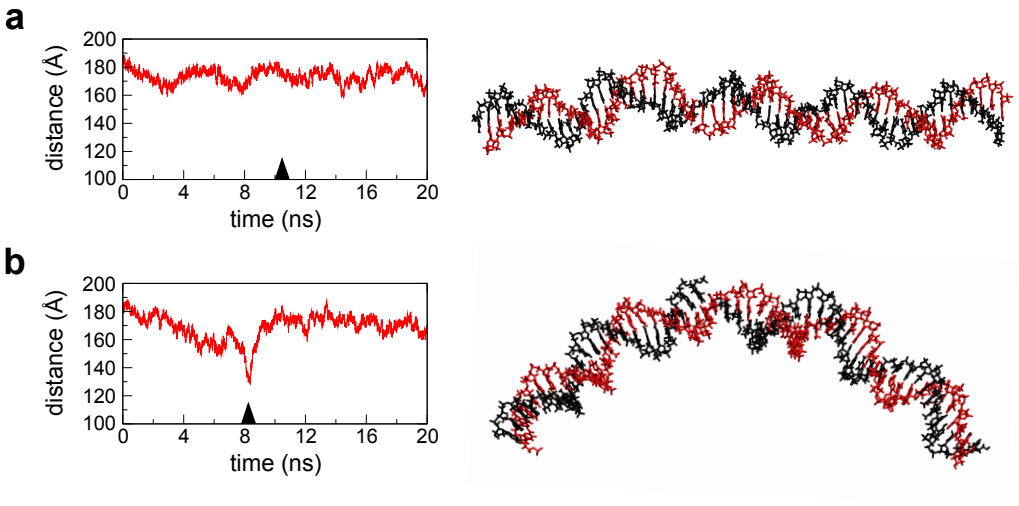


Figure 5.9: *All-atom simulations of* ***(a)*** *duplex DNA and* ***(b)*** *gapped DNA. The plots show the end-to-end distance as a function of time during the simulation. The snapshots represent an example conformation of the duplex DNA, and the highly bent conformation of the gapped DNA. The time points corresponding to the snapshots are indicated with arrowheads.*

5.7 Discussion

Our simulation results added valuable detail to the information provided by the FRET-restrained structure. Whilst the stability of the DNA fragment present in the crystal structures was expected, the high dynamics of the DNA ends on a nanosecond time scale were more surprising, particularly considering the static picture of the Pol-DNA complex that emerged from rigid-body docking. As noted previously (Section 4.6.3), any sub-millisecond dynamics are averaged out in confocal smFRET experiments, and hence the 'static' docked structure in fact represents an average of many inter-converting conformations. The high complex dynamics have implications for the binding mechanism of Pol, investigated in the previous chapter, as they suggest that the DNA substrate can access a range of bend angles even upon complex formation with Pol. Hence, rather than selecting for (or inducing) a specific conformation of the DNA, the polymerase may be able to recognize and bind the DNA in a variety of conformations, as long as the bend angle is sufficient to avoid any clashing between the downstream DNA and Pol.

The unpairing of the non-template flap agrees with the predictions from rigid-body

docking, which indicated that at least two base-pairs of downstream DNA have to be unpaired (Section 4.2.5). Further, recent quenchable FRET experiments using a DNA substrate with a donor-labelled non-template strand and an acceptor-labelled template strand show that contact quenching occurs in the absence of Pol, but is abolished when Pol is added (T. Craggs et al., in preparation). This indicates that in the DNA substrate alone, the flap is base-paired to its template counterpart, but it becomes unpaired when bound to the polymerase. The quenching was observed when the two dyes were at positions +1 and +4, but not when they were at positions +8 and +9, indicating that the unwinding is limited to a maximum of 7 residues of the flap. Notably, the flap is dynamic in our simulations and does not dock to any stable position, nor does it appear to be channelled along the Pol surface by specific Pol-DNA contacts, as was previously hypothesised [10]. Both the unpairing and the dynamics of the flap are likely important features of the strand-displacement mechanism of Pol, allowing the synthesis of new DNA to proceed unhindered, and exposing the old DNA or RNA strand for excision by the 5' exonuclease domain of Pol.

Our simulations directly confirm the suggested involvement of the three-helix bundle, and particularly of Y719, in strand separation by Pol. The position of Y719 close to residues +1 and +2 of downstream DNA for the majority of simulation time is also consistent with several photo-crosslinking experiments that show residues +1 and +2 to be involved [9, 32]. The occasional stacking interactions between the aromatic side chain of Y719 and the template bases observed in the simulations have also previously been suggested based on the cross-linking data [9]. Notably, previous models for the mechanism of strand displacement have hypothesized a process akin to unzipping, with F771 (Y719) acting as a wedge to separate the strands [10]. The stable position of Y719 at the template strand in our simulations, and the high dynamics of the non-template flap, suggest a more passive mechanism for Y719, perhaps ensuring to maintain rather than induce strand separation. However, an active involvement of Y719 may be required during the initial steps of DNA binding, as the DNA substrate is seen to be largely base-paired on its own (Section 4.5).

The interactions between Pol and downstream DNA that we observe in the simulations are likely to be physiologically relevant. Residues R779 and R784 have a homologous

residue in 29 and 48 out of 50 bacterial polymerases analysed, respectively [9]. The proximity of R779 to R784 in the structure, and the similar interactions that they form with downstream DNA in our simulations, suggest that the two residues may be functionally complementary. Whereas our simulations indicate that R779 is more important for contacting DNA in *Bst*, R784 may be the key residue in some of the other bacterial polymerases that lack a positively charged residue at position 779, such as in the *E. coli* homologue. Substitution of the R784-equivalent residue R836 has been shown to increase the binding of downstream DNA to the polymerase site in *E. coli* [9, 178], presumably due to its contribution to the bending and distortion of the downstream DNA. Maintaining the bent conformation of the DNA in the complex could be important for optimizing the chemistry of nucleotide incorporation or for nucleotide discrimination [178], as has been suggested for pol β (Section 4.6.1).

Similarly, residue K549 is part of a (K)KT motif that is found in 33 out of 50 bacterial polymerases analysed [9]. In our simulations, interactions with K549 appear to keep the template strand away from its non-template counterpart, which could be a mechanism that facilitates the process of strand separation. Radioactive competition assays and cross-linking experiments have shown that Pol forms contacts with the first 4 nucleotides of the template strand of downstream DNA [9], which goes beyond the reach of the active-site residues (Y714, S717, Y719 and R789), but could be accounted for by interactions with K549. Whilst the authors of the study were not able to identify the amino acid(s) cross-linking to residue +4, this effect could be due to the dynamics of the template strand and the transiency of interactions with K549 apparent in our simulations. In general, whether they are involved in DNA binding or in strand separation, the dynamic nature of all of the observed Pol-DNA interactions may help the polymerase to move swiftly along its DNA substrate during strand-displacement synthesis.

Finally, our DNA-only simulations suggest that whilst the dynamics of the duplex and gapped DNA are similar, the gapped DNA can occasionally access higher bend angles. This result agrees with the conclusions from the coarse-grained simulations (Section 4.6.3), and explains how Pol can distinguish between a gapped site and undamaged DNA.

5.8 Conclusions and future work

In this chapter, we have used all-atom molecular dynamics simulations to enhance our FRET-restrained structure of the Pol-DNA complex with dynamic and residue-specific information. In this way, we have provided further insight into the mechanisms of DNA recognition, binding and strand displacement by Pol. Specifically, the observed high dynamics of the complex could not be predicted from the docking approaches alone, and may have significant implications for the mechanism of substrate recognition by Pol. The consistent unpairing of the non-template flap agrees with the single-molecule fluorescence data, but also provides further quantitative information as to the degree of unwinding. Similarly, whilst confirming a number of mutagenesis and biochemical studies, our simulations provide a direct observation of strand displacement by Y719, and suggest some mechanistic features not predicted by previous models. In addition, we demonstrate the presence of interactions between specific Pol residues and downstream DNA, which may be important for nucleotide incorporation and strand separation. Finally, the DNA-only simulations confirm the relevance of the Pol recognition and binding mechanisms (discussed in Chapter 4) at the atomistic level.

Notably, due to the relatively short time scale accessible to atomistic systems, it is unlikely that our simulations are able to probe the entirety of the conformational space available to the complex. As a result, it is difficult to establish with certainty the relative importance of the behaviours observed in our simulations, since the underlying conformations and molecular interactions are not energetically equilibrated. Nevertheless, in this report we have focused on the patterns of behaviour that were observed consistently across all or most of the trajectories, and are hence likely to be relevant. Although errors arising from the force-field parameters and the choice of the starting model cannot be excluded, the generally good agreement between our simulations and a variety of experimental data suggests that gross artefacts are unlikely. In addition, the relatively high evolutionary conservation of some of the residues involved in the observed interactions gives further weight to the physiological relevance of our results.

The all-atom simulations presented here are even more speculative than the complex simulations, due to the short simulation times and the lack of repeats. In contrast, the coarse-grained simulations presented in the previous chapter can access much longer time-scales and provide better statistics, but at the expense of accuracy available to atomistic approaches. The very good agreement between the results from the two types of simulations, however, suggests that the coarse-grained model reproduces atomistic behaviour of the DNA sufficiently well, allowing us to infer on DNA dynamics and the Pol recognition mechanism with both high accuracy and statistical reproducibility.

In the short term, we could extend our complex simulations in several ways, such as by repeating them in the presence of *in silico* substitutions of the relevant residues, by testing different and novel force fields, or by employing coarse-grained or biased approaches to extend the simulation time scales. Most crucially, our simulations make a number of predictions that are directly testable experimentally. The unwinding of the non-template flap could be tested by 2-aminopurine assays, which give distinct circular dichroism spectra depending on whether two neighbouring bases are stacked or not. These assays could also be performed with residue-specific mutants (such as in Y719), to determine the mechanistic steps of strand displacement in more detail. Finally, the effect of substitutions of DNA-interacting residues (K549, R779 and R784) on DNA-binding and on the strand-displacement and DNA-synthesis activities of Pol could be tested with both biochemical and single-molecule methods.

5.9 Materials and methods

5.9.1 Model preparation

The protein atoms and the catalytic magnesium ion were extracted from the *Bst* X-ray structure PDB file (code 4BDP, reference [26]). Online server 'WHAT IF' was used to check for errors in the PDB, and build the missing side chains into the structure [179, 180]. PyMol (Schrödinger, LLC) was used to align the FRET-restrained structure with the X-ray structure, based on the protein component only. The downstream DNA in the docked structure was extended to its full length, and its template strand linked to the template strand of a 5-nucleotide fragment of upstream DNA from the X-ray structure (which also includes the templating nucleotide and one nucleotide downstream). Basic molecular sculpting was performed such that the conformation of the DNA backbone was not significantly disturbed, and no steric clashes occurred with the polymerase, which resulted in 6 base-pairs of the downstream DNA being unpaired. The upstream DNA fragment was then extended using the sequence of upstream DNA from the docked structure. This step was justified by the excellent agreement in the position of the upstream DNA between the X-ray and docked structures (see Section 4.8).

For DNA-only simulations, DNA models were generated using the 3D-DART server [145]. DNA atoms were extracted from the PDBs and the terminal phosphate groups removed using PyMol. In the case of gapped DNA, the central nucleotide was removed and a 5' phosphate group generated instead.

5.9.2 Force fields and parameters

All complex simulations and high-temperature DNA simulations were run using the Amber ff99sb force field [87], with modified nucleic acid parameters (parmbsc0, [150, 153]). The crystal structure control and the DNA-only simulations, which were carried out first, were run using Amber ff99sb-1LDN with Amber94 nucleic acid parameters. No parameters were available in either force field for the 5' phosphate groups of DNA, as these are

usually missing in crystal structures due to their high flexibility. The phosphate group had to be modelled at the 5' end of the gap in our DNA substrate, as it is both physiologically relevant and present in our single-molecule experiments. Therefore, the force fields were modified by assuming that the parameters of the β-phosphate of free ADP, available online [181], are a reasonable approximation for the α-phosphate of the gap-proximal thymine residue.

5.9.3 Simulation conditions

All simulations were carried out using Gromacs 4.6 [90]. The X-ray structure control simulations and the complex simulations were done using explicit solvent (TIP3P) in a triclinic box, with a minimum 10-Å solvent edge, in the presence of 10 mM $MgCl_2$. The system was neutralized with addition of magnesium ions, and energy-minimized using steepest-descent minimization. In order to stabilize the temperature of the system, equilibration was performed in the NVT ensemble for 100 ps, with the temperature of 298 K maintained using a Berendsen thermostat [109]. Next, the pressure of the system was stabilized by equilibration in the NPT ensemble for 1 ns, with the temperature of 298 K and the pressure of 1 bar retained using a V-rescale thermostat [182] and a Berendsen barostat [109], respectively. During equilibration, DNA atoms, protein heavy atoms and the catalytic magnesium ion were position-restrained with a force constant of 1,000 kJ-$mol^{-1}nm^{-2}$. DNA was equilibrated for an additional 10 ns with the protein heavy atoms restrained, under the NPT conditions. Atom velocities were preserved between the equilibration steps, and between the equilibration and production steps. Unrestrained production was finally allowed to run for 100 ns, with the temperature of 298 K and the pressure of 1 bar maintained by the V-rescale thermostat and a Parrinello-Rahman barostat [111]. Periodic boundary conditions and the Verlet cut-off scheme were used, and long-range electrostatic interactions were accounted for by the particle-mesh Ewald method [93]. All bonds were treated as constraints with the LINCS algorithm, resulting in a time step of 2 fs. Coordinates were saved to an output trajectory every 5 ps. Repeat simulations were carried

out using different randomly numbered seeds, generating different initial atom velocities each time.

In the case of full-length DNA-only simulations, the conditions were the same except that a square box was used, with dimensions equal to the length of the DNA plus a 10-Å solvent edge. The NVT and NPT equilibration steps were performed, and the production times were 20 ns. In the case of high-temperature DNA simulations carried out as part of the model preparation, the conditions were the same as for the complex simulations except that the temperature during the equilibration and production runs was 400 K, and the production times were 2 ns. All DNA heavy atoms were position-restrained during the production runs, except for the 6 base pairs in the protein-proximal, downstream part of the DNA, which were unpaired in the starting configuration.

5.9.4 Analysis

All analysis was carried out using Gromacs 4.6 or 5.0, and VMD [183]. Trajectories were repaired for periodic boundary conditions, and processed to include only every 10th frame, corresponding to 50-ps steps. Maps of occupancy of the DNA and of polymerase residues during the simulation were created with VMD's volmap density function, using an isovalue of 0.001. RMSD and end-to-end distance measurements were done using standard functions in Gromacs. The flap-to-template H-bonds were quantified by measuring the number of bonds at any one time in the simulation, using a distance cut-off of 0.33 nm, and an angle cut-off of 30°. The position of residue Y719 relative to the DNA was calculated by measuring the distance between the centres of mass of the side chain of Y719 and individual DNA-base moieties. Pol-DNA interactions were detected by measuring the minimum distance between any nitrogen atom of a specific Pol residue and a specific phosphorous atom in the DNA, during the entire simulation. Distances below 0.4 nm were taken as indicating an interaction.

6 Determinants of Pol conformational stability

6.1 Introduction

6.1.1 Project rationale

As described in Section 2.4.2, Pol undergoes a well-defined conformational change as part of its catalytic cycle, which involves the movement of its fingers subdomain from an open to a closed state. Single-molecule FRET analysis has revealed that in its unliganded state, the protein does not assume any one conformation, but can rapidly interconvert between them [5, 13]. In addition, different ligands bias Pol into different conformations, with primer-template DNA inducing the open conformation, matched nucleotides triggering the closed conformation, and mismatched nucleotides stalling Pol in the partially closed conformation. These observations can be rationalized through the proximity of the ligand binding sites to the flexible regions of the fingers subdomain, with Pol-DNA and Pol-nucleotide interactions directly competing with intra-protein interactions. Crystal structures of Pol and MD simulations of the fingers-closing transition have identified many of the interactions involved in the conformational coupling, with conserved residues of the O-helix and particularly residue Y714 appearing to play key roles (Sections 2.4.2 and 5.1.3).

The lack of bias in Pol towards any conformation in the absence of ligands, and its high conformational sensitivity to ligand binding, imply a delicate balance in the energetic stabilities of the open and closed states. Indeed, substituting a single residue involved in the mechanism of fingers-closing can have drastic effects on the conformational equilibrium of Pol. For example, substitution Y766A in *E. coli* increased the bias of the unliganded polymerase towards the open state, whereas substitution Y766F had the opposite effect [5]. Similarly, substitutions Y766A, E710A and E710Q prevented the full transition from the partially closed to the closed state, highlighting the roles of these residues in sensing nucleotide complementarity and ensuring the high fidelity of DNA synthesis (Section 2.4.3).

However, proteins are stabilized in specific conformations as a result of complex networks of inter-residue interactions that are often spread throughout the protein structure [184]. Hence, both the basal conformational bias of Pol and its sensitivity to ligands are likely encoded also in regions of Pol structure far away from the conformationally flexible helices of the fingers subdomain. We sought to identify the stability-conferring residues in Pol, in order to better understand the mechanisms of conformational coupling, both in Pol and in proteins generally. Identifying these residues in Pol would also allow us to test the robustness of Pol stability to single-residue substitutions, and design variants of Pol that are biased towards the open or the closed state. This work could be exploited in experiments to probe the effect of conformational bias, and the dynamics of the open- and closed-state interconversion, on the various functional properties of Pol. In addition, being able to lock polymerases into one or the other conformation through ligand binding at a distant site could provide a means of inhibiting the catalytic activity of certain viral polymerases.

Because stability-determining residues cannot usually be inferred directly from the protein structure, we made use of a computational approach known as the energy decomposition method [184]. We set out to verify the importance of the identified residues by carrying out single-residue substitutions, and by measuring the conformational equilibria of Pol variants using single-molecule FRET. We first attempted these experiments using double-cysteine labelled Pol constructs, and then using unnatural amino acid (UnAA)-labelled constructs, in order to achieve the required labelling specificity. In the following

introductory sections, we give an overview of the theory and applications of the energy decomposition method, and describe the basics of the unnatural amino acid technology.

6.1.2 Energy decomposition method

Most proteins adopt a well-defined native conformation that usually corresponds to the minimum of their free-energy landscape. Protein folding and stability is ensured through a combination of entropic and enthalpic contributions that include the hydrophobic effect, hydrogen bonds, van der Waals and electrostatic interactions within the protein [185, 186]. These contributions serve to counter the effect of conformational entropy, which favours the unfolded state of the polypeptide chain. The folded structure is encoded in the protein sequence, although not all residues contribute equally to its stability. Mutagenesis experiments show that whilst substitutions of most residues do not significantly compromise protein structure or function, some residues are extremely sensitive even to subtle substitutions [187, 188]. Other than by mutagenesis, residue-specific effects on protein stabilization are difficult to probe experimentally, and hence several computational models have been proposed, with different degrees of emphasis on the entropic and enthalpic contributions [184].

Energy decomposition method is one such approach that focuses entirely on the protein interaction energy and neglects entropic contributions [184]. The reasoning behind this approximation is that entropic contributions, such as the hydrophobic effect, participate mainly in the generic collapse of the polypeptide chain into a globule, rather than determining the specific regions of stabilization in the protein structure. In general, EDM involves running an all-atom MD simulation of the protein of interest, and calculating non-bonded (van der Waals and electrostatic) interaction energies between pairs of residues, averaged over the MD trajectory. This analysis allows the construction of an interaction energy matrix M_{ij}, which is a symmetric matrix that can be decomposed into eigenvalues and eigenvectors as follows:

$$M_{\mathrm{ij}} = \sum_{\alpha=1}^{N} \lambda_{\alpha} \mu_i^{\alpha} \mu_j^{\alpha} \tag{6.1}$$

where N is the number of protein residues, λ_α is an eigenvalue, μ_i^α is its associated eigenvector and μ_j^α is the transpose of the eigenvector. If the N eigenvalues are labelled in an increasing order, such that λ_1 is the most negative, then the different terms in the sum of this equation approximate the real interaction energy M_{ij} to an increasing extent. The first term has the largest contribution, and if λ_2 is significantly greater (less negative) than λ_1, then the equation can be simplified into:

$$M_{\mathrm{ij}} \approx \lambda_1 \mu_i^1 \mu_j^1 \tag{6.2}$$

The components of the eigenvector μ_i^1 then indicate to what extent each amino acid contributes to the total stabilization of the protein.

Equation 6.1 applies to an idealized one-domain protein, where a subset of amino acids interacting strongly with each other in the core of the protein structure account for the majority of its stabilization. In more complex multi-domain proteins, where stability-conferring residues are more dispersed, at least one eigenvector is required to describe each domain, and additional eigenvectors are usually needed to account for interdomain interactions [189]. Analysis of the components of the essential eigenvectors reveals the distribution of the stabilization energy across the amino acids, and can be used to identify the hot-spots of protein stabilization. The essential eigenvectors can also be used to construct a simplified energy matrix, in a reverse process of the decomposition described in equation 6.1. The resulting matrix is filtered through a threshold to extract only the significant non-bonded interactions, and a clustering analysis is performed to reveal blocks of structural stabilization. This analysis can be used to identify the possible domains and structural elements within a protein [189].

Despite its fairly simplistic treatment of inter-residue interactions, and the lack of account for entropic effects, EDM has demonstrated good agreement with experimental data. In small, one-domain proteins with a single stabilizing core, EDM analysis has identified 60-80 % of the residues shown to be important for stabilization by mutagenesis experiments [184]. Analysis of three homologous proteins belonging to the calycin superfamily

has identified conserved hydrophobic and polar interactions contributing to stabilization, and has shown agreement with some hydrogen/deuterium-exchange NMR experiments and discrepancy with others [190]. EDM has also been applied to explain the high global stability of the Doppel protein compared to the Prion protein, and the higher sensitivity of the latter to temperature-induced misfolding and aggregation [191]. In this way, EDM analysis could reproduce experimental observations that were previously difficult to understand, considering the two proteins share almost identical 3D structures.

The effect of substitutions on energetic stabilization has also been studied, and it has been shown that substitutions can either affect the strength of stabilizing interactions (i.e. change the principal eigenvalue), or can perturb the actual network of interactions (i.e. change the pattern of eigenvector components) [192]. Substitutions often appear to affect a localized region of the native conformation, even though they may not necessarily themselves be located in that region, which could be explained by cooperative propagation of stability changes. Interestingly, substitutions that induce a higher stabilization of the protein have been observed to give a better fit between the energetic properties of the protein sequence and its fold topology. In this respect, EDM emerges as a useful tool in protein design, allowing a sequence to be 'optimized' to best stabilize a target 3D structure. In addition, the approach has been used to explain the different sensitivity of the HIV-1 protease mutants to known inhibitors, and design a new target site for development of alternative inhibitors, with therapeutic implications [193].

6.1.3 Unnatural amino acid technology

As illustrated in Chapter 4, single-molecule FRET is usually performed with organic fluorophores, which can be attached to cysteine residues of proteins using maleimide chemistry. For intra-molecular FRET, double labelling is performed on double-Cys variants of proteins, generally in a stochastic way that results in all permutations of labelling reactions. Labelling can be biased if the two sites have different reactivities [13, 194], or if one of the sites is reversibly protected [195], but none of these approaches are general, and complete

orthogonality is rarely achieved.

A novel approach that in theory allows perfect orthogonality is based on the unnatural amino acid technology. This approach involves expanding the genetic code of organisms to include amino acids beyond the common 20 natural amino acids, thus allowing the design of proteins with enhanced or novel activities [196]. Normally, each of the 20 amino acids is specifically loaded onto its cognate tRNAs by an aminoacyl-tRNA synthetase (aaRS) that is itself also specific for that amino acid / tRNA pair. The charged tRNA can then enter the ribosome and recognize the correct mRNA codon via its anticodon loop. UnAA technology relies on modifying one or several steps in this process, either *in vitro* or in the context of living cells, in order to achieve UnAA incorporation in response to a specific mRNA codon.

Conceptually, UnAA incorporation can be implemented by adding a new set of the cell-permeable unnatural amino acid, its corresponding tRNA and its cognate aaRS to the existing biosynthetic machinery (Figure 6.1) [196]. In practice, this is usually achieved by importing a heterologous tRNA/aaRS pair from a different domain of life, and mutating the anticodon loop of the imported tRNA to a nonsense (translation-termination) codon. The orthogonality of the tRNA/aaRS pair can be improved, and the heterologous aaRS is subjected to a process of directed evolution, in order to alter its specificity so that it uniquely recognizes the UnAA of interest. In order to incorporate the UnAA into any protein of interest, a nonsense mutation is introduced into its gene at the desired site. The gene-carrying plasmid is then co-transformed into *E. coli* with another plasmid that encodes the additional tRNA/aaRS pair, and protein expression carried out in UnAA-supplemented media.

One of the issues with this approach is that the nonsense codon is still used for translation termination, causing truncation products of the protein of interest to be expressed [198]. A superior strategy would be to use a biologically silent codon for UnAA incorporation, but no such codon exists in native biosynthetic systems. Instead, an *E. coli* strain compatible with this strategy has recently been produced by recoding all 321 of its native *UAG* stop codons to the alternative *UAA* stop codons (Figure 6.2) [199]. Notably,

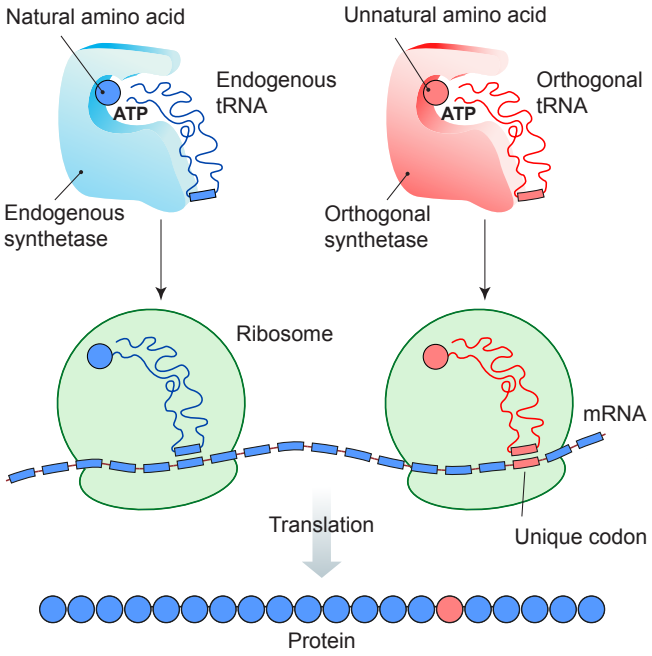


Figure 6.1: *Implementation of unnatural amino acid technology using a tRNA/aaRS pair (red) that is orthogonal to the endogenous pair (blue). The new tRNA recognizes a unique codon in the mRNA, resulting in UnAA incorporation into the polypeptide chain. Adapted from reference [197].*

translation termination in *E. coli* is normally achieved by release factor proteins, which have the following stop-codon specificity: RF1 recognizes *UAG* and *UAA* codons, whereas RF2 recognizes *UAA* and *UGA* codons. Therefore, if RF1 is eliminated from the genome and all native *UAG* codons are recoded to *UAA* codons, then *UAG* can be reassigned as a sense codon, allowing specific incorporation of unnatural amino acids. The recoded strain, termed C321, is freely available for use and compatible with the general protocol of UnAA incorporation described above.

A wide variety of unnatural amino acids have been successfully incorporated [198], including reactive groups that can be specifically labelled using orthogonal chemistries. Commonly used approaches for UnAA-based incorporation of fluorescent dyes have included oxime ligation, Staudinger ligation, and Cu-catalysed azide-alkyne cycloaddition [200]. One of the currently popular approaches, which can be performed at physiological pH and room temperature, is the Cu-free variant of azide-alkyne cycloaddition, in which the reaction is driven by the opening of a strained cyclooctane ring [201]. Due to

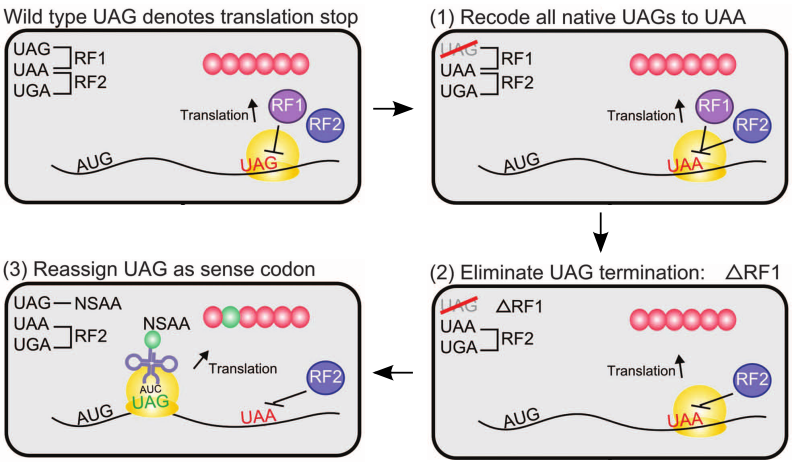


Figure 6.2: *Reassigning UAG codon function. (1) All native UAG codons are mutated to UAA codons. (2) Release factor RF1 is eliminated, preventing UAG termination. (3) Codon UAG can be used for unnatural amino acid incorporation. Adapted from reference [199].*

the bulky shape of strained alkynes that can compromise efficient tRNA charging by the aaRS [202], the reaction is preferably implemented by incorporating an azide-conjugated UnAA into the protein of interest, and attaching an alkyne group to the fluorescent dye [203]. In this thesis, we apply azide-alkyne cycloaddition for UnAA-based labelling by using azidophenylalanine as the UnAA, and dibenzylcyclooctyne (DBCO) as the functional group on the dye to be attached (Figure 6.3).

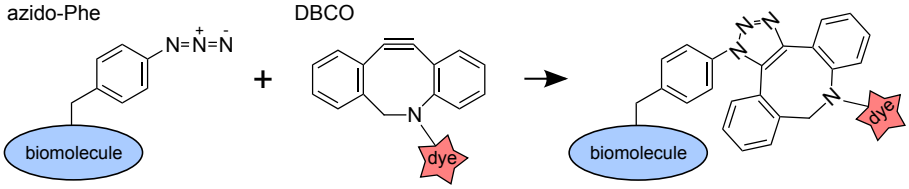


Figure 6.3: *Fluorescent labelling using strain-catalysed azide-alkyne cycloaddition reaction. Azido-Phe is incorporated into the biomolecule using UnAA technology, and is reacted with a DBCO-derivative of the fluorescent dye, resulting in specific labelling at the desired position in the protein.*

6.2 Identifying regions of stabilization

6.2.1 Energy decomposition of open and closed states

As a prerequisite for the EDM analysis of Pol, we ran MD simulations on both an open structure (in the presence of DNA; PDB code 1L3U [8]) and a closed structure of Pol (in the presence of DNA and the correct dNTP; PDB code 1LV5). We used the accelerated MD method (Section 3.2.8) to increase the sampling, and collected 500 ns and 300 ns of production time, for the open and closed conformations, respectively. The non-bonded interaction energy matrices were calculated and decomposed according to the EDM approach, and used to construct the essential interaction matrices (Figure 6.4a, b). The matrices show blocks of interactions indicative of the domain-like nature of Pol, particularly in terms of the 3'-5' exonuclease domain (residue numbers 297-468) and the thumb of the polymerase domain (residues 496-595). The open- and closed-state matrices are similar in terms of the general regions of stabilization (blue) and destabilization (red), as expected given the identical amino acid sequence and the similar conformations of the two structures.

To identify the residues that are most important for the stabilization of the open and closed conformations, we created a vector whose components are the sums of the elements in each matrix column, and therefore indicate the contribution of each residue to the total interaction energy in the structure (Figure 6.4a, b). Most of the stabilizing regions were shared between the open and closed structures, however, some appeared to primarily stabilize one or the other conformation. Specifically, regions stabilizing the open structure include parts of the N- and O-helices of the fingers, the β-sheet region connecting the P- and Q-helices between the fingers and the palm, the β-sheet region at the base of the palm, and the C-helix of the 3'-5' exonuclease domain (Figure 6.4c). In contrast, regions stabilizing the closed structure include parts of helices H1, H2 and I of the thumb subdomain (Figure 6.4d). Some of the highlighted residues in the fingers subdomain are directly involved in the induced-fit mechanism of fingers-closing, and the tip of the thumb subdomain could influence the fingers-closing transition through its effect on DNA binding. However, a direct conformational mechanism for how any of the other regions highlighted

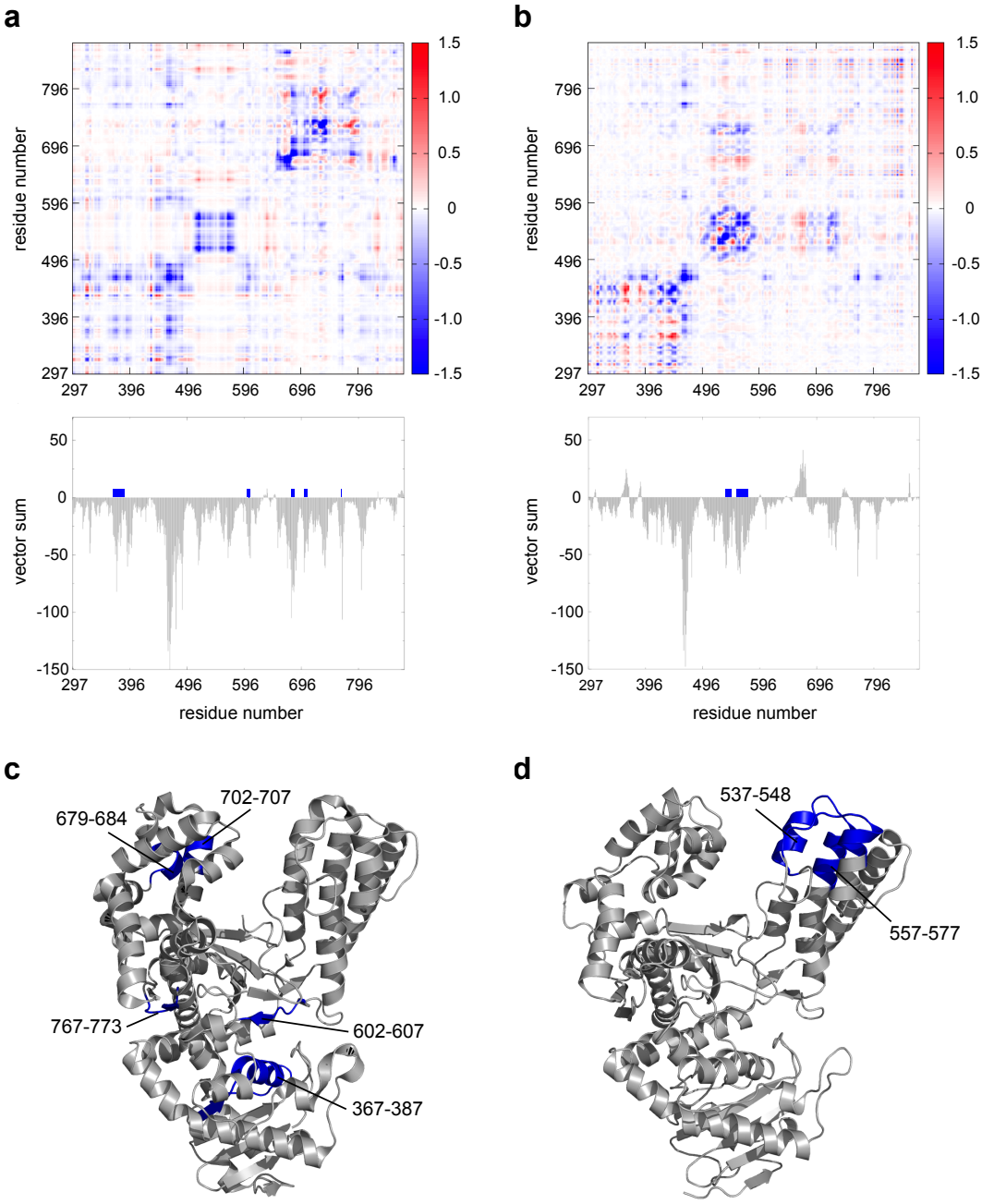


Figure 6.4: *Energy decomposition of open and closed structures of Pol.* ***(a)*** *Top, essential interaction matrix for the open structure. The blue and red colours represent the degree of stabilization (negative values) and destabilization (positive values), respectively. Bottom, vector sum of the above matrix, with the key sequence regions contributing specifically to the open structure stabilization annotated with blue boxes.* ***(b)*** *Essential interaction matrix and its corresponding vector sum, for the closed structure of Pol.* ***(c)*** *Stabilizing regions, corresponding to the blue boxes in (a), mapped onto the open structure of Pol.* ***(d)*** *Stabilizing regions, corresponding to the blue boxes in (b), mapped onto the closed structure of Pol.*

by the EDM analysis could affect the open-closed equilibrium of Pol is not obvious.

6.2.2 Local flexibility analysis

To further explore the stability of the open and closed conformations, we analysed the trajectories for local protein backbone flexibility, in terms of the distance fluctuations of any residue in the protein with respect to its neighbouring residues (see Section 6.7.2). Whilst the two structures produced similar signatures of flexibility, we identified a number of regions that showed higher flexibility in one or the other structure (Figure 6.5a). The majority of these regions mapped onto the fingers domain, as expected, but some mapped to other areas far away from the conformationally flexible helices (Figure 6.5b, shown in orange). We noted that some of the highlighted regions also overlapped with or were adjacent to the stabilizing regions highlighted above (Figure 6.5b, annotated).

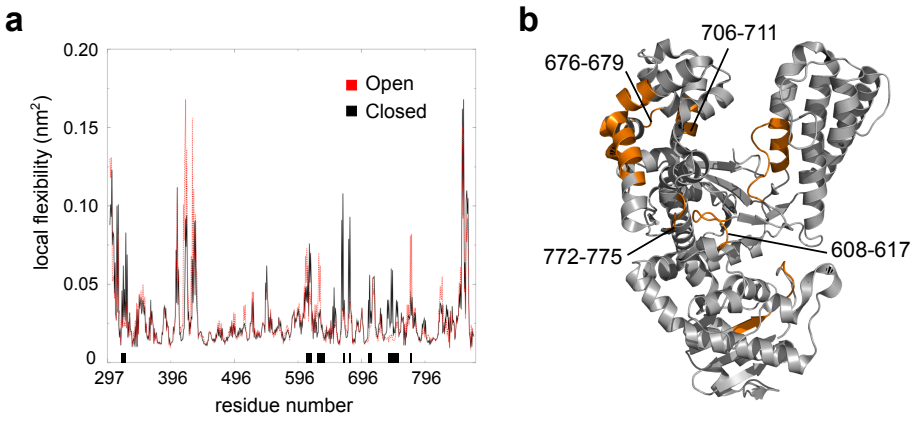


Figure 6.5: *Local flexibility analysis of open and closed structures of Pol.* ***(a)*** *Average local flexibility, as a function of residue number, shown for the open (red) and closed structures (black). Regions that show a significant difference in flexibility between the open and closed conformations are annotated with black boxes.* ***(b)*** *Open structure of Pol, with regions corresponding to the black boxes in (a) shown in orange. The regions that overlap with or are adjacent to the regions of stabilization (Figure 6.4c, d) are annotated with residue numbers.*

6.2.3 Selection of single-residue substitutions

In order to test the predictions of the EDM analysis experimentally, we aimed to find specific residues that stabilize the open or the closed conformation of Pol. We were interested in exploring the *indirect* mechanisms of stabilization, and hence focused on the regions away from the fingers subdomain. In addition, we filtered these regions to include only those that overlapped with or were adjacent to the regions showing significant differences in local flexibility between the open and closed conformations. Based on this analysis, the two β-sheet regions, between the fingers and the palm subdomains ('region 1') and at the base of the palm subdomain ('region 2'), appeared as good targets. Both regions are predicted to stabilize the open conformation of Pol, and hence substituting any residue in these regions would be expected to shift the conformational equilibrium of Pol towards the closed conformation. We selected five such residues as potential candidates for substitutions (Figure 6.6a).

Since the EDM analysis has been performed on the structure of the *Bst* protein, and our experimental pipeline has been established for the *E. coli* protein, we verified the conservation of the residues through sequence and structural alignment of the two proteins. Only three out of the five relevant residues appeared conserved in *E. coli*, and were hence chosen as the residues to be substituted: R822, Y824 and Y659 (Figure 6.6a, b). As a negative control, we looked for regions that showed no significant stabilization in either the open or the closed structure; substitutions in these regions would therefore be predicted to be 'neutral' in terms of their effect on the conformational equilibrium. Of the five relevant residues in *Bst*, three were conserved in *E. coli*. Residue I679 mapped close to the DNA-binding and active sites, and could therefore compromise Pol activity, leading us to choose G575 and N579 as the 'neutral' residues to be substituted.

a

Region 1 residues

Bst	R770	Y772
Eco (sequence)	**R822**	**Y824**
Eco (structure)	**R822**	**Y824**

Region 2 residues

Bst	F371	W382	F606
Eco (seq)	E412	Y423	**Y659**
Eco (str)	Q419	N430	**Y659**

Neutral residues

Bst	G523	N527	I626
Eco (seq)	**G575**	**N579**	I679
Eco (str)	**G575**	**N579**	I679

b

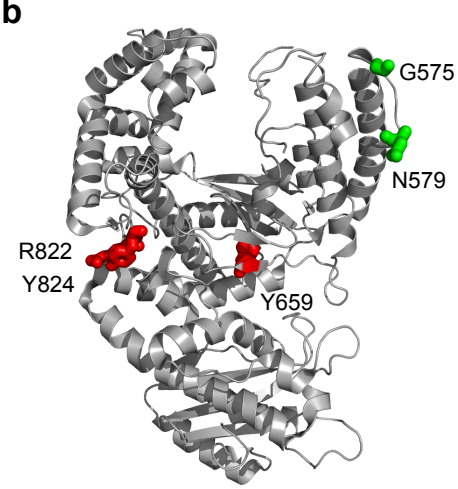


Figure 6.6: *Selection of single-residue substitutions.* ***(a)*** *A subset of Bst residues that, according to EDM analysis, stabilize the open state of Pol (top and centre), or do not affect the open-closed equilibrium (bottom table). E. coli residues corresponding to these residues, based on sequence and structural alignment, are shown in the respective tables. In the case where both sequence and structural alignment indicated that the residue was conserved in the two proteins, the residue is highlighted in bold.* ***(b)*** *Closed structure of E. coli Pol, with the residues selected for substitutions shown in stick representation, and coloured red (stabilization sites) or green (neutral sites).*

6.3 Double-cysteine Pol variants

6.3.1 Site-directed mutagenesis

We initially attempted to test the effect of substitutions on Pol conformational equilibrium using double-cysteine fluorescent labelling, taking advantage of the different reactivities of the previously established fingers and thumb labelling positions (Figure 2.7 and reference [13]). Therefore, as a starting construct we used the *E. coli* Pol (Klenow fragment) gene with mutations encoding the following substitutions: D424A (to eliminate the exonuclease activity of Pol), K550C (the thumb labelling position), L744C (the fingers labelling position) and C907S (to remove the native Cys residue). We aimed to introduce mutations encoding substitutions R822A, Y824A and Y659A into the Pol gene using QuikChange site-directed mutagenesis, an approach that relies on polymerase chain reaction (PCR)-based amplification of the template DNA using mutagenic primers [204]. DNA sequen-

cing results indicated that at least one of the tested colonies transformed with reactions for substitutions Y822A and Y824A contained the plasmid with the desired mutation, whereas all colonies transformed with the reaction for substitution Y659A showed either parental DNA sequence (with no new mutation) or a primer-insertion event. To prevent any primer-template mispairing, the annealing temperature of the reaction for substitution Y659A was increased from 55 °C to 72 °C, which finally also resulted in the introduction of the desired mutation.

6.3.2 Expression and purification

We set out to express and purify Pol variants R822A and Y824A in *E. coli*, using the protocols previously optimized for the 'wild-type' protein (see Section 6.7.6). [1] We obtained a good yield of cells (~1.5 g per litre of culture), suggesting that the substitutions did not affect cell growth. Similarly, His-tag purification on a nickel-nitrilotriacetic acid (Ni-NTA) column proceeded as efficiently as with the wild-type protein, producing ~70 nmol of Pol (R822A) and ~50 nmol of Pol (Y824A). Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) analysis indicated that the majority of contaminants were removed during the purification, that Pol was the major species present in the elution fractions, and that only a very small amount of Pol was lost on the column (Figure 6.7).

6.3.3 Double labelling

We performed double labelling of Pol variants R822A and Y824A with Cy3b and Atto647N, using a previously described double-cysteine labelling protocol [13]. Wild-type Pol was labelled simultaneously as a reference sample, so that any deviations in yields or labelling efficiencies could be controlled for. The higher maleimide reactivity of the thumb position (550) compared to the fingers position (744) allows a degree of selectivity with double labelling, provided that labelling is done in a step-wise fashion [13]. Hence, we labelled

[1] In this chapter, we use the term 'wild-type Pol' to refer to Pol (Klenow fragment) containing substitutions D424A, K550C, L744C and C907S, but not any additional substitutions affecting Pol conformational equilibrium.

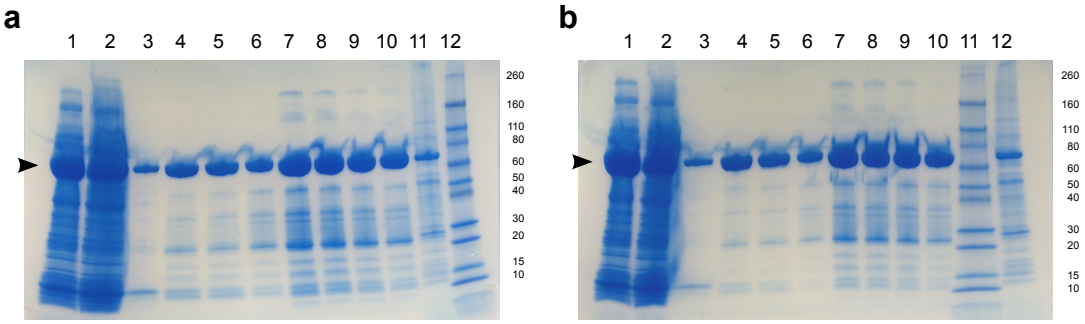


Figure 6.7: *SDS-PAGE analysis of His-tag purification of Pol variants* ***(a)*** *R822A and* ***(b)*** *Y824A. The lanes in (a) correspond to the following samples: (1) sample before binding to Ni-NTA resin, (2) unbound protein (flow-through), (3) 10 mM imidazole wash, (4) 20 mM imidazole wash, (5) 27 mM imidazole wash, (6-10) elution fractions 1-5, (11) Ni-NTA resin after elution, (12) protein ladder. The numbering in (b) is the same except that (11) and (12) are reversed. The bands corresponding to Pol (68 kDa) are indicated with arrowheads. Protein ladders are annotated on the right-hand side of each gel, with sizes given in kDa.*

the thumb position with Atto647N by incubating the purified Pol variants with Atto647N-maleimide for 1 hour at room temperature, and then labelled the fingers position with Cy3b, by additionally incubating the samples with Cy3b-maleimide over-night at 4 °C. Nanodrop analysis indicated reasonable yields of all proteins, with ~5 pmol protein present in the most concentrated fractions (Table 6.1, top). The labelling efficiencies for the wild-type protein were similar to what had been observed before, whereas the efficiency of Atto647N labelling was considerably lower for variants R822A and Y824A. This result suggested that the labelling protocol was carried out correctly, and that the substitutions were affecting the efficiency of the labelling reaction.

In order to better understand the reasons behind the observed low labelling efficiencies, we set to analyse the labelling bias in our samples. We made use of a previously designed assay that relies on partial digestion of labelled Pol constructs by limited exposure to the protease chymotrypsin [13]. Digestion produces a series of fragments that give a characteristic pattern of bands on an SDS-PAGE gel in the green and red fluorescence channels. If singly-labelled Pol (K550C) and Pol (L744C) samples are digested and analysed simultaneously, then some of the bands observed in the doubly labelled sample can be assigned

to fragments labelled at either position 550 or 744. This analysis allows one to estimate the percentage of molecules that are green- or red-labelled at each position, and hence the percentage of molecules that are doubly labelled 'correctly', i.e. with the red dye on the thumb position and the green dye on the fingers position (R550/G744), as opposed to the reverse orientation (G550/R744).

- DNA labelling	**[Pol]** (μM)	**[Cy3b]** (μM)	**[Atto647N]** (μM)	**efficiency (Cy3b)**	**efficiency (Atto647N)**
wild-type	23.0	15.9	21.7	0.69	0.94
R822A	19.1	12.7	9.6	0.66	0.50
Y824A	21.2	10.9	8.9	0.51	0.42

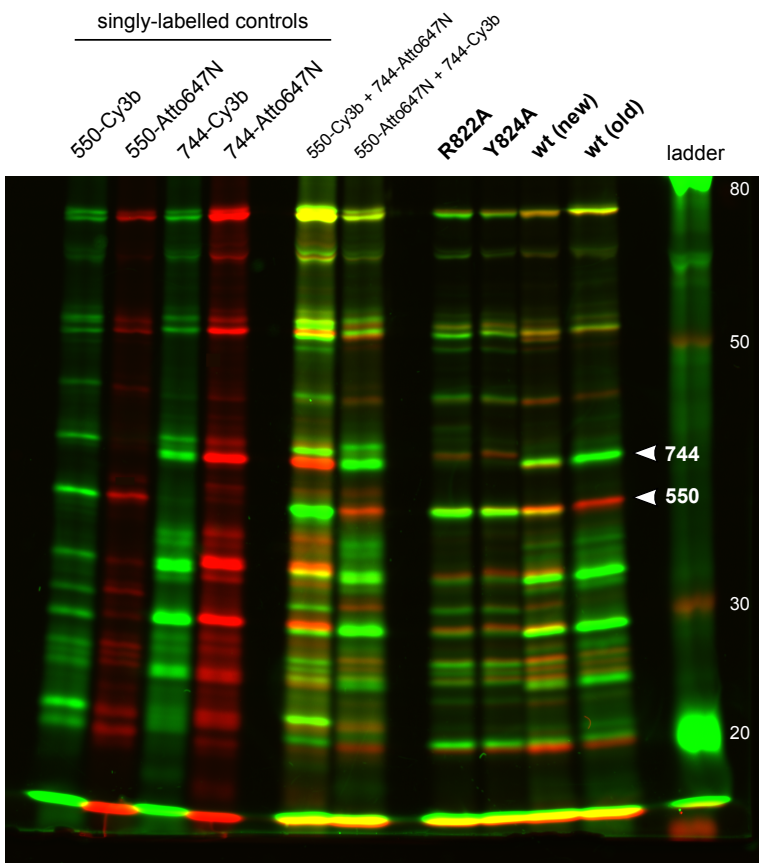
+ DNA labelling	**[Pol]** (μM)	**[Cy3b]** (μM)	**[Atto647N]** (μM)	**efficiency (Cy3b)**	**efficiency (Atto647N)**
wild-type	14.4	5.2	7.9	0.36	0.55
R822A	21.3	9.5	7.4	0.45	0.35
Y824A	15.1	8.8	9.3	0.58	0.62

Table 6.1: *Cy3b and Atto647N labelling of double-Cys Pol variants, in the absence (top) and in the presence of DNA to increase labelling bias (bottom table). Corrected concentrations of protein, Cy3b and Atto647N are given. Labelling efficiencies were calculated from the respective concentrations.*

Applying this assay to our newly labelled samples, we observed a labelling bias of ~75 % towards the R550/G744 species for the wild-type Pol (Figure 6.8), only slightly lower than the previously reported value of 88 % [13]. Surprisingly, however, the bias was on the order of 10 % towards the R550/G744 species for the two Pol variants, suggesting that they were labelled almost in the opposite orientation compared to the wild-type protein. Since it has previously been noted that the G550/R744 species can artefactually adopt the closed state of the protein (see Section 6.5.2), such a high population of 'incorrectly' labelled Pol was not suitable for confocal analysis, and we therefore sought to improve the labelling bias for variants R822A and Y824A.

An approach that has previously been shown to improve the fingers/thumb labelling bias relies on labelling Pol in the closed state [5]. This is achieved by incubating Pol with a primer-template DNA and the complementary nucleotide prior to the labelling reaction.

a



b

Pol variant	- DNA labelling		+ DNA labelling	
	R550 G744	**G550 R744**	**R550 G744**	**G550 R744**
wild-type (old)	/	/	1.00	0.00
wild-type (new)	0.74	0.26	0.85	0.15
R822A	0.08	0.92	0.41	0.59
Y824A	0.11	0.89	0.68	0.32

Figure 6.8: *Estimation of Pol-Cy3b/Atto647N labelling bias using chymotrypsin digestion assay.* ***(a)*** *SDS-PAGE analysis of digested samples, labelled in the absence of DNA. Green and red in-gel fluorescence are shown overlaid. Lanes 1-4 refer to individual singly-labelled controls, lanes 5-6 to 1:1 mixes of singly-labelled controls, and lanes 7-10 to the doubly-labelled samples of interest. Both the newly-labelled wild-type Pol (wt new), and a wild-type Pol sample that was previously labelled in the presence of DNA (wt old), are analysed. The two bands that were used for estimation of 550- and 744-labelling bias are shown with white arrowheads. A protein ladder was run for reference; the bands that are fluorescent in either channel are annotated, with sizes given in kDa.* ***(b)*** *Relative proportion of R550/G744 and G550/R744 species, obtained by labelling in the absence and in the presence of DNA, calculated from in-gel fluorescence intensities.*

The protocol uses a heparin column as the final purification stage, which ensures the removal of both the unreacted dye and the DNA ligand from the Pol sample. Initial attempts to apply this approach to the labelling of variants R822A and Y824A resulted in very poor yields, which could be due to strong binding of the variant proteins to the DNA (preventing the heparin from out-competing the DNA), or due to poor binding to both the DNA and the heparin. Binding of the resulting flow-through to the Ni-NTA resin indeed resulted in partial recovery of the protein, suggesting that some of the protein had been released from the column prior to elution.

Hence, we modified the biased labelling protocol in the presence of DNA by replacing the heparin purification step with a His-tag purification step. Unfortunately, labelling the two Pol variants in this way did not markedly improve any of the labelling efficiencies, and resulted in significantly poorer labelling of the wild-type protein (Table 6.1, bottom). We further analysed the labelling specificity of the samples labelled with this approach, and found the bias to have increased to 40 % in the case of variant R822A, and almost 70 % in the case of Y824A (Figure 6.8). The bias was also increased for the wild-type protein, although not to the level expected for biased labelling of Pol (~100 %, reference [13]).

6.3.4 Confocal analysis

We measured the conformational equilibria of the labelled variants by single-molecule FRET, using confocal (ALEX) microscopy. We first tested the wild-type and variant Pol samples labelled in the absence of DNA, and recorded E/S histograms for the unliganded proteins, the binary complexes (in the presence of primer-template DNA) and the ternary complexes (in the presence of the DNA and correct or incorrect dNTPs). The number of FRET events was low, due to the poor labelling efficiencies, and required the concentration of the labelled protein to be increased to a level that was suboptimal for single-molecule imaging. Despite our single-molecule sorting capabilities, the number of coincidence events arising from singly-labelled species (Figure 6.9a, curve-shaped populations on the E/S histograms) was high enough to interfere with the FRET events (boxed

populations on the E/S histograms). However, we were able to extract the FRET histograms of the different states of the wild-type and the R822A and Y824A variants, and used restrained Gaussian fitting to estimate the proportions of the open, partially-closed and closed conformations in each case (Figure 6.9a, b).

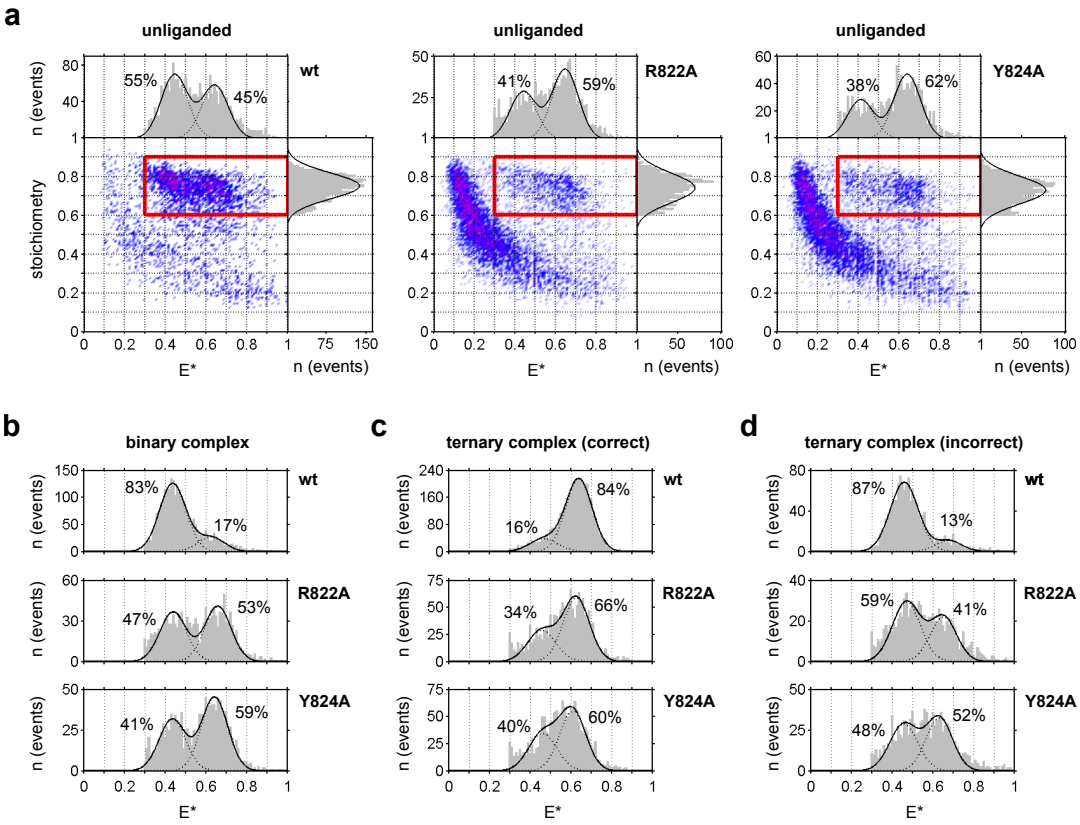


Figure 6.9: *Confocal analysis of double-cysteine Pol variants, labelled with Cy3b and Atto647N, in the absence of DNA.* ***(a)*** *Uncorrected E/S histograms of wild-type Pol (left), Pol variant R822A (centre) and Y824A (right), in the unliganded state. The one-dimensional histograms represent the data defined by the red box. The relative proportions of the open and closed populations, estimated from Gausian fits to the histograms, are indicated.* ***(b-d)*** *Apparent FRET efficiency histograms of Pol variants, (b) in binary complex with primer-template DNA, (c) in ternary complex with the correct dNTP, and (d) in ternary complex with an incorrect dNTP. The relative proportions of the open and closed conformations (binary complex), and of the partially closed and closed conformations (ternary complexes), are indicated.*

The histograms indicated that the newly labelled wild-type Pol behaved similarly to what had previously been observed, showing a ~50:50 ratio of the open and closed con-

formations in the unliganded state, and adopting primarily the open conformation in the binary complex, the closed conformation in the ternary complex with the correct dNTP, and the partially closed conformation in the ternary complex with an incorrect dNTP. By comparison, the two variants showed milder sensitivity in response to the different ligands, and were more biased towards the closed conformation in the unliganded, binary and ternary-incorrect states. However, whereas both variants showed an additional degree of closing in the ternary-correct state compared to the unliganded state, this effect was less pronounced than in the wild-type protein, suggesting that the variants may be compromised in DNA or dNTP binding. The two variants behaved similarly to one another, with variant R822A giving a slightly more wild-type-like behaviour than variant Y824A.

We further tested the Pol samples labelled with the biased approach, in the presence of the DNA ligand, which contained smaller populations of 'incorrectly' labelled Pol. Unfortunately, however, preliminary confocal analysis of these samples indicated an even poorer separation of the coincidence and FRET events, and showed broad FRET distributions (Figure 6.10). A significant number of high-FRET events (E* > 0.8) were observed, suggesting that protein aggregation or non-specific labelling could be involved. We therefore halted further confocal analysis of Pol variants until better-quality samples could be obtained.

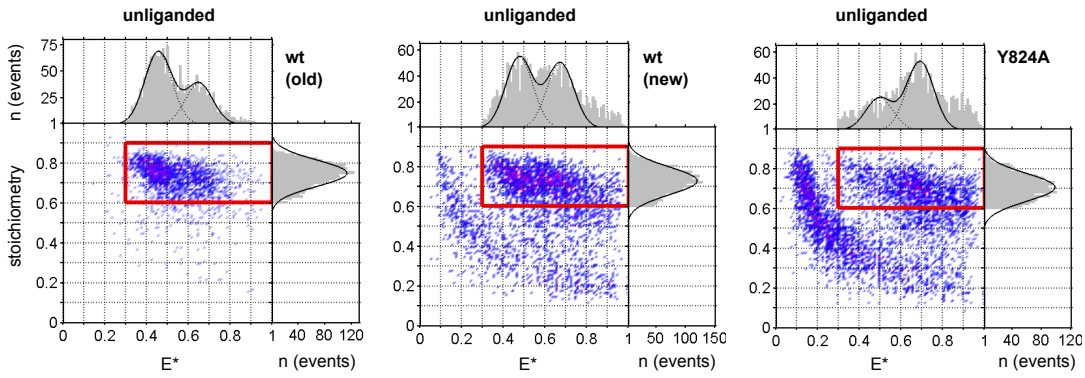


Figure 6.10: *Confocal analysis of double-cysteine Pol variants, labelled with Cy3b and Atto647N, in the presence of DNA. Shown are uncorrected E/S histograms of previously labelled wild-type Pol (left), newly labelled wild-type Pol (centre) and Pol variant Y824A (right), in the unliganded state.*

Although additional work could have been done to optimize the double-cysteine labelling of Pol, we noted that it was unlikely that we could obtain identical labelling bias for the wild-type protein and the variants. Since the investigated substitutions affect the labelling bias, and the labelling bias affects the apparent conformational equilibrium, the direct effect of substitutions on the conformational equilibrium (that was not due to labelling artefacts) therefore could not be evaluated using this approach. Hence, in the next section we attempted to develop a new labelling strategy that would allow a truly orthogonal double labelling of Pol and its variants.

6.4 Unnatural amino acid-modified Pol variants

6.4.1 Project design

In order to achieve orthogonal labelling of Pol and its variants, we turned to the unnatural amino acid technology (Section 6.1.3). UnAA-based double labelling for smFRET can be implemented either using a maleimide-UnAA pair [205], or using two UnAAs that react with different chemistries [206]. We chose to replace the Cys at the thumb position of Pol with an unnatural amino acid, whilst retaining the Cys at the fingers position for maleimide labelling. We used the UnAA azidophenylalanine, which can be reacted specifically with dibenzylcyclooctyne derivatives of commercially available dyes, such as Cy5 or Atto647N (Figure 6.3).

As described in Section 6.1.3, expression of unnatural amino acid-modified proteins is often compromised by premature termination during translation. One way to avoid this issue is to ensure that the affinity-purification tag of the protein is encoded at its C-terminal end, such that only fully translated constructs are purified. However, in the case of Pol the C-terminal end of the protein is close to the active site, and hence attaching a His-tag at this end could interfere with Pol function. Furthermore, C-terminal tagging does not address the issue of poor expression yields that usually result from premature termination. We therefore took an alternative approach and made use of the *E. coli* strain C321,

in which codon *UAG* has been reassigned for unnatural amino acid incorporation [199]. To allow transformation and expression using standard protocols, we prepared chemicompetent cells of the C321.dA.exp strain, a variant of C321 optimized for expression at 37 °C.

Notably, the C321 cell strain is not compatible with the pET expression system used for expressing double-Cys Pol variants (Section 6.7.6), as it lacks a chromosomal copy of the T7 RNAP gene. In addition, arabinose induction of protein expression is preferred to the standard isopropyl β-D-1-thiogalactopyranoside (IPTG) induction, as it first allows simultaneous expression of the tRNA/aaRS pair from one of the commercially available plasmids. Therefore, we opted to use vector pBAD24 [207], a popular vector that hosts an arabinose-inducible promoter. We designed a Pol gene containing all previous mutations (Section 6.3.1), with the mutation encoding K550C replaced to encode the *UAG* codon, and ordered the gene to be commercially synthesised and cloned into the pBAD24 vector. In addition, Cys-less variants of both Pol and full-length Pol were ordered to allow flexibility in future project development. We verified the sequence of the synthesised genes by commercial DNA sequencing.

6.4.2 Expression trials

We first tested the expression of Pol and full-lenth Pol Cys-less genes from the pBAD24 vector, in both the cell strain used previously for Pol expression (HMS174) and the new cell strain (C321). Initial attempts to transform pBAD plasmids containing Pol and flPol into C321 cells failed, presumably due to the lack of required modifications in the commercially synthesised DNA. We therefore had to first transform the pBAD plasmids into XL-1 Blue Supercompetent cells and isolate the replicated plasmids, before transforming them into C321 cells. We then tested the expression using both normal arabinose induction, followed by 3 hours of expression, and by using a commercially available autoinduction medium, supplemented with arabinose, which allowed over-night expression. We carried out these expression trials on a small scale (in 5 ml cultures), as they were intended only for analytical purposes, in order to find the suitable conditions for large-scale expression.

Prior to manual induction, both HMS and C321 cells grew well (Figure 6.11a), although the latter required more time to reach an optical density (OD) suitable for induction (~2 and ~3 hours, respectively). Samples were taken from the manually induced cultures after 1.5 and 3 hours of expression, and from the autoinduced cultures after the over-night expression. SDS-PAGE analysis indicated successful expression of Pol and flPol in all cases, with a higher apparent protein yield obtained from the C321 cells compared to the HMS cells (Figure 6.11b). The manual induction produced a cleaner expression than the autoinduction, with a better ratio of the amount of Pol compared to the amounts of other proteins in the medium. Finally, more Pol had accumulated after 3 hours, compared to 1.5 hours of expression, highlighting the importance of the long expression time. As expected, the expression of flPol was slightly less efficient than the expression of Pol, due to the different size of the two constructs.

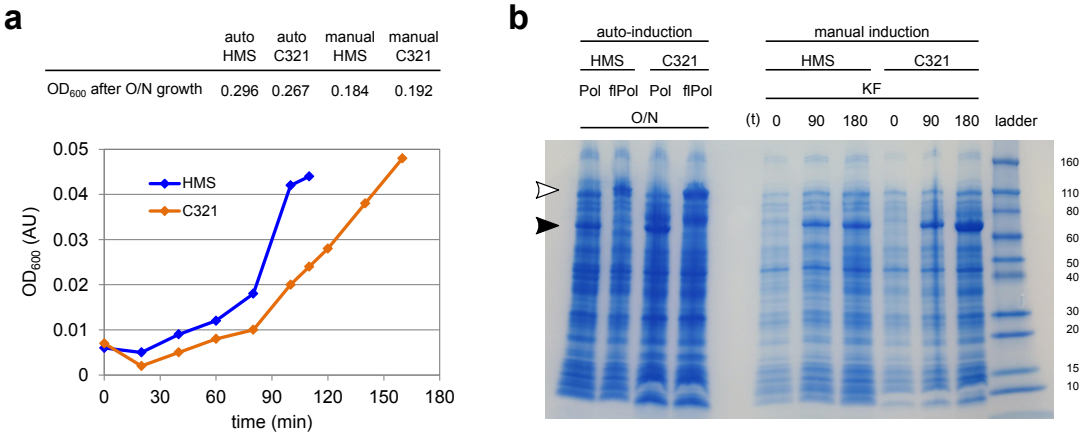


Figure 6.11: *Expression trials for Cys-less Pol and flPol in pBAD24 vector, in HMS and C321 cells.* ***(a)*** *Top, OD_{600} values after over-night growth of HMS and C321 cells, in the autoinduction (auto) or manual-induction media (manual). Bottom, growth curves of day cultures of HMS and C321 cells, prior to manual induction of Pol expression. The OD_{600} values were measured using a 1-mm light path.* ***(b)*** *SDS-PAGE analysis of Pol and flPol expression. Expression was performed either using auto-induction or manual induction, in HMS or C321 cells, and samples taken either after over-night expression (O/N) or at a specific time (t) following manual induction (0, 90 or 180 min). The positions of Pol and flPol are indicated with black and white arrowheads, respectively. A protein ladder was run for reference and is annotated, with sizes given in kDa.*

Next, we tested the expression of unnatural amino acid-modified Pol, both using autoinduction and manual induction. Two different plasmids were tested for expression of the tRNA/aaRS pair that allows azidophenylalanine incorporation: pEVOL and pDULE2. Therefore, we co-transformed C321 cells with pBAD24-Pol and either pEVOL-AzF or pDULE2-AzF, innoculated successful transformants in lysogeny broth (LB) or the UnAA-supplemented autoinduction medium, and induced expression with 0.2 % arabinose. SDS-PAGE analysis indicated that expression was successful in all cases except in the case of pDULE2-based autoinduced expression (Figure 6.12), which could have been due to an experimental error. Regardless, manual induction produced higher yields of Pol than autoinduction, and thus appeared to be the induction approach of choice. There was no significant difference between pEVOL- or pDULE2-based expression, in terms of the yield of Pol or the presence of other protein contaminants. No expression was observed in the absence of arabinose or the unnatural amino acid (which served as negative controls), suggesting that expression was tightly regulated and that no amino acid misincorporation took place. As expected, there was no obvious band corresponding to a premature termination product of Pol, indicating that all of the expressed Pol was fully translated.

6.4.3 Expression, purification and labelling

We performed large-scale expression of UnAA-modified Pol in 1-litre culture, choosing pEVOL as the tRNA/aaRS plasmid and opting for manual induction with arabinose. All expression was carried out in the dark to avoid damage to the light-sensitive azide, and no reducing agent was included in the lysis buffer to prevent azide reduction. Cell growth was faster than in small cultures, needing only 2 hours to reach an OD of 0.6, at which point expression was induced. We obtained 2.6 grams of cells per 1 litre of culture, a higher amount than normally observed with HMS cells, and the SDS-PAGE analysis indicated successful expression of Pol (data not shown). We purified the UnAA-modified Pol immediately after expression, using the same protocol as for double-Cys variants, but avoiding the use of any reducing agent.

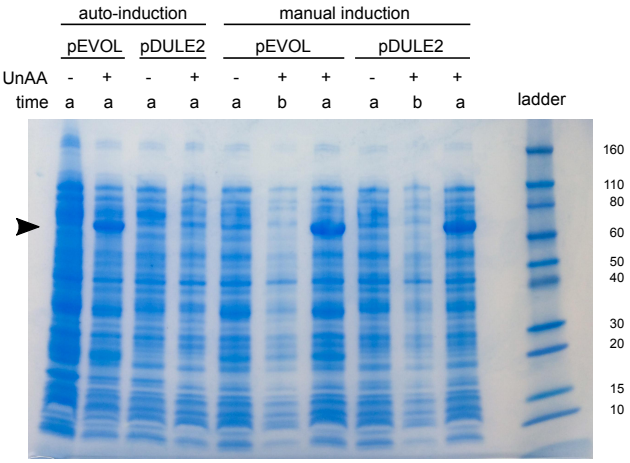


Figure 6.12: *SDS-PAGE analysis of UnAA-modified Pol expression in C321 cells. Expression was performed using either auto-induction or manual induction, with plasmid pEVOL or pDULE2, in the presence (+) or absence (-) of UnAA in the medium, and samples were taken either before (b) or after expression (a). The position of Pol is indicated with a black arrowhead. A protein ladder was run for reference and is annotated, with sizes given in kDa.*

We next attempted to label the UnAA-modified Pol with a DBCO-derivative of Cy5. We used Cy5 instead of Atto647N for these preliminary labelling attempts, as DBCO-Atto647N can only be synthesised on request and is therefore more expensive to use. Labelling was performed with 10- or 20-fold excess of the dye, at 4 °C or 25 °C, and for 24 or 48 hours (Figure 6.13a). We achieved a 50-55 % labelling efficiency regardless of the conditions used, suggesting that the reaction proceeds to completion already within 24 hours and under milder conditions. Increasing the temperature to 37 °C led to aggregation of Pol, as did a further increase in the molar excess of the dye, due to the resulting high concentration of organic solvent in the protein buffer (25 %). Using the chymotrypsin digestion assay, we obtained a labelling bias for DBCO-Cy5 of ~97 % for the 550 versus the 744 position.

We further attempted to label the Cy5-labelled Pol with Cy3b. To this aim, the Cy5-labelled sample was reduced with DTT, dialysed and incubated with 20-fold excess of Cy3b-maleimide. We used a higher concentration of Cy3b than normally (5-fold excess), to account for the fact that the starting concentration of the Cy5-labelled Pol was lower than usual. The latter was in turn was due to the Cy5 dye-removal step that was neces-

a

T (°C)	time (h)	dye excess	efficiency
4	24	10x	0.52
4	48	10x	0.53
4	48	20x	0.52
25	24	10x	0.54
25	48	10x	0.56
25	48	20x	0.56

b

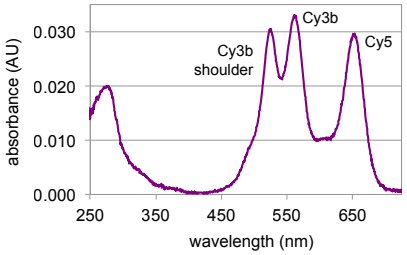


Figure 6.13: *Double labelling of UnAA-modified Pol. (a) Labelling efficiencies for DBCO-Cy5 labelling of UnAA-modified Pol, observed with different temperatures, incubation times and dye amounts. (b) Absorption spectrum of Pol, labelled with Cy3b and Cy5. The main Cy3b and Cy5 peaks are labelled, along with the additional Cy3b shoulder peak.*

sary for the quantification of labelling efficiencies. Labelling was performed over-night at 4 °C, and yielded a labelling efficiency of 92 % for Cy3b. However, the absorption spectrum of the doubly labelled Pol showed a very high shoulder of Cy3b that had not been observed before (Figure 6.13b), and could potentially arise from non-specific attachment of Cy3b to the protein, or from dimerization of two Cy3b molecules. The sample was too dilute to allow more thorough analysis, such as the estimation of labelling bias using the chymotrypsin digestion assay.

We carried out a preliminary confocal analysis of the doubly labelled Pol-Cy3b/Cy5 sample in the unliganded state and in the binary and ternary complexes. We noted that the ratio of the number of FRET to the number of coincidence events was markedly increased compared to the double-Cys samples (Figure 6.14), suggesting that a higher number of doubly-labelled species was present. However, the observed FRET populations occurred at significantly higher FRET values than what would be expected for the open and closed states of Pol, even accounting for the effects of dye linker length and the different Förster radii of Cy3b/Cy5 and Cy3b/Atto647N dye pairs (Section 8.2.1). It is possible that the non-standard conditions under which the sample was produced (including the presence of a high molar excess of Cy3b) compromised the quality of the sample. In addition, the potential non-specific attachment of Cy3b to the protein, or Cy3b oligomerization, could account for the unusual FRET signatures observed.

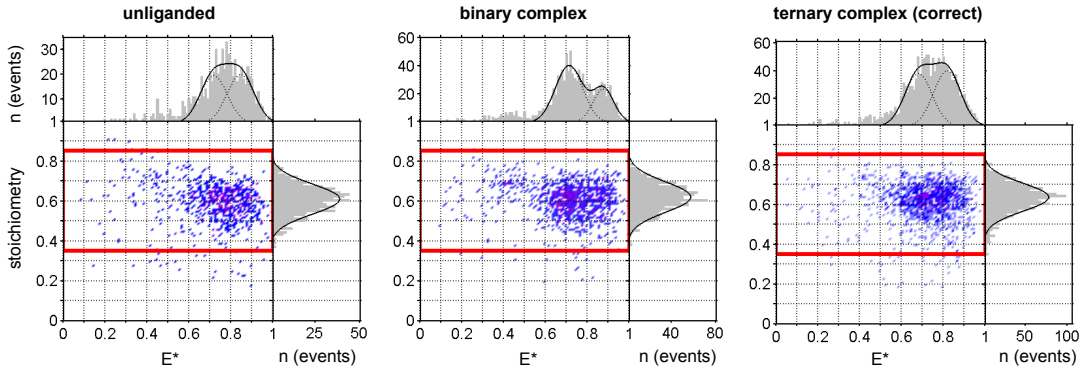


Figure 6.14: *Confocal analysis of UnAA-modified Pol, labelled with Cy3b and Cy5. Shown are uncorrected E/S histograms of Pol in the unliganded state (left), in binary complex with primer-template DNA (centre), and in ternary complex with the correct nucleotide (right).*

6.4.4 Single-point mutagenesis

In order to introduce mutations encoding substitutions R822A, Y824A and Y659A into the gene encoding UnAA-modified Pol, the same primers were used that were designed for the double-Cys variants (Section 6.3.1). In addition, new primers were designed for two 'neutral' substitutions, N579A and I679A. The mutagenesis conditions were kept the same as before, except that an annealing temperature of 72 °C was used for all mutations, in order to avoid any non-specific products. DNA sequencing results confirmed successful introduction of mutations encoding substitutions R822A, Y824A, Y659A and N579A into the gene encoding UnAA-modified Pol. Attempts to introduce mutation encoding substitution I679A consistently failed and occasionally resulted in a primer insertion event, suggesting that the mutagenic primer for this mutation may need to be redesigned. Unfortunately, although most of the plasmids encoding UnAA-modified Pol variants were successfully produced, it was not possible to proceed with the expression, purification and labelling of the variants by the end of the project timeline.

6.5 Discussion

6.5.1 Energy decomposition and selection of substitutions

The energy decomposition method has previously successfully identified stability-determining regions in simple, single-domain proteins [184, 190, 193], but has never been applied to a complex molecular machine such as Pol. However, the ability of EDM to identify the different domains of Pol and to detect the effects of DNA and nucleotide binding on Pol stability (Figure 6.4 and G. Colombo et al., in preparation) suggests that a qualitative description of Pol interaction energetics using this method is possible. In addition, substitutions of residues important for the replication fidelity of Pol (E658 and Y714) have shown energetic signatures that are consistent with crystal structures and single-molecule data (G. Colombo et al., in preparation), indicating that the method is able to detect subtle perturbations of energetic stabilization in Pol.

It is important to note that the specific application of EDM in this project is subject to some caveats. First, EDM has thus far only been used to identify regions of general stabilization in protein structures, but not to predict their role in biasing the protein towards one or the other conformation. The latter approach, which we use in this thesis, is not straightforward because it relies on a direct comparison of the essential interaction matrices of the two states of the protein. This analysis is compromised by the fact that some of the differences between the two matrices may arise from the conditions of structure crystallization and the presence of different ligands in the two structures. In addition, because most of the strongly stabilizing regions will be shared between the two structures, it is necessary to focus on the less strongly stabilizing regions, which is turn are more prone to errors resulting from the simplistic nature of the energy decomposition method.

The second caveat is common to all mutagenesis-based studies, and arises from the observation that even single-residue substitutions can dramatically compromise protein structure and function [188]. Since all of the selected residues used in this study were predicted to stabilize only one or the other conformation of Pol, the effect of their substitutions on the general structural stability of Pol would be expected to be minor. On the other hand,

the proximity of the fingers domain to the nucleotide- and DNA-binding sites in Pol (and the conformational coupling between the two) implies that any substitution affecting the fingers-closing transition may also compromise the ability of Pol to bind the DNA and nucleotide substrates. Such substitutions may therefore complicate the established protocols of Pol sample preparation and the assays used to measure Pol conformational equilibrium. Many of these issues have indeed been observed here, for example prompting us to develop new labelling strategies with improved specificities, which were not needed for characterization of the wild-type protein.

Finally, it should be noted that the stability-determining residues were predicted based on the computational analysis of the *Bst* polymerase, and may not necessarily be relevant in the *E. coli* protein, which was the one tested experimentally. To avoid this issue, we focused only on the residues that showed both sequence and structural conservation in the two proteins; however, it is possible that EDM analysis of the *E. coli* protein would predict different hot-spots of stabilization. Unfortunately, crystal structures of the open and closed conformations of *E. coli* Pol are not available, making it difficult to perform the same type of EDM analysis on this homologue.

6.5.2 Double-cysteine Pol variants

In order to investigate the effect of selected substitutions on the conformational equlibrium of Pol, we initially carried out standard double-Cys labelling of Pol variants in the absence of DNA. We found the efficiency of Atto647N labelling to be significantly lower for the variants than for the wild-type protein, and the chymotrypsin digestion assay suggested that the labelling bias of the variants was almost the opposite of that of the wild-type Pol, with G550/R744 being the major species. Although there is a degree of uncertainty in the interpretation of the chymotrypsin assay, due to potential misassignment of the bands and the effect of substitutions on the mobility of labelled fragments, it is likely that the labelling bias of Pol variants is considerably different than that of the wild-type Pol.

Double-cysteine labelling in the presence of DNA, previously shown to increase the

labelling bias towards species R550/G744 to 100 % [5], also improved the bias for the variants. However, the bias was still significantly lower for the variants compared to the wild-type Pol, and the labelling efficiencies were low for both the variants and the wild-type protein. The reasons for the poor labelling of the wild-type protein remain unclear, and should be investigated further. Since the labelling protocol has previously been successfully used in the laboratory, the low efficiencies could be due to protein or dye degradation, and care should be taken that newly prepared samples are used in the next labelling attempt. Interestingly, the variants showed aberrant behaviour in terms of the binding to the heparin column, suggesting that their DNA-binding properties are different from those of the wild-type Pol, presumably also affecting the ability of the DNA ligand to promote biased labelling in these proteins.

The low labelling efficiency of Pol introduces a large number of coincidence events into confocal traces that have to be distinguished from the FRET populations, which can usually be done using single-molecule sorting approaches. However, the different labelling bias of the wild-type Pol and its variants is inherently problematic. This conclusion arises from the previous observation that species G550/R744 can produce an artefactually closed conformation in the unliganded, binary and ternary complexes [13]. The effect most likely arises from the structural similarity of Atto647N to a dNTP, which implies that the dye can induce a fingers-closing transition when in the vicinity of the fingers domain. Therefore, in addition to any *direct* effect of the substitutions on the conformational equilibrium of Pol, there will also be an *indirect* effect, whereby the substitutions affect the labelling bias, which in turn affects the degree of fingers-closing. Since we are only interested in the direct effect, and it is not possible to deconvolve the two effects from each other, it is necessary to achieve a close-to-100 % labelling bias towards the R550/G744 species for both the wild-type Pol and its variants. We reasoned that this would only be possible if we used orthogonal chemistries to label the two sites.

Since labelling the wild-type protein in the closed conformation (i.e., in the presence of DNA and the correct dNTP) results in an increased bias towards species R550/G744, the apparent preference of the variants to produce species G550/R744 could mean that they

favour the fingers-open conformation. In contrast, the confocal results on Pol variants (labelled in the absence of DNA) show them to be biased towards the closed conformation in the majority of binding states, potentially as a result of the additive effects of an intrinsic fingers-closing bias and the bias induced by the labelling artefacts. This interpretation would confirm the predictions of the EDM analysis that residues R822 and Y824 contribute to the energetic stabilization of the open conformation of Pol. However, the lack of further fingers-closing in the variants in the ternary-correct state suggests that the observed shift in the equilibrium may be more due to an insensitivity to the DNA and nucleotide ligands than due to an intrinsic fingers-closing bias. The apparent compromised binding of Pol variants to the heparin column would also be consistent with the latter interpretation. On the other hand, confocal analysis of the Pol samples labelled in the presence of DNA shows a significant population of high-FRET species, which could be due to labelling issues or protein aggregation. Based on these results, it is clear that the investigated substitutions affect the structure and/or dynamics of Pol in some way, but it remains difficult to make conclusive statements about their effect on Pol conformational equilibrium.

6.5.3 Unnatural amino acid-modified Pol variants

Encouragingly, we have managed to express unnatural amino acid-modified Pol in *E. coli* C321 cells, obtaining similar expression yields as with the wild-type protein and avoiding the issue of premature termination during translation. The labelling of the UnAA-modified Pol with DBCO-Cy5 was also generally successful, although the labelling efficiencies plateaued at 50-55 %, and could not be increased further by using a larger excess of the dye, a higher temperature or a longer incubation time. This effect has been documented before [203], and it has been suggested that the local surface microenvironment of the protein could be a critical factor in determining the efficiency of the cycloaddition reaction. Specifically, hydrophobic patches on the protein could facilitate docking of the hydrophobic cyclooctyne group of DBCO, promoting a proximity effect and the click reaction between the azide and the alkyne. Regardless, a 50 % labelling efficiency at the 550

position would be sufficient for confocal analysis, particularly provided a higher efficiency of Cy3b-maleimide labelling at the 744 position.

The results of the chymotrypsin assay suggest that the preference of DBCO-Cy5 for labelling the azide group over the sulfide group is high. Assuming that Cy3b-maleimide showed a similar preference for the sulfide group over the azide group, the resulting proportion of the R550/G744 species would be expected to be ~100 %, hence eliminating the issue of labelling bias. Unfortunately, the doubly-labelled Pol-Cy3b/Cy5 sample was too dilute to allow estimation of Cy3b labelling bias using the chymotrypsin assay. However, Cy3b labelling of Cy5-labelled UnAA-modified Pol produced good labelling efficiencies, and eliminated the occurrence of coincidence events observed with double Cys-labelled Pol. This is likely due to the orthogonality of the reactions, which ensures a higher percentage of doubly-labelled species than maleimide-only labelling. Curiously, the absorption spectrum of Cy3b/Cy5-labelled Pol indicated an unusual shoulder for Cy3b, and the preliminary confocal analysis showed broad FRET histograms at high FRET values. Additional sample characterization will be required to analyse its labelling profile and its oligomeric state, and the labelling protocol may have to be further optimized to enable a reliable confocal analysis of the variants.

6.6 Future work

The future development of the project appears well defined. Wild-type Pol expression, purification and double labelling (with Cy5 and Cy3b) will be carried out in a continuous fashion, to prevent sample degradation and dilution between the purification and labelling steps. Next, labelling will be repeated with Atto647N and Cy3b, using the protocol optimized for Cy5 and Cy3b, because the same functional groups will be used in the reaction. Finally, the UnAA-modified Pol variants that have been successfully generated at the level of plasmid DNA will be expressed, purified and doubly labelled using the same protocol, and their conformational equilibrium assessed. If the predictions of the EDM analysis are confirmed, then EDM could be trusted as a method for studying the stabilization determ-

inants in Pol. In addition, these results would provide a general demonstration of the EDM approach as a cost-efficient means of predicting structural stabilization in complex multi-domain proteins, a feature of great interest to the field of protein design, particularly in terms of therapeutic implications.

The computational aspects of the project could be extended in several ways. Energy decomposition analysis could be performed on Pol variants with *in silico* substitutions, allowing their effect on Pol interaction energetics to be tested. The effect of substitutions on the conformational ensemble populated by Pol could also be observed directly from the trajectories, since it is now becoming possible to observe full fingers-opening events in Pol using unbiased MD simulations (Section 5.1.3 and reference [36]). Notably, the sampling would not be sufficient to measure the conformational equilibrium of Pol in this way, but it could allow us to distinguish the behaviour of Pol variants from that of the the wild-type protein, and to begin to understand the long-range molecular effects of the destabilizing substitutions. Coarse-grained models and biased MD approaches could also be used to extend the time scales of the simulations and enhance conformational sampling.

In addition to the relevance for this project, our ability to express, purify and label unnatural amino acid-modified Pol presents a significant contribution towards all studies relying on orthogonal labelling of Pol. The conformational dynamics of wild-type Pol could now be studied more accurately in TIRF-based assays, without the interference of artefactual data resulting from the presence of species G550/R744. Further, being able to specifically label Pol and full-length Pol at any position significantly expands the possibilities of structure determination by single-molecule FRET that were presented in Chapter 4. Unnatural amino acid-based labelling of Pol would also be an invaluable approach for studying the protein in the context of the living cell, since it may allow in-cell labelling and avoid the need for protein internalization, as we discuss in Chapters 7 and 8.

6.7 Materials and methods

6.7.1 MD simulations and energy decomposition

All simulations were performed with the AMBER 12 suite of programs [89] and the ff12sb force field [208]. *Bst* crystal structures with PDB codes 1L3U and 1LV5 [8] were used for the open- and closed-state coordinates, respectively. Solvent was modelled explicitly, using the TIP3P water model, and a square simulation box was used, with a minimum 12-Å solvent edge. All Mg^{2+} ions and water molecules present in the crystal structures were retained for the simulations. Energy minimization was performed in two steps: first, a restrained minimization of 4000 steps was done, keeping the backbone protein, DNA and magnesium atoms fixed with a force constant of 100 kcal/mol. Next, unrestrained minimization was done, consisting of 2000 steps of steepest-descent minimization and 2000 steps of conjugate-gradient minimization. Equilibration was performed for 2.5 ns at 300 K, and was used to estimate the mean total energy and define the boost parameters, as described [209]. A tuning parameter α of 0.20 was used to modulate the depth and local roughness of basins in the modified potential. Production runs were carried out in the NPT ensemble, at 300 K temperature and 1 atm pressure, with particle-mesh Ewald electrostatics and a 10-Å cut-off for van der Waals interactions. All covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm, and a time step of 2 fs was used.

Energy decomposition analysis was performed using the resulting open- and closed-state trajectories, with the AMBER suite of programs, as described in Section 6.1.2 and in reference [189]. To account for the complex and modular structure of Pol, 7 eigenvectors were used in the construction of essential energy matrices.

6.7.2 Local flexibility analysis

Local flexibility analysis [210] was carried out by first computing a matrix of distance fluctuations A_{ij}, as follows:

$$A_{ij} = \langle (d_{ij} - \langle d_{ij} \rangle)^2 \rangle \tag{6.3}$$

where d_{ij} is the distance between the C_α atoms of residues i and j at a specific time point in the trajectory, and $\langle d_{ij} \rangle$ is the time average of the same distance across the trajectory. The flexibility parameter p of each residue i was then calculated relative to its neighbouring residues running from $i-2$ to $i+2$, and an additional sigmoid function was included to restrict the sum to residues within 5 Å:

$$p(i) = \sum_{j=i-2}^{i+2} A_{ij} \frac{1 - tanh(\langle d_{ij} \rangle - 5)}{2} \tag{6.4}$$

Thus, highly flexible residues show significant time-dependent variation in terms of their distances to neighbouring residues, resulting in high values of the local flexibility parameter $p(i)$.

6.7.3 *Bst* and *E. coli* Pol alignment

Sequence alignment of *Bst* and *E. coli* polymerases was done in Clustal X [211], using FASTA sequences of PDB entries 1L3U and 1KLN [8]. Structure alignment was carried out in PyMol, using the same PDBs and commands 'align' (sequence-dependent alignment) and 'super' (sequence-independent alignment).

6.7.4 Single-point mutagenesis

As starting constructs for mutagenesis, we used either a pAS1-based plasmid [212] carrying N-His_6-tagged Pol (KF) gene with mutations encoding D424, K550C, L744C, C907S, or a pBAD24-based plasmid [207] carrying the same gene with the same mutations except for K550(UAG). Mutagenic primers were designed using QuikChange Primer Design online tool [213] and were synthesised commercially (IBA). Mutagenesis reactions were performed in 20 µl volumes, using 1x Pfu buffer, 0.2 mM dNTP mix, 0.5 ng/µl plasmid DNA, 0.5 µM top primer, 0.5 µM bottom primer and 0.5 µl of the commercially supplied PfuTurbo DNA poymerase (Agilent Technologies). The PCR cycle involved an initial denaturation step at 95 °C for 2 min, followed by 30 cycles of denaturation (95 °C; 30 sec), annealing (55 °C or 72 °C; 30 sec) and extension (72 °C; 10 min). Reactions were kept at

4 °C, until 6 µl of each reaction was incubated with 1 or 2 µl of DpnI enzyme (New England BioLabs) at 37 °C for 3 hours. Following digestion of methylated DNA, DpnI was inactivated by 20-min incubation at 80 °C.

All of the DnpI-treated PCR mix was used to transform 100 µl of XL-1 Blue Supercompetent cells, by 30-min incubation on ice, 45-s heat shock at 42 °C, and 2-min incubation on ice. The cells were plated out on LB agar plates containing 50 µg/ml carbenicillin, and the plates incubated at 37 °C over-night. Successful transformants were inoculated in 5 ml LB (50 µg/ml carbenicillin) and incubated at 220 rpm and 37 °C over-night. The cells were spun in a swinging-bucket rotor centrifuge (Beckman GS-6R, rotor GH 3.8) for 5 min, at 4200 rpm and 22 °C. DNA isolation was performed using QIAprep Spin Miniprep Kit (Quiagen), according to manufacturer's instructions. Samples were eluted in 50 µl of the 'EB' buffer, and their concentrations and 260/280 nm absorption ratios measured using Nanodrop. Samples were stored at -20 °C until they were used.

For DNA sequencing, 5 µl of 100 ng/µl DNA was prepared per reaction, along with 5 µl of each of the sequencing primers at 4 µM concentration. Primers used for double-Cys Pol variant genes were previously designed. The UnAA-modified Pol gene was sequenced using commercial primers pBADF and pBADR, and a new primer (MS23) was designed to cover the central region of the gene. Sanger sequencing was performed commercially (Source BioScience), and the resulting chromatograms and sequences were analysed in SnapGene (GSL Biotech LLC).

6.7.5 Competent-cell preparation

Chemicompetent C321 cells were prepared by streaking cell stock onto an LB agar plate, and inoculating a colony in 3 ml LB, in the absence of antibiotic. The cells were grown over-night at 37 °C, transferred to 200 ml LB and grown at the same temperature until an OD of 0.5. Following 30-min incubation on ice, the cells were spun down in a swinging-bucket rotor centrifuge (GS-6R, Beckman) for 15 min at 3000 rpm and 4 °C, and then resuspended in 8 ml of cold 100 mM $CaCl_2$ and 15 % glycerol. The cells were then incubated on ice for

20 min, spun down as before, and resuspended in 4 ml of the same buffer, before they were split into aliquots, flash-frozen and stored at -80 °C.

6.7.6 Expression and purification of double-Cys Pol

Double-Cys Pol variants were expressed using the pET expression system, whereby the gene of interest is put under the control of a T7 promoter, and an IPTG-inducible copy of the T7 RNAP gene is inserted into the host genome. Therefore, pET plasmids carrying Pol genes were transformed into HMS174 (DE3) cells, and single colonies inoculated in 25 ml LB, supplemented with 50 μg/ml carbenicillin. The cultures were grown over-night at 220 rpm and 37 °C, and then used to inoculate 1 litre of LB, supplemented with carbenicillin. The cultures were grown to an OD_{600} of 0.6, at which point expression was induced with 0.5 mM IPTG. After 2 hours of expression, the cells were harvested by spinning the cultures in a swinging-bucket rotor centrifuge (GS-6R, Beckman) for 20 min at 3000 rpm and 4 °C. The cells were then resuspended in cold 50 mM Tris pH 7.5, and spun down in an ultracentrifuge (Sigma 3K30, rotor 12150-H) for 15 min at 8,000 rpm at 4 °C. Finally, the pellet was resuspended in lysis buffer, containing 50 mM Tris pH 7.2, 300 mM NaCl, 1 mM β-mercaptoethanol, 10 mM imidazole, 2 mg/ml lysozyme and 0.02 mM phenylmethylsulfonyl fluoride (PMSF). The cells were stored in the lysis buffer over-night at -80°C.

The frozen cells were thawed, and a further 25 μl of PMSF added to prevent any protein digestion. Sonication was then performed, with 6 cycles of 5-second 'on' and 10-second 'off' time, and the cell debris spun down for 20 min at 15,000 rpm and 4 °C (Sigma 3K30). 3 ml of Ni-NTA resin (Quiagen) per gram of cells was spun in a swinging bucket rotor centrifuge (GS-6R, Beckman) for 5 min at 4,000 rpm and 4 °C, and then washed 3 times with Ni-NTA buffer, consisting of 50 mM Tris pH 7.2, 300 mM NaCl, 1 mM β-mercaptoethanol and 10 mM imidazole. The supernatant containing the cell lysate was combined with the washed Ni-NTA resin, and the protein allowed to batch-bind for 1 hour on a rotating wheel at 4 °C. The resin was spun down, resuspended in the Ni-NTA buffer, applied to a plastic column and washed with 10 column volumes (CV) each of Ni-NTA buffers containing 10

mM, 20 mM and 27 mM imidazole. The protein was eluted with 5 CV of Ni-NTA buffer containing 100 mM imidazole, and the fractions analysed using Nanodrop and SDS-PAGE. The concentrated fractions were pooled, and then dialysed into 50 mM Tris pH 7.4, 1 mM DTT, over-night at 4 °C. The dialysed samples were combined in a 1:1 ratio with 2x glycerol storage buffer (80 % glycerol, 50 mM Tris pH 7.5, 2 mM DTT) and stored at -20 °C.

6.7.7 Expression and purification of UnAA-modified Pol

The following genes were synthesised commerically and cloned into the pBAD24 vector: UnAA-modified Pol (D424A, K550(UAG), L744C, C907S), Pol Cys-less (D424A, C907S) and full-length Pol Cys-less (D424A, C907S). All samples were dissolved in TE buffer (10 mM Tris pH 8.0, 1 mM EDTA) to 100 ng/µl concentration. The plasmids were transformed into XL-1 Blue Supercompetent cells, replicated and isolated, to make them suitable for expression in C321 cells. For Pol and flPol Cys-less expression trials, HMS and C321 cells were transformed with Pol and flPol plasmids, and used to inoculate 5-ml cultures of either the auto-induction medium (Overnight Express, Novagen, supplemented with 50 µg/ml carbenicillin and 0.05 % arabinose) or LB (with 50 µg/ml carbenicillin). In the case of manual induction, expression was triggered with 0.05 % arabinose at an OD of 0.5, and expression allowed to proceed for 3 hours.

Expression of the UnAA-modified Pol was tested by co-transforming C321 cells with pBAD24-Pol and either pEVOL-AzF or pDULE2-AzF, and plating them out on plates containing 50 µg/ml carbenicillin and either 35 µg/ml chloramphenicol (pEVOL) or 50 µg/ml spectinomycin (pDULE2). The autoinduction medium was supplemented with the antibiotics, 0.2 % arabinose and 1.03 mg of the unnatural amino acid azidophenylalanine; the latter was added 40 min after innoculation. For manual induction, over-night cultures were prepared in LB medium supplemented with the antibiotics. Following inoculation in new media and re-growth, induction was triggered at an OD of 0.5 with 0.2 % arabinose. In this case, the unnatural amino acid was added 30 min before induction.

Large-scale expression of UnAA-modified Pol was carried out in the same way as in

the expression trials, except that 25-ml over-night cultures were grown, and 1-litre cultures were used for expression. Cells were resuspended in 50 mM Tris pH 7.2, 300 mM NaCl, 10 mM imidazole, 2 mg/ml lysozyme and 1 cOmplete Protease Inhibitor Cocktail tablet (Roche) per 10 ml buffer. His-tag purification was carried out as before, but with no β-mercaptoethanol present in the purification buffers.

6.7.8 Double-cysteine labelling

Purified Pol samples were reduced with 5 mM DTT for 1 hour at room temperature, transferred into 10K Slide-A-Lyzer dialysis cassettes (Life Technologies), and dialysed into 1 litre of fresh 1 mM tris(2-carboxyethyl)phosphine (TCEP) over-night at 4 °C. Samples were further dialysed into 1 litre of 120 μM TCEP for 3x 1 hour at 4 °C, removed from the cassettes and their concentration measured on Nanodrop. For biased labelling, a primer-template DNA was used with dideoxy-terminated 3' end and cytosine as the templating base. The DNA was annealed by combining the primer and template strands at 127 μM and 160 μM concentrations, respectively, in 50 mM Tris pH 7.4 and 100 mM $CaCl_2$. For each labelling reaction, 10 nmol protein was combined with >10 nmol annealed DNA, in a buffer consisting of 5 mM $CaCl_2$, 1 mM EDTA and 1 mM dGTP (the correct nucleotide). Binding of the primer-DNA to Pol was allowed for 10 min at room temperature.

Atto647N-maleimide (Atto-tec) was dissolved in dimethly sulfoxide (DMSO) and added to each protein (+DNA) sample, at 12 nmol per reaction. Labelling was carried out in 200-μl reactions for 1 hour, at 300 rpm and 22 °C, until DMSO-dissolved Cy3b-maleimide (GE Healthcare) was added at 34 nmol per reaction. The second labelling step was carried out over-night on a rotating wheel at 4 °C, and was quenched by adding 1 mM DTT. Unreacted dye was removed on a heparin column, which was set up in a Pasteur pipette with ~300 μl of heparin agarose resin per 10 nmol protein. The resin was washed with water, and then equilibrated with 10 CV of heparin buffer, containing 20 mM Tris pH 7.4, 1 mM EDTA, 2 % glycerol and 1mM β-mercaptoethanol. The protein-dye mix was loaded, and the column washed with 10 CV of the heparin buffer, and finally with 10 CV of heparin

buffer containing 50 mM NaCl. The protein was eluted using heparin buffer containing 400 mM NaCl.

Alternatively, unreacted dye was removed on a Ni-NTA column, following 30-min batch-binding of the protein to 100 µl of resin. The column was washed with 50 CV of Ni-NTA buffer, containing 50 mM Tris pH 7.4, 25 mM NaCl and 10 mM imidazole. Further washing was done with 10 CV of Ni-NTA buffer containing 500 mM NaCl, to remove any remaining DNA bound to the protein, followed by 10 CV of the low-salt Ni-NTA buffer, to prevent elution in high salt. The protein was eluted in Ni-NTA buffer containing 200 mM imidazole. Samples were dialysed first into 1 litre of 50 mM Tris pH 7.4, 25 mM NaCl, 1 mM DTT, for 3x 1 hour, and then into 500 ml of the same buffer containing 40 % glycerol, over-night. The glycerol-dialysed samples were stored at 4 °C.

All fractions were analysed on Nanodrop, to measure the absorbance at 280 nm, 570 nm and 669 nm, and hence the concentrations of Pol, Cy3b and Atto647N. Concentrations were calculated assuming extinction coefficients for Pol (KF) of 58,790 $M^{-1}cm^2$ at 280 nm, for Cy3b of 130,000 $M^{-1}cm^2$ at 669 nM [214] and for Atto647N of 150,000 $M^{-1}cm^2$ at 570 nm [144], and assuming A_{280}/A_{max} correction factors of 0.14 and 0.05 for Cy3b and Atto647, which were measured experimentally using individual dyes. Labelling efficiencies were estimated from the corrected concentrations of Pol, Cy3b and Atto647N.

6.7.9 Azide/cysteine labelling

UnAA-modified Pol was labelled by combining 5 nmol protein with 50 (or 100) nmol of DMSO-dissolved DBCO-Cy5, in 100-µl reactions containing 50 mM Tris pH 7.2 and 300 mM NaCl. The reactions were protected from light and incubated for 24 or 48 hours, at 4 °C, 25 °C or 37 °C. Unreacted dye was removed by applying each sample to an equilibrated Micro Bio-Spin 6 column (Bio-Rad), centrifuging at 1000x *g* for 4 min, collecting the eluate and repeating the process on a new column. The samples were analysed using Nanodrop or a UV spectrophotometer (Cary 50 Bio, Varian), and absorbance at 280 nm and 649 nm measured. Concentrations and labelling efficiencies were calculated assum-

ing the extinction coefficient for Cy5 of 250,000 $M^{-1}cm^2$ at 649 nm [214] and the A_{280}/A_{max} correction factor of 0.05.

Cy5-labelled UnAA-modified Pol sample was quenched with 5 mM DTT for 1 hour at room temperature, and dialysed into 1 mM TCEP and then 120 µM TCEP, as before. The dialysed sample was incubated with 20-fold molar excess of DMSO-dissolved Cy3b-maleimide over-night, on a rotating wheel at 4 °C. Unreacted dye was removed using a heparin column, and labelling efficiencies estimated as before.

6.7.10 Chymotrypsin digestion assay

25 µg of bovine chymotrypsin (Sigma-Aldrich) was dissolved in 50 µl of 1 mM HCl, 2 mM $CaCl_2$ and stored in aliquots at -80 °C. Reaction buffer was prepared as 100 mM Tris pH 8, 10 mM $CaCl_2$, and the quench buffer as 25 mM EDTA pH 8, 2 % SDS and 2 mM PMSF. For each reaction, 1 µl of 20-30 µM protein was added to 10 µl reaction buffer, 1 µl 1 % SDS and 1 µl 10 mg/ml BSA, before 1 µl of chymotrypsin was added at 10x dilution. After a 40-second incubation, the reactions were quenched with 33 µl quench buffer, and combined with 47 µl of 2x transparent protein loading buffer, consisting of 250 mM Tris pH 6.8, 2 % SDS, 20 % glycerol and 1 mM DTT. The reactions were stored at -20°C before they were analysed. In addition to the samples of interest, singly-labelled Pol 550-Cy3b, 550-Atto647N, 744-Cy3b and 744-Atto647N samples that were previously prepared were also digested, along with a 1:1 mix of Pol 550-Cy3b and 744-Atto647N, and a 1:1 mix of Pol 550-Atto647 and 744-Cy3b.

The samples were heated for 5 min at 95 °C before being loaded on an SDS-PAGE gel. Mini gels were bought precast (4-20 % acrylamide, Mini-PROTEAN TGX, Bio-Rad) and were run in Tris-glycine SDS at 300 V for 15 min. Midi and maxi gels were poured manually to make a 10 % acrylamide resolving layer and a 4 % acrylamide stacking layer, and were run at 200 V for 2.5 hours (midi), or at 20-40 mA for 6 hours (maxi). In-gel fluorescence was recorded in the green and red channels, using appropriate filters (Pharos FX Plus Molecular Imager, Bio-Rad). Specific bands were selected that were observed only

in one of the Pol 550 or Pol 744 singly-labelled controls, and their intensities quantified in Fiji. The intensities were corrected for the different detection efficiencies in the green and red channels, and the different labelling efficiencies of Cy3b and Atto647N, which allowed calculation of the relative amount of each species (550-Cy3b, 550-Atto647N, 744-Cy3b, 744-Atto647N), and hence the ratio of the two labelling orientations (R550/G744 and G550/R744) in each sample.

6.7.11 Confocal microscopy

Doubly-labelled Pol samples were diluted to 50-200 pM concentration in Pol buffer (4.8.2). Binary complexes were formed by adding 100 nM of CJ281 primer-template DNA, a stem-loop DNA with dideoxy-terminated 3' end and adenine as the templating base. Ternary complexes were prepared using the same DNA and 1 mM dTTP (correct nucleotide) or 1 mM dGTP (incorrect nucleotide). Single-molecule measurements were performed as in Section 4.8.2; 6-12 datasets of 10 min were recorded for each sample, and E/S histograms extracted as before. The lower-FRET population of the ternary-incorrect complex of wild-type Pol was fitted to obtain the FRET efficiency and sigma values of the partially closed state, which were in turn used to obtain the parameters of the closed state. The lower-FRET population of the binary complex of wild-type Pol was then fitted to obtain the parameters of the open state, and the remaining FRET populations in other E/S histograms restrain-fitted with Gaussian functions using these parameters. The relative proportions of the species were calculated from the areas of the FRET populations in the histograms.

6.8 Contributions

- Accelerated MD simulations, EDM and local flexibility analyses, and initial selection of *Bst* substitutions were done by Meli Massimilliano and Giorgio Colombo.
- Sequencing primers for genes encoding double-Cys Pol variants were designed by members of Cathy Joyce's group.
- DNA sequencing was done commercially.
- The chymotrypsin digestion assay was previously optimized by Cathy Joyce. Singly labelled Pol variants used as reference samples in the chymotrypsin assay were also purified and labelled by members of her group.
- Doubly labelled wild-type Pol sample used as a reference for optimal labelling was previously prepared by Tim Craggs and Johannes Hohlbein.
- Design of the gene encoding the unnatural amino acid-modified Pol, and the optimization of the unnatural amino acid labelling were done by David Bauer.
- Vector pBAD24 was provided by Pawel Zawadzki.
- The gene encoding the UnAA-modified Pol was synthesised and cloned into vector pBAD24 commercially.
- Cell strain C321 and plasmid pEVOL were provided by Nicolae Solcan.
- Initial stocks of C321 chemicompetent cells were provided by Florence Wagner.
- Plasmid pDULE2 was provided by David Bauer.

7 Cell internalization of Pol

7.1 Introduction

7.1.1 Project rationale

So far, we have used a reductionist approach in studying Pol, which has involved isolating the protein from the living cell, and investigating it *in vitro*. Reductionist approaches have been the most common means of studying biomolecular structure and function, as they allow the control of experimental conditions and can deconvolve complex biological systems into separate components and effects. However, *in vitro* studies can show behaviours and mechanisms that do not apply in native environments and under physiological conditions [14]. For example, the high density and viscosity of the cytosol, and the associated crowding effects, imply that the rates and equilibria of macromolecular interactions will be severely affected [215, 216]. In addition, many cellular reactions occur under non-equilibrium conditions, with a constant supply of reactants and free energy. The spatial organization of the cell, regulatory interactions and cellular feedback networks further complicate the interpretation of *in vitro* studies.

With these considerations in mind, we set to study Pol in the context of the living cell, at the single-molecule level. We aimed to establish single-molecule tracking and single-

molecule FRET imaging of Pol, which would ultimately open the door to a plethora of studies. First, these capabilities would enable us to test the *in vitro* structure of Pol in cells, including the arrangement of the fingers, thumb and palm subdomains. The same approach could be extended to the structure of full-length Pol, to probe the relative position of the 5' nuclease domain, and determine which (if any) of the two presented crystal structures is physiologically relevant. In addition, we aimed to detect changes in the localization and conformation of Pol upon an external trigger, such as addition of a DNA-damaging agent, and to measure the DNA search and binding times. Finally, it would be of great interest to measure the conformational dynamics of Pol in live cells during DNA repair, hence probing enzyme activity in real time under physiological conditions.

The significant technical challenges underlying these project aims require prior optimization of the sample preparation procedures, and the cell-delivery and cell-imaging methodologies. In this chapter, we optimize the cell internalization of organically labelled Pol and flPol, and investigate their diffusion in cells by means of single-molecule tracking. With these capabilities established, we move on to studying Pol by means of single-molecule FRET in Chapter 8. We begin with a brief introduction to single-molecule detection in cells, particularly in terms of the technical requirements involved.

7.1.2 Single-molecule detection in cells

In Chapter 3, we discussed the ability of single-molecule methods to unravel both the static and dynamic heterogeneity in populations of molecules. Molecular heterogeneity is particularly prominent in the complex and ever-changing environment of the living cell (Figure 3.3). Many cellular processes, such as the activation and deactivation of gene expression, also exhibit stochastic reaction events, and cannot be accurately synchronized across different molecules or cells [14]. In addition, some macromolecules exist in low copy numbers: a particular gene exists in only one copy, and some important proteins such as DNA polymerases or transcription factors are present in low numbers, necessitating detection at the single molecule level [55]. Over the last few decades, single-molecule detection in cells has

allowed the monitoring of interactions, copy numbers and diffusion patterns of proteins in processes ranging from DNA replication and transcription to protein translation and membrane transport [53–56].

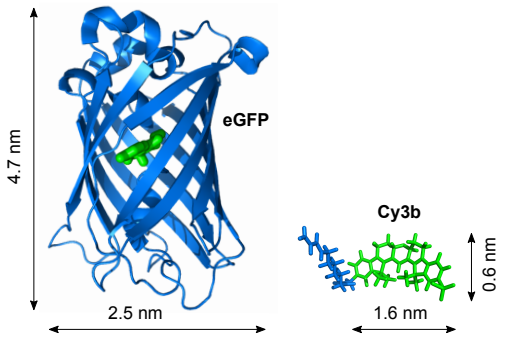
Fluorophores

Single-molecule imaging in cells has traditionally been performed with fluorescent proteins (FPs), derivatives of the green fluorescent protein (GFP) [217]. FPs are β-barrelled proteins, hosting a tripeptide-based fluorophore in their core that forms as a result of an autocatalytic cyclization reaction. The ability to genetically encode FPs as protein fusions has enabled highly specific and efficient (close to 100 %) labelling of almost any protein of interest in live cells. A number of FP variants are available with different spectroscopic properties [218], including photoactivatable proteins, which can be converted from a non-fluorescent to a fluorescent state upon irradiation with low-wavelength light. The latter have allowed super-resolution imaging in cells, by means of photo-activated localization microscopy (PALM) [219].

However, FPs fail to meet a number of criteria important for efficient single-molecule detection in cells (Figure 7.1). First, they have poor photostability, which limits the possibility of imaging proteins at biologically relevant time scales, which are on the order of seconds or minutes. Second, they are characterized by relatively low brightness, often preventing detection against the high level of cellular background fluorescence, known as autofluorescence [220, 221]. Whilst low brightness can be compensated for by using high excitation powers, these in turn limit the survival of the fluorophore and can also be toxic for the living cell [222]. Finally, FPs are large fusions and may interfere with the structure, dynamics and activity of the protein of interest.

Organic fluorophores, such as Cy3(b), Cy5 and the Atto and Alexa dye series, used traditionally for *in vitro* experiments, are smaller, brighter and more photostable than FPs (Figure 7.1) [218, 223–225]. Organic-dye labelling has been employed for a variety of cellular applications, ranging from stochastic optical reconstruction microscopy (STORM) [226] to single-molecule FRET. Whilst a range of dyes are available, covering the

whole visible spectrum, dyes emitting at the red end of the spectrum are preferred, because autofluorescence generally decreases with increased wavelength [220], and hence a better signal-to-noise ratio can be achieved. Finally, some organic fluorophores are reasonably hydrophobic [227], and should be avoided if an exact account of intracellular dynamics is required.



	fluorescent proteins	organic fluorophores
size[a]	~20 nm^3	< 1 nm^3
brightness[b]	3 – 60 (up to 95)	10 – 110
photostability[c]	sub-second	second-minute
rotational freedom	low	medium-high
labeling position	N- or C-terminal	any surface-exposed residue
labeling environment	*in vivo*	*in vitro*
internalization	not required	required

Figure 7.1: *Comparison of fluorescent proteins and organic fluorophores, in terms of properties relevant for single-molecule detection in cells.*[1] *Example fluorescent protein (eGFP, PDB code 2YOG, reference [228]) and organic fluorophore (Cy3b) are shown and sized to scale. The fluorescent component of each molecule is depicted in green; the protein backbone of eGFP and the dye linker of Cy3b are in blue.* [a] *Volume estimated from typical fluorophore dimensions (eGFP; Cy3b and Atto647N).* [b] *An approximation based on the product of typical extinction coefficients and quantum yields [218, 223, 225].* [c] *Approximate survival time under focused laser illumination used for single-molecule tracking, such as in references [15, 229].*

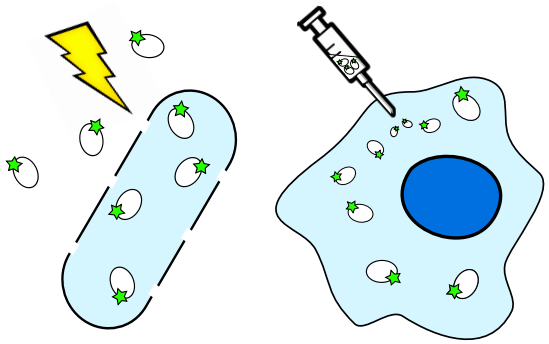
[1] The table includes fluorophore properties relevant specifically for in-cell smFRET, which we discuss in Chapter 8.

Cell internalization

Whereas FPs can be encoded genetically, organic fluorophores are not easily amenable to endogenous labelling, and hence labelling is performed *in vitro*, such as via thiol or amine modifications. An alternative approach relies on using polypeptide tags to which organic fluorophores can be attached (such as SNAP-tag or HaloTag [230]), but these are generally large fusions that may interfere with biomolecular dynamics. *In vitro* labelled biomolecules can be internalized into cells by several means. In the case of microbial cells, heat shock has been used to introduce organic fluorophore-labelled DNAs into *E. coli* [231], whereas electroporation has been shown to work for internalization of labelled DNAs and proteins into both bacteria and yeast [229].

Electroporation is a highly versatile method and has been used traditionally for transformation of prokaryotic and eukaryotic cells with plasmid DNA [232]. Fluorescently-labelled biomolecules can be internalized by incubating them with electrocompetent cells under conditions of low ionic strength, followed by the application of a high-voltage electric field (Figure 7.2). This results in the formation of transient pores in the cell membrane that allow the labelled biomolecules to be internalized, and the cells are thoroughly washed to remove any non-internalized molecules [229]. Many cells ($> 10^8$ in the case of bacteria) can be electroporated simultaneously, allowing high-throughput imaging. However, the method is restricted to small and medium-sized monomeric biomolecules, due to the limited size of the membrane pores. As has previously been indicated [229], and as we further show in this thesis, the concentration of the incubated biomolecule and the electroporation voltage can both be varied to achieve the desired loading, ranging from a few molecules (suitable for single-molecule experiments) to several hundreds (suitable for single-cell imaging). Similarly, the electroporation voltage affects cell viability, with high viability observed under voltage conditions suitable for single-molecule experiments. The structure and activity of the internalized biomolecules can also be preserved, as demonstrated by observing T7 RNAP-induced EmGFP expression in cells electroporated with unlabelled RNAP [229].

In larger eukaryotic cells, microinjection, syringe loading and scrape loading have been used for internalization [233–235]. Microinjection can provide efficient internalization without the need for cell washing or recovery, and can maintain high cell viability. However, cells have to be microinjected and imaged individually, resulting in a relatively low throughput of ~100 cells per day (Figure 7.2). This approach can be advantageous to electroporation when delivery to a specific cell compartment is needed, or for internalization of large biomolecules and complexes, but it is unfortunately not applicable to microbial cells due to their small size.



	electroporation	**microinjection**
system	most cell types	large cells
compartment delivery[a]	not possible	possible
throughput[b]	very high	~100 cells / day
molecule size	max ~100 kDa	not limited
cell loading	tunable	tunable
cell washing[c]	required	not required
imaging[d]	~30 min delay	immediate
viability[e]	up to 95 %	~75 %

Figure 7.2: *Comparison of electroporation and microinjection as internalization methods for single-molecule detection in cells. The prokaryotic and eukaryotic cells are not drawn to scale.* [a] *Delivery of labelled molecules to a specific cell compartment, e.g. cytoplasm or nucleus.* [b] *Number of cells that can be loaded with the labelled molecule and imaged in a typical experiment. In the case of electroporation, the throughput is only limited by parallel imaging capabilities.* [c] *Removal of non-internalized molecules from cell suspension, after internalization and prior to imaging.* [d] *Delay between internalization and imaging. Microinjection is performed on the microscope set-up and hence almost no delay occurs.* [e] *Proportion of cells that appear intact after internalization. [224, 229].*

In principle, the need for internalization could be avoided if biomolecules were labelled inside cells, such as using unnatural amino acid modifications (Section 6.1.3). Unfortunately, most of the UnAA approaches rely on reactions that proceed slowly and inefficiently under physiological conditions, or that have side reactions in cells [200]. However, recent studies have shown promising results with tetrazine-based cycloaddition reactions, which are fast, specific and can label proteins in live *E. coli* and mammalian cells, with minimal background [236–238]. The challenge now lies mainly in the design of fluorophores that have both exceptional photophysical properties, and are membrane-permeable to enable cell delivery and wash-out. Use of fluorogenic probes, which are natively non-fluorescent but are activated upon conjugation [239, 240], may help in addressing the latter.

Imaging

Single-molecule fluorescence in cells can be imaged using either confocal or total internal reflection fluorescence (TIRF) microscopy (Section 3.1.5). TIRF microscopy is preferred for our applications, as it enables long observations of many molecules at a time, limited only by the photostability of the fluorophore. When working with large cells, highly inclined and laminated optical sheet illumination (HILO) can be used instead of TIRF to increase the depth of illumination whilst maintaining low background fluorescence [241]. With TIRF and HILO imaging, point spread functions (PSFs) arising from single molecules can be fitted and the their centroid positions determined with high precision (~20 nm). Repeated localization of the labelled molecule in successive frames additionally allows single-molecule tracking, with time resolution limited by the frame rate of the camera (~10 ms).

In both confocal and widefield-based imaging approaches, light scattering and autofluorescence properties of the cell type of interest should be considered. Particularly at low wavelengths, autofluorescence can interfere with single-molecule tracking, and it may be difficult to distinguish the molecule of interest from the fluorescence background. In this respect, lifetime-based fluorescence imaging may be preferable to intensity-based imaging in some applications, because autofluorescence exhibits a specific lifetime [242]. Notably, the environment of the cell may also affect the photophysical properties of the dyes. For

example, owing to the reducing nature of the cytosol [243], photobleaching is less pronounced in cells, allowing longer observation times than in the absence of reducing agents *in vitro* [229].

7.2 Protein internalization by electroporation

Whilst the electroporation protocol has previously been used to deliver labelled proteins into *E. coli* [229], it has not been thoroughly characterized for this purpose. In this section, we optimize the protocol for the internalization of small and medium-sized soluble proteins. In particular, we explore the buffer conditions compatible with electroporation, and measure the effect of electroporation voltage on protein internalization efficiency and cell viability. We use the results of these studies as the starting point for the selection of optimal conditions for the internalization of Pol in the second part of the chapter.

7.2.1 Buffer conditions for electroporation

Most proteins have a preference for specific buffer conditions, which maintain their structural stability and activity over time. At the same time, electroporation has to be carried out under conditions that ensure a high efficiency of protein internalization and preserve cell viability. In particular, the ionic strength of the buffer in which the cells are electroporated should be high enough to maintain the integrity of the protein sample but low enough to avoid the occurrence of an electrical short circuit (arcing). The electroporation time constant, which describes the exponential decay of the applied voltage with time, is indicative of whether a certain level of electrical discharge has occurred in the medium. For a standard electroporator (such as MicroPulser Electroporator, Bio-Rad), 4.00 ms can be taken as a conservative, and 3.00 ms as a generous estimate for the lower bound of an acceptable electroporation constant that allows good cell loading and preserves viability. For reference, the electroporation constant of pure deionized water is approximately 6.00 ms.

To determine the highest salt concentration that can be used for protein electroporation, we measured the electroporation constant for buffers containing 50 mM Tris and

between 0 and 150 mM NaCl, diluted 20 times in water to simulate the dilution under conditions of cell electroporation. As anticipated, the time constant decreased with increasing salt concentration (Figure 7.3a), and at higher concentrations there was a greater probability of arcing events. Seeing as the electroporation voltage affects both protein internalization efficiency and cell viability (discussed below), we also tested the effect of voltage on the electroporation time constant, and observed the constant to decrease with increasing voltage. Whilst the effect was minor at low ionic strengths, it was significant at higher ionic strengths (> 60 mM NaCl). We conclude that a working buffer of 50 mM Tris pH 7.4 and up to 50 mM NaCl (or equivalent) is appropriate for successful electroporation at any voltage up to 1.80 kV. Higher ionic-strength buffers may only be used at low voltage settings.

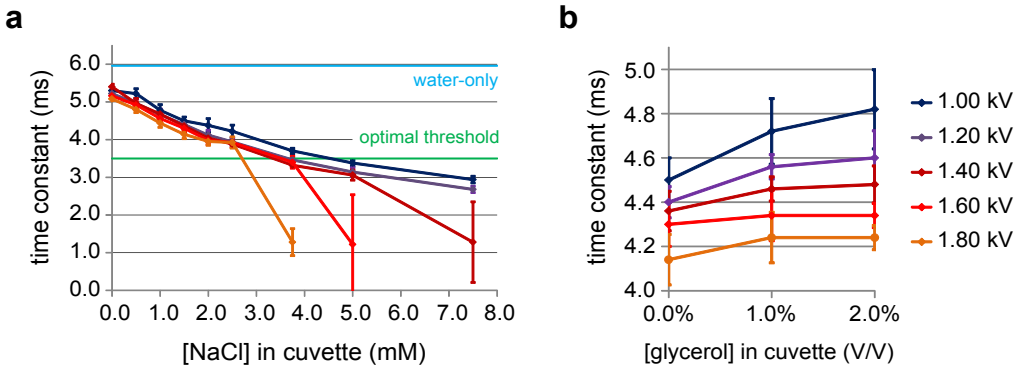


Figure 7.3: *Effect of* ***(a)*** *NaCl and* ***(b)*** *glycerol concentrations in stock buffer on electroporation time constant. NaCl and glycerol concentrations shown refer to the final concentrations present in the electroporation cuvette, and the error bars represent standard deviation over 5 measurements. The mean time constant for electroporation of water at 1.4 kV (6.00 ms) is shown with a blue line, and the optimal time-constant threshold (3.50 ms) is indicated with a green line. Different voltage conditions are coloured as denoted in the legend.*

Finally, we noted that protein storage buffers often contain glycerol to maintain protein integrity and mitigate the effects of sample freezing and thawing. We therefore tested the effect of the concentration of glycerol in the buffer (containing 50 mM Tris pH 7.4, 30 mM NaCl) on the electroporation time constant. A minor but consistent positive effect was observed when glycerol concentration was increased from 0 % to 40 % (Figure 7.3b),

suggesting that storing stock protein samples in a glycerol-based buffer is appropriate for the purposes of electroporation, and is unlikely to compromise cell loading.

7.2.2 Effect of voltage on internalization efficiency

The applied voltage is thought to affect the efficiency of electroporation, in terms of the number and size of membrane pores that are created [244]. In order to test the effect of voltage on protein internalization efficiency, we selected a small protein, the 10-kDa (ω) subunit of bacterial RNA polymerase, and labelled it with Cy3b using maleimide chemistry. SDS-PAGE analysis of the labelled RNAP ω showed two bands: a main band corresponding to the full-length protein, and a secondary band that could correspond to its proteolytic fragment. The concentration of contaminating dye in the sample, estimated from in-gel fluorescence, was 1 % (Figure 7.4a). With a starting concentration of RNAP ω of 2.5 μM, the protein was internalized at high efficiency, whilst both the non-electroporated and empty-cell controls showed virtually no fluorescence, as seen from both the cell images and the intensity distributions (Figure 7.4b, c).

Next, we varied the electroporation voltage from 1.00 to 1.80 kV and measured the distributions of cell-averaged intensities resulting from the internalization of 2.5 μM protein. We calculated the percentage of loaded cells by considering cells that are significantly (by at least 3 standard deviations) brighter than empty cells. As expected, increased voltage led to increased loading (Figure 7.4d), although there was significant variation in loading between experiments. Notably, the non-electroporated control occasionally included cells that exhibited high fluorescence, most likely corresponding to cells with compromised membranes that allow internalization even without electroporation. This phenomenon increases the effective background level of fluorescence, which has to be taken into account when interpreting the loading results. Hence, while 'absolute' loading ranged from 40 % to 70 % of the cells being loaded, loading corrected for non-electroporated cells was between 15 % and 45 %.

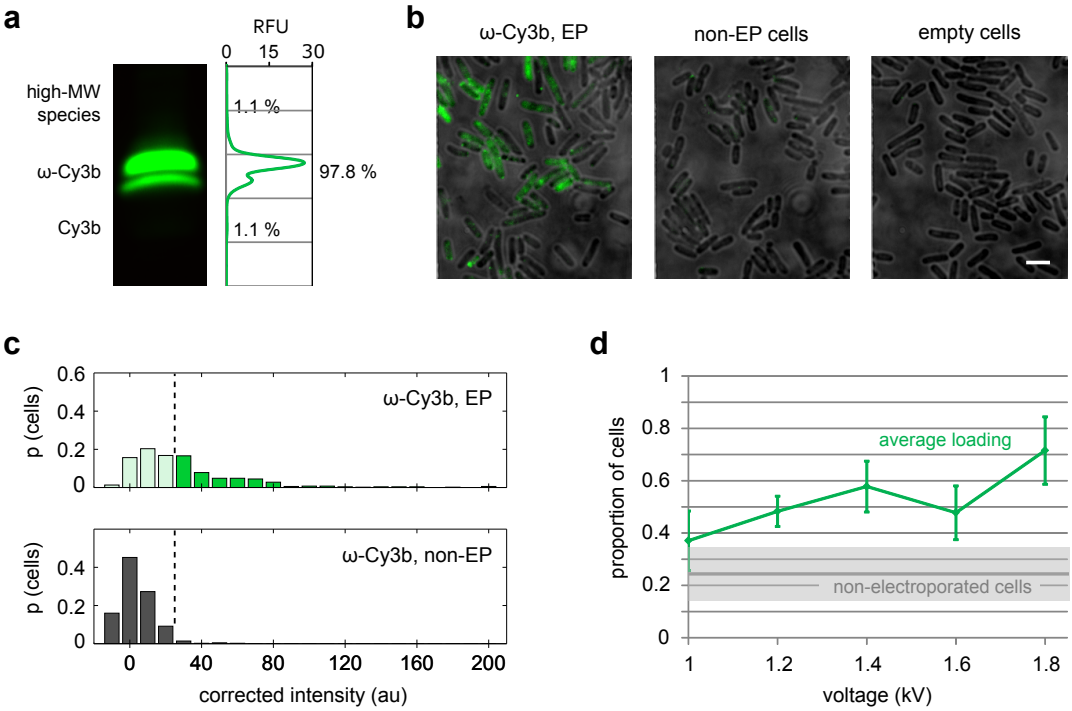


Figure 7.4: *Effect of voltage on protein internalization efficiency.* ***(a)*** *In-gel fluorescence of an SDS-PAGE gel, showing the main ω-Cy3b band and a secondary band. High-molecular weight (MW) species and the contaminating dye are not visible with the naked eye. The fluorescence is quantified (RFU, relative fluorescence units), and the relative amounts of species are indicated.* ***(b)*** *Example fields of view of cells electroporated with 2.5 µM RNAP ω at 1.4 kV voltage, and imaged in widefield mode, with 532-nm excitation at 1 mW power, with 50-ms exposure. Non-electroporated (non-EP) and empty cells are also shown. Scale bar, 3 µm.* ***(c)*** *Distribution of corrected cell-averaged intensities for EP and non-EP cells, corresponding to (b), given in proportion of the total cell count. Intensities corresponding to the loading threshold are indicated with vertical dashed lines, and histogram bars below the threshold shown half-transparent for electroporated cells.* ***(d)*** *Effect of voltage. Loaded cells correspond to cells exhibiting average fluorescence intensity higher than the intensity of non-electroporated cells plus 3 standard deviations. The proportion of loaded cells relative to the total cell count is plotted, with the standard deviation represented by error bars. Intensity of non-electroporated cells is shown for reference, with the standard deviation represented by the grey box. Imaging was done in HILO mode, with 532-nm excitation at 600 µW, 100-ms exposure. > 900 cells (3 independent internalizations of > 300 cells each) were analysed for each voltage condition.*

7.2.3 Effect of voltage on cell viability

In addition, we explored the effect of voltage on cell viability. Following cell electroporation, washing and recovery in EZ rich defined medium, we examined cells in the white-light mode over 1-2 hours, and observed four different classes of cells: (i) *dividing cells*, which divided into daughter cells during the course of imaging; (ii) *growing cells*, which grew in length but did not divide into daughter cells; (iii) *identical cells*, which neither divided nor grew, but their membranes appeared intact; and (iv) *damaged cells*, which showed visibly damaged cell membranes. We quantified the proportion of cells in each class, and observed that the number of growing and dividing cells decreased with voltage, whilst the number of identical and damaged cells increased (Figure 7.5a). In addition, since we noted that a cell-filtration treatment may be needed to remove non-internalized fluorescence after Pol internalization (Section 7.3.2), we repeated the experiments with the additional cell-filtration step prior to cell recovery. The results show a similar trend as in the absence of filtration, but with the viability further compromised at increased voltage (Figure 7.5b).

To rule out the possibility that it is only damaged cells that become loaded with fluorescent molecules, we set out to perform the viability and loading measurements simultaneously, and analysed loading for each 'viability class' of cells separately. We performed these experiments with the additional cell-filtration step, as this appeared to be the more relevant protocol for Pol internalization. The results show that the applied voltage positively affects loading for all classes of cells (Figure 7.5c). The effect of voltage on loading is stronger for damaged cells than for identical cells, and stronger for identical cells than for growing or dividing cells, suggesting that there is a level of negative correlation between cell internalization and viability. However, since the percentage of damaged cells is the sample is low at any voltage (< 10 % without and < 20 % with the cell-filtration treatment), the majority of loaded cells will correspond to identical, growing or dividing cells.

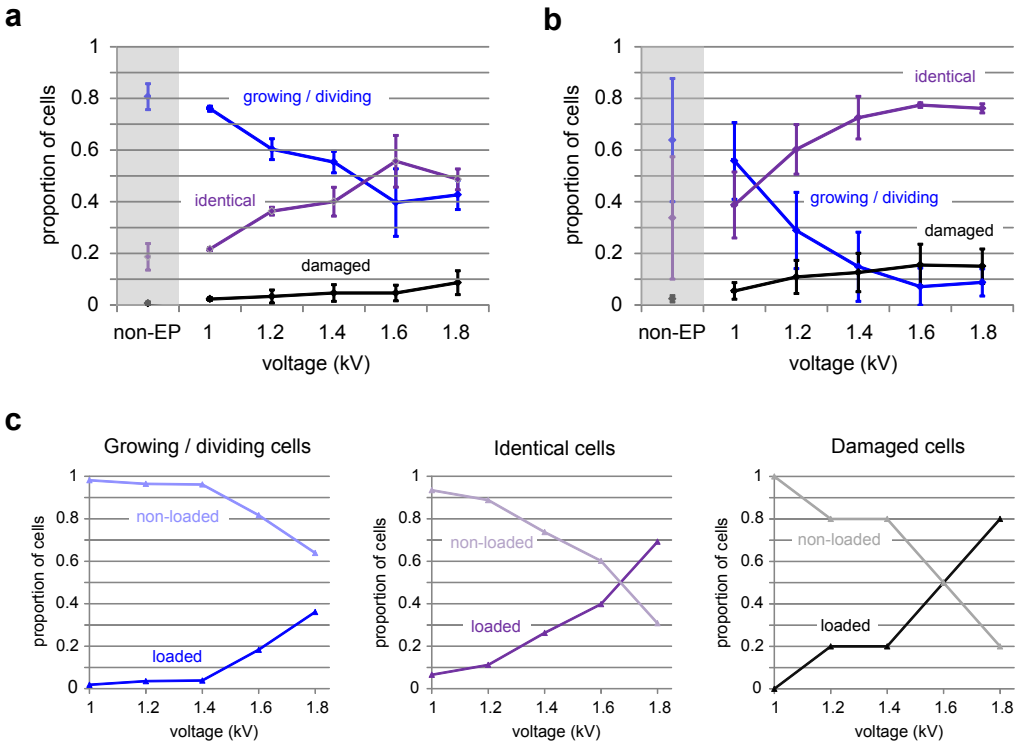


Figure 7.5: *Effect of voltage on cell viability,* ***(a)*** *without and* ***(b)*** *with cell-filtration treatment. The non-electroporated cells are shown for reference, and the standard deviation is represented by error bars. > 300 cells per voltage condition were analysed.* ***(c)*** *Effect of voltage on proportion of loaded and non-loaded cells, analysed separately for each class of cells (growing / dividing, identical and damaged), with the cell filtration treatment. > 300 cells per voltage condition were analysed.*

7.3 Pol internalization

7.3.1 Red-labelled Pol

We began our internalization experiments with Pol (Klenow fragment), due to its smaller size compared to the full-length Pol (68 kDa vs. 103 kDa). We used Pol variant L744C that had previously been singly labelled with two different red dyes, Alexa647 and Atto647N, at > 75 % labelling efficiency. Internalization of Pol-Alexa647 into live *E. coli* showed a relatively high internalization efficiency, with no significant fluorescence outside of the cells, or in the non-electroporated control (Figure 7.6a, c). Cell-averaged fluorescence intensities, corrected for empty-cell fluorescence, showed that more than 60 % of the cells exhibited fluorescence above the background of non-electroporated cells. Pol-Atto647N, on the

other hand, was internalized with much lower efficiency, and showed fluorescent spots that did not always co-localize with cells (Figure 7.6b). These results suggest that Pol-Atto647N may bind non-specifically to the cell membrane, likely due to the high hydrophobicity of the dye, or that it is prone to aggregation. Attempts to optimize Pol-Atto647N internalization were not successful, and we therefore opted for Pol-Alexa647 as the red-labelled Pol construct of choice.

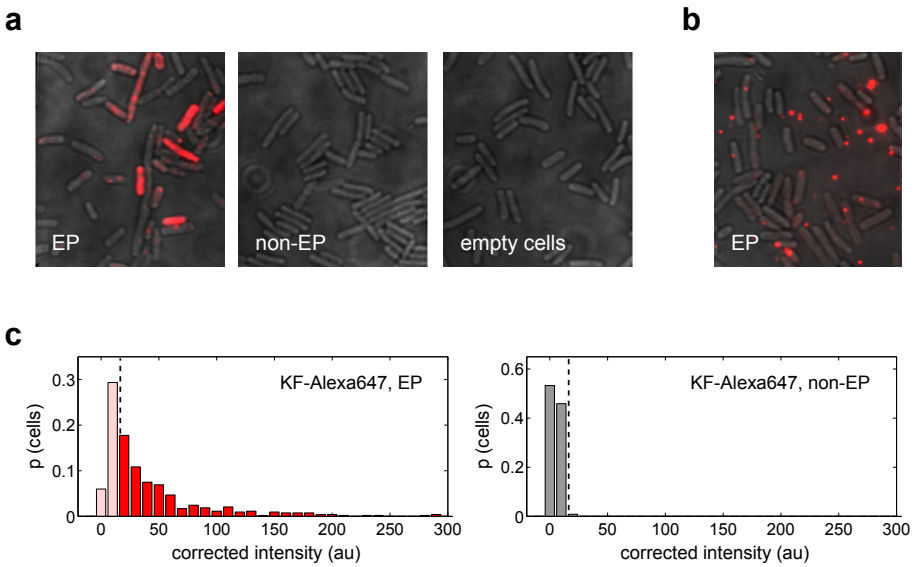


Figure 7.6: *Internalization of red-labelled Pol.* ***(a)*** *Example fields of view of cells electroporated with Pol-Alexa647, non-electroporated and empty cells.* ***(b)*** *Cells electroporated with Pol-Atto647N.* ***(c)*** *Distribution of corrected cell-averaged intensities for EP and non-EP cells incubated with Pol-Alexa647, given in proportion of the total cell count. Intensities corresponding to the loading threshold are indicated with vertical dashed lines, and histogram bars below the threshold shown half-transparent for electroporated cells. Electroporation with (a) 500 nM or (b) 150 nM protein, at 1.8 kV voltage, widefield mode, 637-nm excitation at (a) 1 mW or (b) 3 mW, (a) 50-ms or (b) 100-ms exposure.*

7.3.2 Green-labelled Pol

Non-internalized fluorescence

We next attempted to internalize Pol samples that had previously been labelled with Cy3b, either on the fingers (L744C) or on the thumb position (K550C). We noted that the electroporated cell samples contained non-internalized fluorescent spots that could not be washed off the cells using standard protocols. Stepwise photobleaching of the spots indicated that they arose from single molecules, and hence could not correspond to protein aggregates (Figure 7.7a). Consistent with this observation, gel filtration results indicated that most of the Pol-Cy3b molecules were in a native, monomeric state. To probe the aggregation state of the Pol-Cy3b sample more accurately, we devised a confocal microscopy assay in which we measured single-molecule bursts arising from Pol molecules *in vitro* (Figure 7.7b). We followed the total number of photons per burst, which is dependent on the number of fluorescent molecules arriving in the confocal volume, and hence can reveal the aggregation state of the protein [245, 246]. We further tested the effect of electroporation on Pol-Cy3b aggregation by electroporating Pol-Cy3b under the same conditions as in the cell electroporation experiments but in the absence of cells. The results show that only very few high-molecular weight species (> 500 photons) can be observed either before or after electroporation, suggesting that Pol-Cy3b is neither generally prone to aggregation, nor is its aggregation induced by electroporation.

Since the Pol-Cy3b sample quality appeared optimal and could not be further improved, we sought to address the issue of non-internalized fluorescence using curative procedures. We tested a variety of harsher methods of cell-washing, including cell filtration, detergent treatment and protease treatment. Cell filtration proved to be most successful, with the optimized protocol consisting of 3 steps of cell resuspension and centrifugation at 800 x g over a 0.22 μm filter. When 500 nM Pol-Cy3b was internalized, the addition of the cell-filtration step after cell recovery and washing significantly reduced the level of non-internalized fluorescence, which was particularly clear when the fluorescence focus was at the level of the agarose pad (Figure 7.7c). In addition, changing the recovery medium

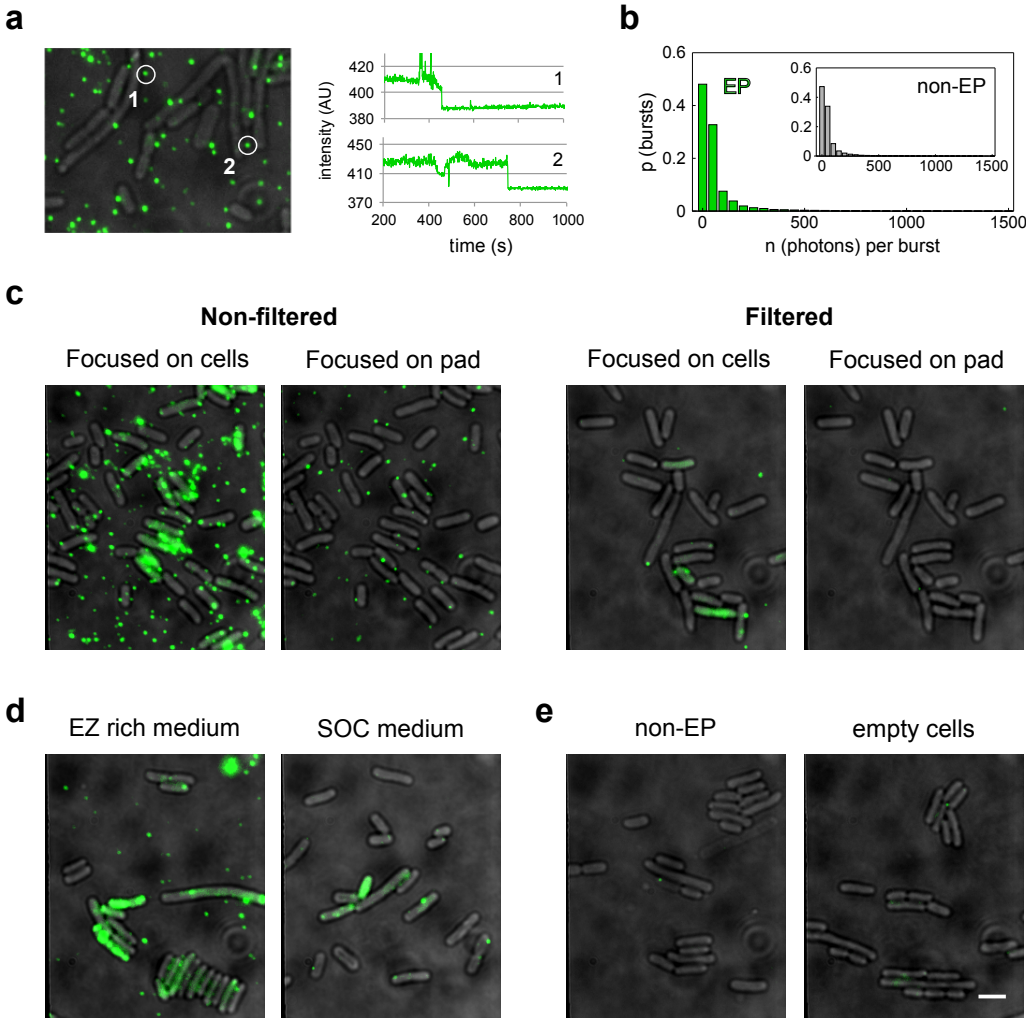


Figure 7.7: *Internalization of green-labelled Pol: non-internalized fluorescence.* ***(a)*** *Example field of view for initial internalization of Pol-Cy3b, showing the issue of non-internalized fluorescent spots. Photobleaching curves for two example molecules (circled) are shown on the right.* ***(b)*** *Distribution of total number of photons per burst, measured by confocal microscopy, of Pol-Cy3b after electroporation of the sample in vitro. Inset, photon distribution obtained without sample electroporation.* ***(c)*** *Example fields of view for Pol-Cy3b internalization, without or with the cell-filtration treatment, with the fluorescence focus either on the cells or the agarose pad.* ***(d)*** *Comparison of Pol-Cy3b internalization with EZ rich defined or SOC medium used as the recovery medium after electroporation.* ***(e)*** *Non-electroporated and empty cells, imaged under the same conditions. Non-EP cells were treated as EP cells, using the optimized internalization protocol. Electroporation with (a, c, e) 500 nM or (d) 750 nM protein, at 1.0 kV voltage, HILO mode, 532-nm excitation at (a) 5 mW or (c-e) 1 mW, (a) 10-ms or (c-e) 50-ms exposure. Scale bar, 3 µm.*

from the EZ rich defined defined medium to Super Optimal broth with Catabolite repression (SOC) medium resulted in cleaner electroporated samples, with smaller numbers of non-internalized spots on the agarose pad (Figure 7.7d). Using the optimized internalization protocol, non-electroporated cells exhibited fluorescence intensities similar to those of empty cells (Figure 7.7e).

Dye contamination

We noted that the level of dye contamination of different Pol-Cy3b samples affected the average cell fluorescence, suggesting that at least in the case of ensemble fluorescence experiments, contaminating dye will interfere with Pol imaging. We reasoned that, despite the low concentration of contaminating dye in the samples, its small size could cause it to be internalized at high efficiency. To quantify the effect of dye contamination in Pol samples on cell fluorescence, we carried out comparative internalization of Pol and free-dye samples. Cells were electroporated with 1.5 μM of a Pol-Cy3b sample containing 15 % dye contamination, or with Cy3b free dye at a concentration corresponding to 15 % contamination. We measured cell-averaged (per-pixel) intensities for > 400 cells per sample, and corrected them for the mean fluorescence of empty cells. Both the cell images and the intensity histograms of the segmented cells showed that the two samples exhibited similar levels of fluorescence (Figure 7.8a). These results suggest that contaminating dye comprises a significant, and sometimes the major, proportion of fluorescence observed.

We therefore devised a protocol for the purification of contaminating dye from Pol samples, which consisted of binding Pol to the Ni-NTA column, and washing the column with a large amount (100 CV) of buffer. Due to the large amount of buffer used, the washing was expected to be effective for removing both the free dye and the dye that was bound non-specifically to the protein. We developed the protocol by varying the buffer composition and the number of washing steps, and measuring in-gel fluorescence of purification fractions on an SDS-PAGE gel. Figure 7.8b shows the in-gel fluorescence for the original sample prior to purification, the flow-through, and the fractions obtained after 30 CV (wash 1), and 100 CV of buffer wash (wash 2). The relative intensities of the different

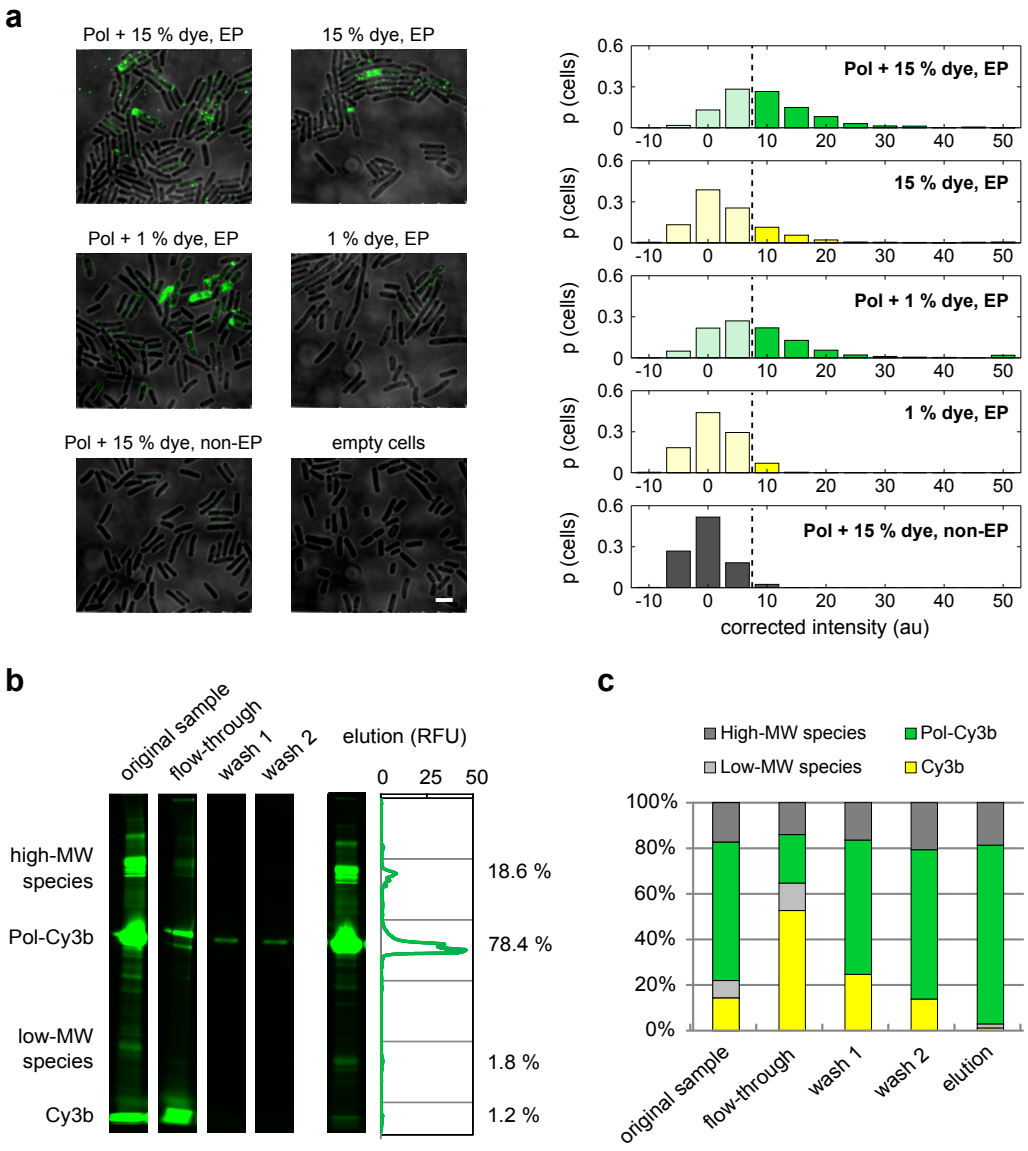


Figure 7.8: *Internalization of green-labelled Pol: dye contamination.* ***(a)*** *Left, internalization of Pol-Cy3b before dye purification (15 % dye contamination) and after purification (1 % cont.), and internalization of Cy3b free dye at concentrations corresponding to 15 % and 1 % contamination. Empty cells and the non-electroporated control for Pol-Cy3b before dye purification (15 % cont.) are also shown. Scale bar, 3 µm. Right, distribution of cell-averaged (per-pixel) intensities, corrected for the mean fluorescence of empty cells, and given in proportion of the total cell count. Intensities corresponding to the loading threshold are indicated with vertical dashed lines. > 400 cells per sample were segmented. Electroporation at 1.4 kV voltage, widefield mode, 637-nm excitation at 1 mW, 50-ms exposure.* ***(b)*** *In-gel fluorescence of an SDS-PAGE gel showing His-tag purification of Pol-Cy3b. The different contaminants are marked: high-, low-MW protein species and free dye.* ***(c)*** *Relative amounts of different species present in the samples, quantified from band intensities, as shown for the last lane in (b).*

bands in each lane can be extracted and correspond to the relative amounts of the main Pol-Cy3b species, and the protein and dye contaminants in each sample (Figure 7.8b, shown for the last lane). From this analysis, it is evident that His-tag purification of Pol-Cy3b results in a gradual but significant decrease in the amount of contaminating dye present (Figure 7.8c, yellow bars). The lowest dye contamination that we could achieve was 1.2 %, more than 12-fold lower than the starting amount of contamination. Other low-molecular weight contaminants that were present in the sample (Figure 7.8c, light grey bars) could also be removed using this procedure, whereas high-molecular weight contaminants remained (Figure 7.8c, dark grey bars). However, the latter are unlikely to be an issue for internalization by electroporation, as they will not be internalized as efficiently as Pol.

To assess whether the remaining level of dye contamination still affected the average cell fluorescence, we performed comparative internalization of the dye-purified Pol sample (containing 1 % dye contamination) and a free-dye sample at a concentration corresponding to 1 % dye contamination. Whilst the distribution of cell-averaged intensities for the Pol-Cy3b sample was similar as before purification, the distribution of intensities for the free-dye sample was significantly lower (Figure 7.8a). In particular, cells loaded with the free dye at 1 % concentration exhibited intensities on the level of fluorescence of empty cells, corresponding to the background autofluorescence. Hence, 1 % dye contamination does not compromise cell loading with Pol-Cy3b, and constitutes a workable condition under which one can be confident that the observed internalized fluorescence corresponds to Pol and not the contaminating dye.

7.3.3 Pol tracking

In order to allow single-molecule detection, we internalized Pol-Alexa647 and Pol-Cy3b at low concentration (1-3 molecules per cell). Single molecules could be observed above the autofluorescence background, and were either stationary or diffusing in the cells (Figure 7.9a). We used custom-made software to localize single molecules and track their diffusion in time (Figure 7.9b); the tracks were used to calculate the mean square deviation,

and hence the average diffusion coefficient of each track. This analysis allowed us to obtain the apparent diffusion-coefficient histograms of Pol-Alexa647 and Pol-Cy3b (Figure 7.9c, d). In the case of Pol-Alexa647, the diffusion coefficient distribution ranged from 0.0 to 2.5 $\mu m^2/s$, with no clear separation between the immobile and diffusing populations of molecules. With Pol-Cy3b, faster-diffusing molecules were seen (up to 3.0 $\mu m^2/s$), and the bimodal distribution of the immobile and diffusing species was more evident. Removing tracks from the analysis that did not colocalize with the cells, or that appeared after all Pol molecules had bleached, eliminated some but not all of the immobile tracks. Carrying out the analysis on empty cells yielded very few tracks for Pol-Alexa647 (~2 % of the number in electroporated cells) and somewhat more for Pol-Cy3b (~10 %), due to the higher autofluorescence in the green channel.

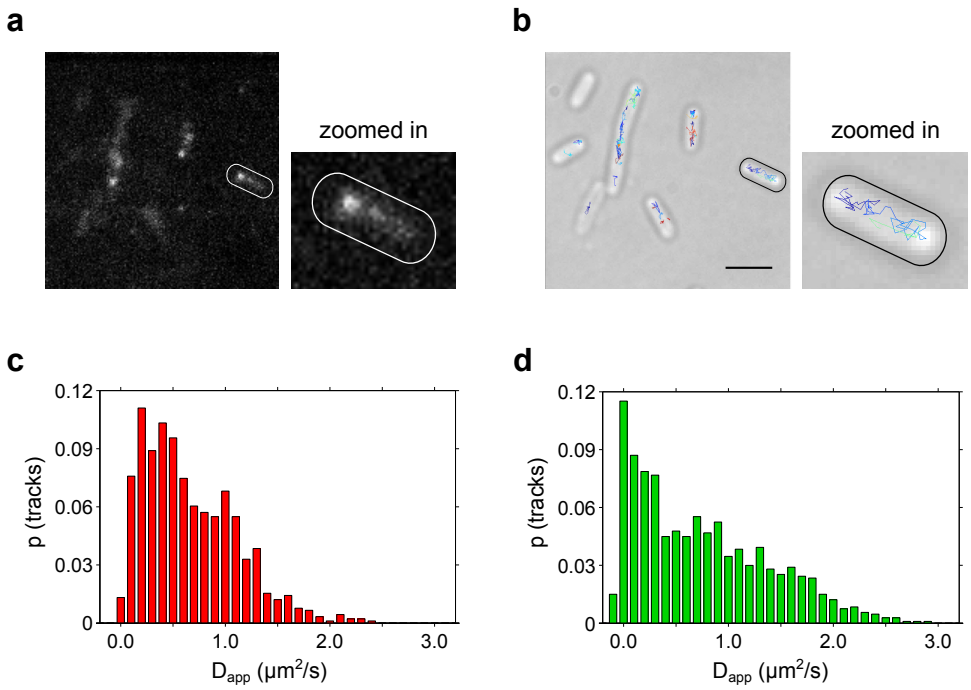


Figure 7.9: *Single-molecule tracking of Pol.* ***(a)*** *Example field of view in green channel, of cells electroporated with Pol-Cy3b, with an enlarged view of a single cell (circled) shown on the right.* ***(b)*** *Same field of view as in (a), but with single-molecule tracks overlaid on the white-light image. Scale bar, 3 μm.* ***(c-d)*** *Apparent diffusion-coefficient histograms for (c) Pol-Alexa647 and (d) Pol-Cy3b. Electroporation with 500 nM protein, at 1.8 kV voltage, HILO mode with (c) 637-nm excitation at 1 mW or (d) 532-nm excitation at 5 mW power, (c) 15-ms or (d) 10-ms exposure.*

7.4 Full-length Pol internalization and tracking

Further to the optimization of Pol (Klenow fragment) internalization and tracking, we attempted to internalize the 103-kDa full-length Pol into live *E. coli*. We used flPol variants that had previously been singly labelled on the thumb position (K550C) with either Cy3b or Alexa647. With both samples, only low internalization efficiencies could be achieved, and non-internalized fluorescent spots were seen that localized both to and outside of the cells (Figure 7.10a, b), similarly to what had initially been observed with the internalization of Pol (KF)-Cy3b. In this case, however, stepwise photobleaching indicated that the non-internalized spots consisted of many molecules, most likely corresponding to Pol aggregates.

We analysed our flPol-Alexa647 sample on a gel filtration column, which indeed showed a significant aggregate peak, corresponding to ~50 % of the molecules being aggregated. To avoid aggregation, we purified a new batch of flPol K550C, and labelled it with Alexa647 under milder conditions than previously. This included performing most steps of the labelling protocol at 4 °C, adding salt to all dialysis and storage buffers, and avoiding some of the dialysis steps. The new sample showed ~8 % aggregation by gel filtration, suggesting an improvement in sample quality. However, internalization of the new flPol sample gave similar results as previously, showing high-intensity spots indicative of non-internalized flPol aggregates.

To test whether electroporation itself was inducing aggregation, we applied the confocal aggregation assay described earlier (Section 7.3.2) to full-length Pol-Alexa647. The results showed that whilst some high-molecular weight species (> 500 photons) could be observed in the flPol-Alexa647 sample already before electroporation, the proportion was significantly increased upon sample electroporation, with ~15 % of the bursts giving > 500 photons and ~6 % of them showing > 1500 photons (Figure 7.10c).

Since the electroporation treatment itself was inducing flPol aggregation, we tried to find a means of removing flPol aggregates from the cell suspension after electroporation. Notably, due to the large size of the aggregates, we reasoned that the majority of them would

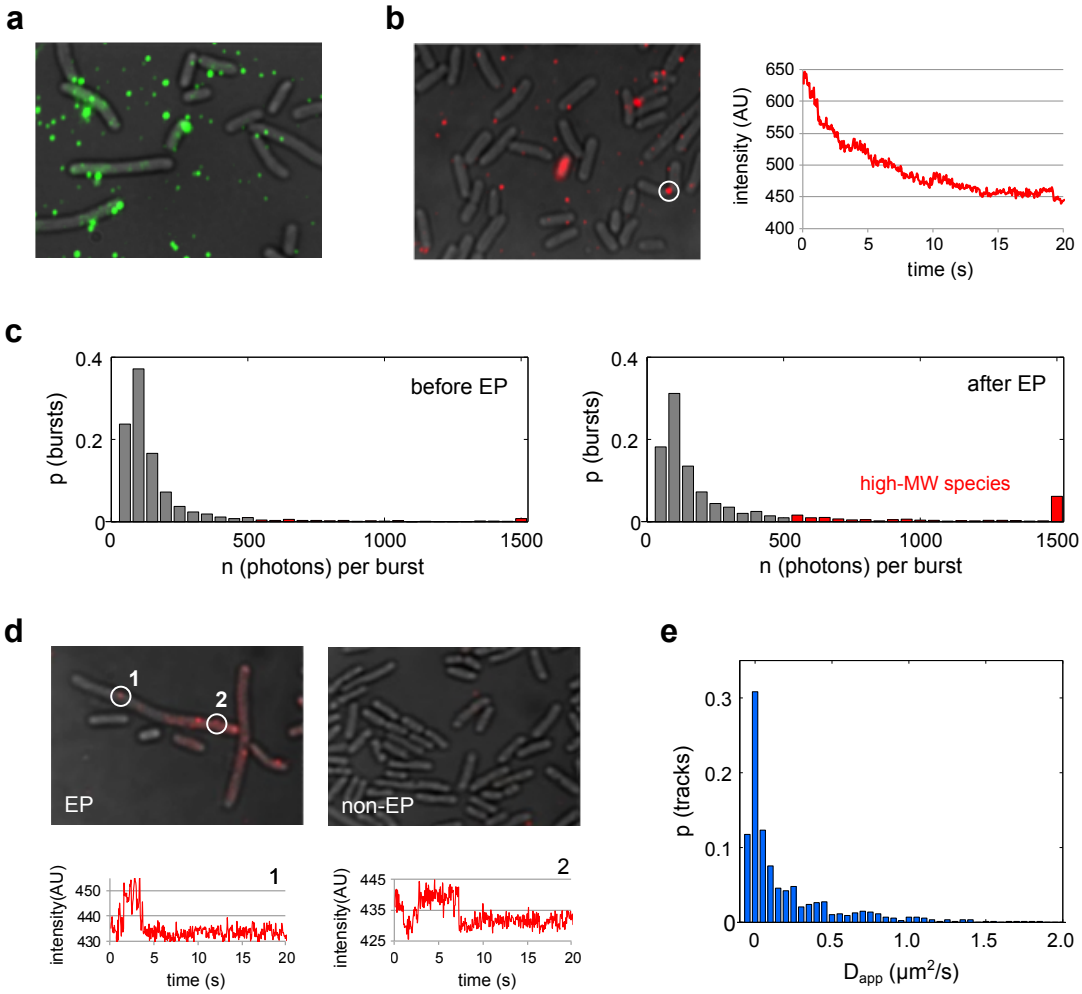


Figure 7.10: *Internalization of full-length Pol.* ***(a-b)*** *Example fields of view of cells electroporated with initial (a) flPol-Cy3b and (b) flPol-Alexa647 samples that contained protein aggregates. A photobleaching curve of a high-intensity spot circled in (b) is shown on the right.* ***(c)*** *Distribution of total number of photons per burst, measured by confocal microscopy, of flPol-Alexa647 before (left) and after electroporation treatment in vitro (right). The bars likely corresponding to Pol aggregates (> 500 photons) are highlighted in red.* ***(d)*** *Left, example field of view of cells electroporated with the newly labelled flPol-Alexa647N sample, and filtered to remove protein aggregates. Photobleaching curves for two example molecules (circled) are shown at the bottom. Right, non-electroporated control for flPol-Alexa647 internalization, also subject to the filtration treatment.* ***(e)*** *Apparent diffusion coefficient distribution of flPol-Alexa647. Electroporation with (a) 500 nM, (b) 1.5 μM or (d, e) 150 nM protein at 1.8 kV voltage, widefield mode with (a) 561-nm excitation at 3 mW, or 637-nm excitation at (b, d) 1 mW or (e) 3 mW power, (a, b, d) 50-ms or (e) 15-ms exposure.*

be non-internalized and hence would be subject to this treatment. In one attempt, cells were attached to polyethylenimine (PEI), a cationic polymer that binds to the negatively charged outer surface of bacterial cells. Although this approach allowed the majority of non-internalized fluorescence to be washed off the agarose pad, photobleaching analysis of cell-localized spots indicated that they corresponded mainly to aggregates, suggesting that the aggregates were either bound non-specifically to the outside of the cells, or that some of them were internalized.

In another attempt, non-internalized flPol aggregates were removed from the cells afer electroporation by means of filtration, using the same protocol as previously for the non-internalized monomeric Pol (KF)-Cy3b (Section 7.3.2). This allowed the majority of non-internalized fluorescence to be removed, but unlike with PEI attachment, the cell-localized spots photobleached in single steps, indicative of flPol monomers (Figure 7.10d). In addition, we were able to record an apparent diffusion-coefficient histogram for Pol, which showed a small number of diffusing molecules, with apparent diffusion coefficients of up to 2.0 $\mu m^2/s$ (Figure 7.10e).

7.5 Discussion

7.5.1 Optimization of protein internalization

A number of general observations can be made for electroporation-based protein internalization. It appears that the specific buffer components and salts can be varied, as long as the overall ionic strength of the buffer remains low. When higher ionic strengths are required to ensure the stability and activity of the protein of interest, the protein should be stored in its optimal buffer at high concentration, so that a larger dilution factor can be achieved when the protein sample is added to cell suspension for electroporation. Alternatively, the protein sample can be diluted into a low-ionic strength buffer and concentrated just prior to electroporation. Notably, the required level of internalization (e.g. dictated by whether imaging is done at the single-cell or single-molecule levels) will determine the amount of

sample that has to be used, and hence the maximum ionic strength that can be present in the buffer. Similarly, higher ionic strengths can be used at lower electroporation voltages (1.00 - 1.20 kV), and glycerol can be added to slightly increase the threshold of acceptable ionic strength.

Protein internalization efficiency is linearly correlated with the applied voltage, consistent with the idea that higher voltage is associated with an increase in the number and size of pores in the bacterial membrane [244]. This dependency allows one to tune the voltage to suit the needs of the experiment, again in terms of whether imaging is done at the ensemble or single-molecule levels. However, we noted a considerable variation in the internalization efficiency between experiments, with some of the non-electroporated cells exhibiting significant fluorescence intensities. The latter effect was observed with both the commercially supplied and the home-made electrocompetent cells. It is likely that these cells feature membranes that have been overly compromised during the induction of electrocompetency, hence allowing protein internalization without electroporation. A population of such cells is likely to always be present in a batch of electrocompetent cells, highlighting the need for non-electroporated controls in electroporation experiments.

Finally, viability decreases with the applied voltage, likely because the increased number and size of membrane pores further compromise the cell integrity. We assume growing and dividing cells to be healthy, and identical cells to be in a viable but stressed state (resulting from the electroporation shock), and in need of further recovery to resume growth and division. However, a more robust viability assay would be needed to confirm these assumptions, such as testing for expression of a fluorescently labelled protein in real time. Cells exposed to the filtration treatment display lower viability than cells washed only by cycles of centrifugation and resuspension, but this may be a necessary price to pay for the removal of non-internalized fluorescence. Notably, the major fraction of loaded cells always corresponds to non-damaged cells, although most of the damaged cells will appear loaded at higher voltages. If absolute viability is required, the loaded cells of interest can always be tested for their viability by following their division after recording the fluorescence data.

7.5.2 Pol internalization

Red-labelled Pol

Using the above guidelines, we demonstrated efficient internalization of Pol-Alexa647 into *E. coli*. The lack of non-internalized fluorescent species and the clean non-electroporated controls indicate that Pol-Alexa647 is not prone to non-specific binding to the cell membrane, and that the standard protocols of cell washing are sufficient. Unfortunately, this is not the case for Pol-Atto647N, which displayed apparent non-specific binding to the cells and the agarose pad, likely due to the hydrophobicity of the dye [227]. Non-specific binding of Atto647N has previously been noted with the internalization and tracking of Atto647N-labelled DNA constructs, although much cleaner internalizations could be achieved in that case [229]. We noted that, whereas Atto647N is a superior fluorophore to Alexa647 in terms of its brightness [225] and photostability [224], its propensity for non-specific binding would make it unsuitable for in-cell applications, and hence we did not proceed with further optimization of Pol-Atto647N internalization.

Green-labelled Pol

Internalization of Pol-Cy3b proved difficult, mainly due to the occurrence of non-internalized fluorescent spots in the electroporated samples. This phenomenon is not observed for DNA internalization, and likely arises from non-specific binding of the labelled protein to the cell membrane. It is possible that Pol-Alexa647 is less susceptible to non-specific binding than Pol-Cy3b because of the negatively charged nature of Alexa647, which could repel it from the cell membrane. Notably, we observed a large number of non-internalized spots that did not colocalize with the cells, begging the question of why these molecules are not removed in the process of cell centrifugation. Since protein aggregation was excluded, one possibility is that the fluorescent molecules can bind transiently to the cells during the washing procedure, and then dissociate in subsequent steps of sample preparation. Alternatively, the molecules could bind to cellular debris present in the cell suspension, which is spun down during centrifugation.

Encouragingly, we showed that cell filtration constitutes an effective means of removing the non-internalized fluorescence, allowing clean internalizations of Pol-Cy3b. Fluorescence contamination is reduced both at the level of the cells and at the level of the agarose pad, suggesting that both the molecules that bind non-specifically to the cells and the molecules that float in the cell suspension are removed. The reason why filtration is superior to simple cell centrifugation is unclear, but it is likely either because the size of the membrane pore allows the cellular debris to pass through, and/or because the constant flow of buffer provides an additional force to remove the membrane-bound molecules from the cells. We also noted that SOC medium is superior to EZ rich defined recovery medium in terms of preventing the occurrence of non-internalized fluorescence. It is possible that the different chemical compositions and ionic strengths of the two media affect the efficiency of interactions between the labelled protein and the cell membrane.

Dye contamination

Internalization of Pol-Cy3b made us aware of the issue of dye contamination in our samples, which interfered with Pol imaging. Notably, contaminating dye can either be present as 'free' unreacted dye, or can be bound non-specifically to the protein, via hydrophobic or electrostatic interactions. The free-dye contamination is particularly problematic in terms of the internalization by electroporation. The reason lies in the small size of organic dyes compared to proteins, which causes dyes to be internalized at a higher efficiency than proteins, and the fact that the internalized free dye may be difficult to distinguish from the labelled protein inside the cell. The non-specifically bound dye, on the other hand, is not necessarily an issue; it is expected to be internalized at an efficiency similar to the internalization efficiency of the protein, and may or may not dissociate from the protein once internalized. We attempted to quantify these two populations of the contaminating dye in our Pol-Cy3b samples using native PAGE; however, it was not possible to achieve a clear separation of the dye from the protein on the gels.

We succeeded in removing most of the contaminating dye with His-tag purification. Washing the column with a large volume of a mild buffer proved more effective than wash-

ing with buffers containing high salt or non-ionic detergents, which mainly promoted dissociation of Pol from the column. His-tag purification likely works primarily against the free-dye contamination, although the large flow of buffer may also help to remove any protein-bound dye, depending on the dissociation kinetics of the interaction. The resulting 1 % *total* dye contamination quantified from SDS-PAGE, which corresponds to a maximum of 1 % *free* dye contamination, did not show a significant cell-averaged signal, suggesting that this level of dye contamination constituted a workable condition for Pol-Cy3b internalization and imaging.

It should be noted that dye contamination is very common in organically labelled protein samples, and is therefore an important general factor to consider in electroporation-based protein internalization. Dye contamination will interfere with single-cell imaging and with single-colour single-molecule tracking, although in the latter case it may be possible to distinguish the dye from the labelled protein, depending on their diffusion coefficients. Since the highest acceptable level of dye contamination will depend on the specific protein and organic dye that are used, appropriate controls of side-by-side internalization of the protein and free dye should be carried out for each new labelled protein and construct under study. Unfortunately, we only realized the importance of dye contamination at the end of the project timeline, and therefore did not thoroughly control for this phenomenon in the case of red-labelled Pol, or the green- or red-labelled full-length Pol. The level of dye contamination in these samples was on the order of 6-9 %, implying a significant effect on cell fluorescence intensities. The reported intensities for these internalization experiments should therefore be interpreted with caution.

7.5.3 Pol tracking

We were able to track single internalized molecules of Pol-Alexa647 and Pol-Cy3b, and measure their apparent diffusion-coefficient distributions. The diffusing populations were centred at ~1.0 $\mu m^2/s$, consistent with the previously reported value for FP-tagged full-length Pol diffusion in cells (~0.8 $\mu m^2/s$) [15]. However, whilst only 3 % of the FP-tagged

flPol were previously seen to be immobile in undamaged cells, we observe much larger immobile populations with the organically labelled Pol. This effect could be either due to non-specific binding of labelled Pol to intracellular structures, or due to artefacts generated by the tracking software. The higher brightness of Cy3b compared to Alexa647 allows faster-diffusing species to be probed, hence providing a better resolution between the diffusing and immobile species. Pol-Cy3b would therefore likely be the construct of choice to probe intracellular Pol dynamics, despite the interference of the autofluorescence background. It should be noted that diffusion of Pol-Alexa647 and Pol-Cy3b was measured prior to the optimization of dye contamination, and hence the fast-diffusing populations may include a contribution from the free dye molecules. We expect free-dye diffusion to be too fast to be detectable using our time resolution (10-15 ms), but slower apparent dye diffusion may result from occasional non-specific binding of the dye to cellular structures.

7.5.4 Full-length Pol internalization and tracking

Despite the large size of full-length Pol (103 kDa), we reasoned that internalization of flPol on the level required for single-molecule experiments may be possible, considering the previous demonstration of T7 RNAP internalization (98 kDa) [229]. However, the aggregation propensity of flPol has significantly limited the possibilities for its internalization by electroporation. The tendency of flPol to aggregate with time, particularly when organically labelled, has previously been noted in our laboratory, and our confocal assay for aggregation suggests that electroporation further promotes aggregation. Notably, this is not a general effect of electroporation, since Pol (KF)-Cy3b is not sensitive to electroporation-induced aggregation, but it may affect aggregation-prone proteins such as the full-length Pol. Interestingly, a recent study noted an effect of electroporation on siRNA aggregation, attributed to the release of aluminium ions from the electroporation cuvette [247], and similar mechanisms could be in play in electroporation-induced protein aggregation.

It is likely that the limited membrane-pore size of the electroporated cells prevents internalization of higher-order flPol oligomers, thus effectively working as a filter to remove

the aggregates. Indeed, using our cell-filtration treatment to remove the non-internalized fluorescence, we observed spots corresponding mainly to monomeric molecules in the cells. However, the majority of the internalized molecules were immobile, unlike what has been observed with FP-tagged flPol [15], likely indicating that the structure of Pol was compromised and that it was binding non-specifically to cellular structures. Therefore, any physiological studies of Pol structure and dynamics in cells using this approach will require further sample optimization, as wells as modifications to the electroporation protocol to minimize the aggregation effects [247].

7.6 Conclusions and future work

In this chapter, we have characterized the electroporation protocol for the internalization of proteins, in terms of the optimal buffer conditions and the effect of voltage on protein internalization efficiency and cell viability. We have also addressed the issues of non-internalized fluorescence and dye contamination, and developed optimized protocols that enable clean and efficient internalization of Pol, as well as providing useful guidelines for electroporation-based protein internalization in general. Both red- and green-labelled Pol can be internalized into cells and tracked at the single-molecule level, allowing Pol dynamics and localization to be studied on a much longer time scale than previously possible [15]. In addition, these experiments set the stage for single-molecule FRET studies of (fl)Pol structure and conformational states, which we attempt in the next chapter.

In the future, the use of infrared dyes should be explored, as these would allow Pol tracking without the interference of autofluorescence. To this aim, we have labelled Pol with the infrared dye Alexa750, and the optical modifications to our microscope set-ups are underway. Photoactivation of organic dyes will also likely prove useful, as it would allow autofluorescence to be bleached prior to the start of imaging. Photoactivation would additionally allow true super-resolved tracking [226], whereby a high concentration of fluorescently labelled protein could be internalized without compromising single-molecule detection, thus significantly improving the throughput of cellular imaging.

7.7 Materials and methods

7.7.1 Sample preparation

C-terminal His_6-tagged RNAP ω (C68) was previously expressed using the pET expression system and purified on a Ni-NTA column under denaturing conditions. The protein was reduced either as previously described [248], or using Reduce-Imm column (Pierce) according to manufacturer's instructions. Fluorescent labelling was performed using Cy3b-maleimide (GE Healthcare) as in reference [248]. The labelled protein was purified from excess dye on a Ni-NTA column.

Pol (KF) samples were previously expressed and purified as described in Section 6.7.6, and singly labelled with Alexa647, Atto647N or Cy3b as described in reference [13]. Full-length Pol was expressed and purified fresh, using the same protocol except that the last dialysis step was avoided. TCEP reduction was also carried out as before except that 50 mM NaCl was added to all dialysis buffers. Full-length Pol was labelled with Alexa647 by incubating 7.5 nmol of protein with 15 mol of DMSO-dissolved Alexa647-maleimide (Thermo Fisher Scientific) in 300 µl reaction volume, over-night on a rotating wheel at 4 °C. Heparin purification was done as before, and the samples combined in a 1:1 ratio with 2x glycerol storage buffer and stored at -20 °C. Labelling efficiencies, quantified from UV-Vis spectra, were between 75 % and 90 %. For electroporation, the labelled proteins were dialysed into a low-salt buffer consisting of 50 mM Tris pH 7.4, 25 mM NaCl, 1 mM DTT and 50 % glycerol.

7.7.2 Internalization by electroporation[2]

Electrocompetent cells (Electro MAX DH5α-E, Invitrogen) were diluted 1:1 in water and stored in 20-µl aliquots. Fluorescently labelled protein was added to an aliquot of cell suspension at 50 nM to 2.5 µM concentration and transferred to a pre-chilled 1-mm elec-

[2]A movie demonstrating the optimized protocol of internalization by electroporation, along with the guidelines for cell imaging and data analysis, is available at `http://www.jove.com/video/52208/internalization-observation-fluorescent-biomolecules-living/` [249].

troporation cuvette. For dye-contamination experiments, Cy3b-maleimide dye was inactivated with 10 mM DTT for 10 min and diluted in water before being added to cells. Electroporation was performed at 1.0 to 1.8 kV voltage in a standard electroporator (MicroPulser, Bio-Rad). Cells were recovered by incubation with 500 µl of pre-warmed SOC medium (Invitrogen) or EZ rich defined medium (Teknova) for 3 min at 37 °C. After recovery, cells were pelleted for 1 min at 3,300 x g and 4 °C, and washed with phosphate buffered saline (PBS) solution containing 100 mM NaCl and 0.005 % Triton X100. Washing was repeated 2 more times with the same buffer, and 3 more times with PBS only. In the case of viability analysis, cells were further recovered in EZ rich defined medium for 1-2 hours at 37 °C. Non-electroporated control samples were treated identically except that no electroporation was performed. Empty-cell samples were prepared by diluting electrocompetent cells 5-10x in PBS. 5 µl of cells was applied to pads containing 1 % agarose (Bio-Rad Certified Molecular Biology Agarose) and 1x M9 minimal medium. In the case of viability analysis, M9 salts were replaced with the fluorescence-friendly EZ rich defined medium to ensure cell growth and division.

7.7.3 Widefield and TIRF imaging

Samples were imaged on a customized inverted Olympus IX-71 microscope with a TIRF set-up. The agarose pads were sandwiched between two coverslips and placed on the objective with the cell-covered side facing downwards. For viability analysis, the objective was heated to 37 °C (Objective Heater System, Bioptechs) to promote cell growth and division. Beams from a 532-nm Nd:YAG (Samba, Cobolt AB) and a 637-nm diode laser (Stradus, Vortran) were combined and collimated before being focused onto the back focal plane of the objective. The incident angle of the beam was adjusted such that either widefield or HILO illumination was achieved. Fluorescence from the sample was collected through the same objective, separated from the excitation light using a long-pass and a notch filter, and split into red and green channels using a dichroic mirror (630DRLP, Omega). The two channels were imaged onto separate halves of the chip of an EM-CCD

camera (iXon +, 887-BI, Andor technology). Videos were recorded with manufacturer's software, using the kinetic mode with 50-100 ms exposure. White-light images were obtained using a white-light lamp (IX2-ILL100, Olympus) and a condenser (IX2-LWUCD, Olympus) attached to the microscope as an illumination source.

7.7.4 Buffer-only electroporation

For buffer optimization experiments, buffers containing 50 mM Tris pH 7.4, 0-150 mM NaCl and 0-40 % glycerol were diluted 20 times in water, to simulate the dilution under conditions of cell electroporation. Electroporation was performed at 1.0 - 1.8 kV in the absence of cells, using the same cuvette for each buffer condition, and the electroporation time constant was measured each time. Pure deionized water was tested for reference.

7.7.5 Internalization and viability analysis

Internalization images were obtained in Fiji by overlaying inverted white-light images and false-coloured fluorescence images, averaged over 10 frames. Cells were segmented using an adapted version of programme 'Schnitzcells' [250], and cell intensities quantified and normalized for the cell area by means of a custom-written MATLAB script. Intensities were corrected for the mean intensity of empty cells, to account for cellular autofluorescence. Viability was analysed by manually comparing white-light images of cells taken every 20-40 min, and classifying cells as growing / dividing, identical or damaged. For comparative analysis of loading and viability, Micromanager was used to set the microscope stage to image an area of the pad first in the fluorescence mode for the loading analysis, and then in the white-light mode for the viability analysis.

7.7.6 Treatment of non-internalized fluorescence and aggregates

Stepwise photobleaching was performed using normal laser intensities, and the resulting fluorescence of single spots measured in Fiji using the 'Plot Z-axis Profile' function. Gel filtration was carried out using a Superdex 200 column (10/300 GL, GE Healthcare Life

Sciences), on a fast protein liquid chromatography system (AKTA, GE Healthcare Life Sciences). The gel filtration buffer was prepared from 50 mM Tris pH 7.4, 150 mM NaCl and 1 mM DTT.

For the aggregation assay, Pol samples were diluted in water to the same concentration as in cell electroporation experiments, and electroporated under the same conditions. The electroporated sample was diluted to 100-200 pM in Pol buffer, and single-molecule measurements performed as in Section 4.8.2. Two to four datasets of 10 min were recorded for each sample, and the data analysed as described before.

Cell filtration was performed after the first cycle of cell washing. Cells were transferred to an Ultrafree-MC centrifugal filter tube (0.22 μm pore diameter) and spun 3 times for 3 min at 800 x g and 4 °C, before the remaining cycles of cell washing were carried out. PEI attachement was performed just prior to cell imaging. Branched polyethylenimine (60-750 kDa, Sigma-Aldrich) was diluted from a 50 % stock in water to a working concentration of 1 %. PEI was applied to gasket wells, left to adsorb for 10 min, and was then washed 5 times with PBS. 10 μl of cells was applied to the PEI before it had dried out, the cells were left to bind for 3 minutes and were washed 5 times with PBS.

7.7.7 Analysis and removal of dye contamination

Pol samples were added to 4-6x the recommended amount of Ni-NTA resin (binding capacity 10-15 mg/ml), and incubated on a rotating wheel for 30 min at 4 °C. The resin was washed with 100 column volumes (CV) of the Ni-NTA buffer (50 mM Tris pH 7.1, 25 mM NaCl, 10 mM imidazole), and the fluorescent protein eluted with 20 CV of Ni-NTA buffer containing 200 mM imidazole. Eluate was dialysed into 50 mM Tris pH 7.5, 25 mM NaCl, 1 mM DTT, 50% glycerol, and stored at -20 °C. Fractions from different steps of the purification procedure were run on a denaturing SDS-PAGE gel (Mini-PROTEAN TGX Precast Gels, Bio-Rad), using a transparent sample buffer (250 mM Tris pH 6.8, 20 % glycerol, 2 % SDS, 1 mM DTT). In-gel fluorescence was imaged (Molecular Imager PharosFX Plus System, Bio-Rad), and the gels stained with Coomassie Brilliant Blue to confirm the

identity of protein bands. Fluorescent bands were quantified in Fiji using the 'Plot Profile' function, and the peaks integrated in OriginPro (OriginLab).

7.7.8 Single-molecule tracking

Single-molecule diffusion was analysed essentially as described [15, 229]. A fixed localization intensity threshold was applied on the bandpass-filtered fluorescence image [251], and the positions of point spread functions fitted by 2D elliptical Gaussian functions. Single-molecule tracking was carried out using a custom-written MATLAB script, based on an existing algorithm [252]. The localized PSFs in consecutive frames were linked to a trajectory provided that they appeared within a defined window (7 pixels or 0.69 μm). Transient PSF disappearance due to fluorophore blinking or missed localization was accounted for by using a memory parameter (1 frame). To eliminate noise, only tracks with a minimum of 4 steps were considered, and the apparent diffusion coefficient D_{app} was calculated for each track from the mean square displacement $\langle \Delta r^2 \rangle$, according to: $D_{app} = \langle \Delta r^2 \rangle / 4\Delta t$. The apparent diffusion coefficients were corrected for the localization standard deviation and plotted in histograms. It should be noted that, due to the effects of cell confinement and motion blurring, the coefficients calculated in this way do not directly correspond to accurate microscopic diffusion coefficients [253].

7.8 Contributions

- Parts of the introduction have been published in reference [254].
- Parts of the results and discussion have been published in references [255] and [249].
- RNAP ω was previously purified and labelled by members of Nikolay Zenkin's group.
- Experiments to determine the effect of voltage on the internalization efficiency of RNAP ω and cell viability were carried out jointly by myself (sample preparation), Anne Plochowietz (loading measurements and analysis) and Louise Aigrain (viability measurements and analysis).
- Pol variants were previously purified and labelled by members of Cathy Joyce's group. Optimized preparation of the full-length Pol was done by myself.

8

Probing Pol structure in cells

8.1 Introduction

8.1.1 Project rationale

Following on from the optimization of Pol internalization and tracking in the previous chapter, we set to study Pol in live cells by single-molecule FRET. The first part of this project aims to establish smFRET detection of doubly-labelled Pol in cells, and test if the structure and conformational states of Pol observed *in vitro* are relevant in the context of the cell. This 'structural smFRET' capability would open a plethora of possibilities, such as studying the domain arrangement in full-length Pol, measuring the kinetics of DNA binding and synthesis by Pol, and following enzyme conformational changes in real time.

In the second part of the project, we establish an indirect detection of full-length Pol in cells via smFRET, building upon the studies of Pol binding to gapped DNA, presented in Chapter 4. In particular, to test if the proposed mechanism of DNA recognition is physiologically relevant, we probe the bending of gapped-DNA in the cell, and investigate the existence of the Pol dimer species. Before we present the results of these studies, in the following introductory section we explore the method of in-cell single-molecule FRET, highlighting its underlying technical challenges and reviewing its applications.

8.1.2 In-cell single-molecule FRET

With FRET being the ultimate 'molecular ruler', its application in the context of the living cell has significant potential for understanding macromolecular structure and function under physiological conditions. Ensemble FRET measurements using fluorescent proteins are feasible *in vivo* [256], and have been used to study protein interactions and signaling cascades, and as a basis for biosensors [257–260]. However, single-molecule FRET detection in live cells was long limited due to technical challenges. With the obvious advantages of single-molecule detection, significant effort has been invested to address these challenges, in terms of the labelling, internalization and smFRET imaging in live cells, as we discuss below. These efforts have recently enabled smFRET detection in both microbial and mammalian cells [229, 231, 261, 262].

Technical considerations

Fluorophore requirements for smFRET detection in cells are similar to those for general single-molecule detection (Section 7.1.2), with some additions. First, it should be possible to attach the probe to any position of interest, without affecting biomolecular function. Second, the probe should exhibit high rotational freedom, since fixed orientations of the probe can give rise to anisotropy effects and result in large distance errors. Hence, even more so than is the case for general single-molecule detection, organic fluorophores are superior to FPs as probes for smFRET imaging in cells (Figure 7.1). The chosen FRET dye pair should have sufficient spectral overlap, and an appropriate Förster radius to allow FRET to occur in the desired distance range. Although red-green dye pairs are the most popular, red-infrared dye pairs can be used to avoid issues with autofluorescence [231].

smFRET in cells can be imaged using either confocal or TIRF microscopies. With TIRF, single molecules can be tracked simultaneously in the donor and acceptor channels, allowing smFRET of diffusing molecules to be followed over time [261]. Background autofluorescence does not generally interfere with smFRET tracking, because only the molecule of interest will give a signal in the FRET channel. However, correcting for autofluorescence

is nevertheless important for estimating correct FRET efficiencies, since autofluorescence often varies among cells and between the donor and acceptor channels. In addition, loss of the acceptor fluorophore caused by photobleaching or intracellular degradation can skew the distributions of FRET efficiencies [229]. Alternating-laser excitation microscopy can be used to correct for donor-only species in such situations [70].

Intracellular effects on the photophysical properties of the dyes can also cause a shift in the value of the Förster radius (R_0) of a dye pair. Measurements with the Atto640-Atto740 dye pair in bacterial lysate indicated significantly lower R_0 values compared to their *in vitro* counterparts, due to the differences in the quantum yield, the donor-acceptor cross-talk, and the refractive index of the medium [231]. In addition, the anisotropies of the two dyes were measured to be higher in the lysate, likely due to the high viscosity of the cytosol. The higher anisotropies translate into a broader distribution of the orientational factor of the dye pair, and hence a broader distribution of possible R_0 values in the cell. Studies so far have found different levels of agreement between *in vitro* and in-cell smFRET values [229, 231], depending on the system studied and the conditions used. Hence, parallel in-cell and *in vitro* measurements should always be carried out on well-characterized FRET standards (such as doubly labelled DNA fragments), before any FRET states observed in cells can be interpreted with certainty.

Current applications

Two studies have demonstrated single-molecule FRET capability in prokaryotic systems so far. Fessl and coworkers used heat shock to internalize Atto680- and Atto740-labelled DNA constructs of 8 to 16 base-pairs in length into *E. coli*, and showed that they adopt the B-DNA, rather than the A-DNA conformation in cells [231]. In our laboratory, we demonstrated internalization by electroporation of Cy3b- and Atto647N-labelled single-stranded and double-stranded 45-mer DNAs [229], and their imaging in *E. coli* using TIRF and HILO microscopies (Figure 8.1a). DNA constructs with different separation between the labels produced sufficiently different FRET signatures in cells that they could be distinguished from each other.

In mammalian systems, early studies demonstrated smFRET detection on and at the cell surface, between organic dye-labelled ligands, or between a ligand and an FP-tagged membrane protein [263, 264]. The first smFRET characterization inside live mammalian cells used microinjection to deliver a membrane-fusion protein SNAP-25 into mammalian kidney cells, followed by TIRF imaging [261]. The protein was seen to incorporate into a folded complex with its SNARE protein partners at the cell membrane, resulting in the formation of a high-FRET species (Figure 8.1b). Single-molecule tracking was demonstrated for the membrane-tethered, but not for cytosolic SNAP-25.

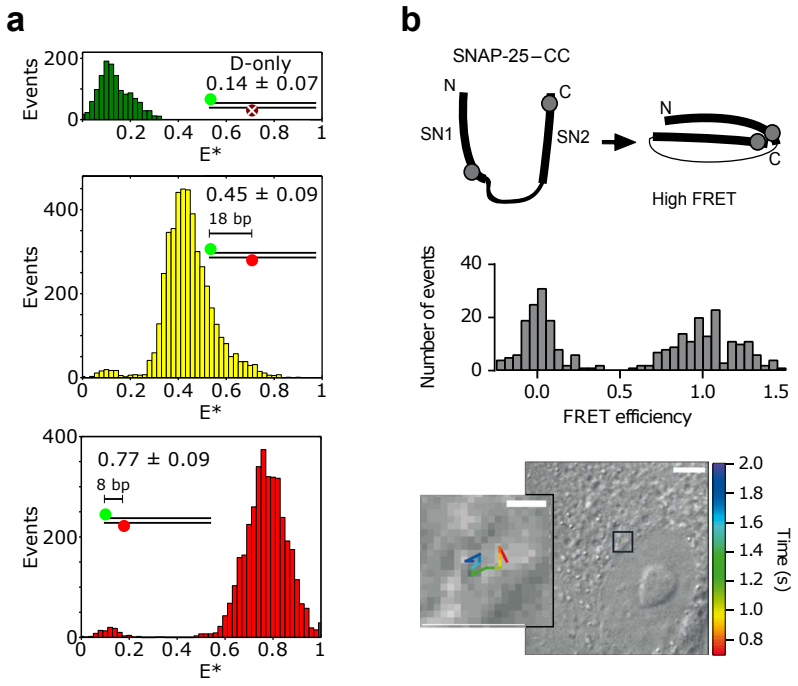


Figure 8.1: *Applications of smFRET in live cells.* ***(a)*** *Demonstration of smFRET in bacteria, using double-stranded 45-mer DNA with different separation between labels. Shown are smFRET efficiency histograms for donor-only (top), intermediate-distance (center) and short-distance species (bottom). Average FRET values based on Gaussian fitting are indicated.* ***(b)*** *Use of smFRET to probe SNARE complex formation in mammalian cells. Top, a schematic of internalized construct SNAP-25-CC, which forms a high-FRET species upon folding into a complex at the cell membrane. Centre, single-molecule FRET efficiency histogram of SNAP-25 in cells. Bottom, example single-molecule tracking of SNAP-25, with intensity timecourse. A close-up of the boxed region in the larger image is shown in the inset. Adapted from references [229, 261].*

In a recent study from Ben Schuler's laboratory [262], microinjection was used to internalize a series of doubly labelled proteins into HeLa cells, and investigate their conformational dynamics and folding using confocal microscopy. The authors showed that the intrinsically disordered protein (IDP) ProTα gives the same mean FRET efficiency in the nucleus, in the cytosol and extracellularly, indicating that this IDP remains unstructured inside the cell. Further, they could observe cold denaturation of the protein frataxin, and measure folding kinetics of protein GB1 on a millisecond timescale.

Future applications

With additional technical developments, in-cell smFRET could be used for a number of novel applications [254]. The most immediate potential lies in *intra*-molecular smFRET, which could probe biomolecular interactions, and interaction-induced structural changes. If a sufficient number of intramolecular distances were measured, then smFRET could be used as a structural tool, allowing *in vitro* X-ray and NMR structures to be tested for their physiological relevance. Structure determination will be most straight-forward with nucleic acids, as they can easily be labelled at any position, with structured RNAs appearing as excellent targets [265]. In addition, the kinetics of intramolecular dynamics could be studied with 10-20 ms temporal resolution using widefield approaches, or faster using confocal microscopy, albeit at the expense of long observation times. In this way, smFRET could be used to probe conformational changes in proteins and molecular machines in real time, allowing their activities to be followed under native conditions.

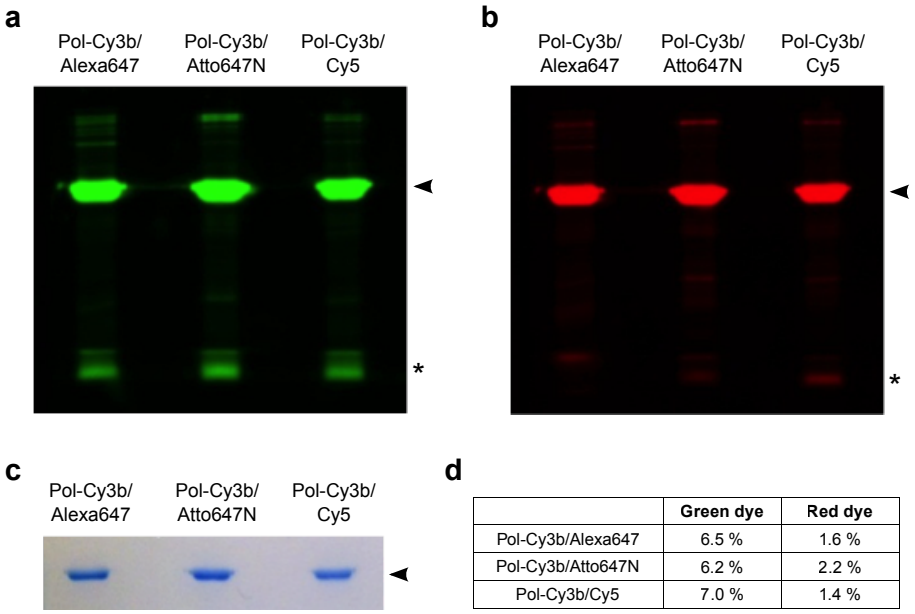
8.2 Polymerase domain structure

In this section, we use smFRET to probe the structure of Pol in cells, in terms of the relative arrangement of the fingers and thumb subdomains. We chose to start with the Pol (KF) construct due to the current issues with the internalization and single-molecule tracking of full-length Pol. We aimed to detect the open and closed conformations of Pol, to confirm the physiological significance of the X-ray and smFRET results. In addition, this capability

could allow us to test whether the conformational dynamics of Pol are affected in cells, an open question with significant implications for the mechanisms of DNA recognition, synthesis and nucleotide discrimination by Pol.

8.2.1 *In vitro* characterization

To probe the structure of Pol in cells, we used Pol samples labelled with donor and acceptor fluorophores on the fingers (L744C) and thumb positions (K550C), respectively. The donor dye used was always Cy3b, whereas three different acceptor dyes were tested (Alexa647, Atto647N and Cy5), as we reasoned that some dyes may perform better than others in the context of in-cell smFRET. When analysed by SDS-PAGE, all samples produced clear bands at the expected molecular weight for Pol, which were fluorescent in both green and red channels (Figure 8.2). Dye contamination was quantified from the gels and was found to be on the level of 1-2 % for the red dyes, and 6-7 % for Cy3b.

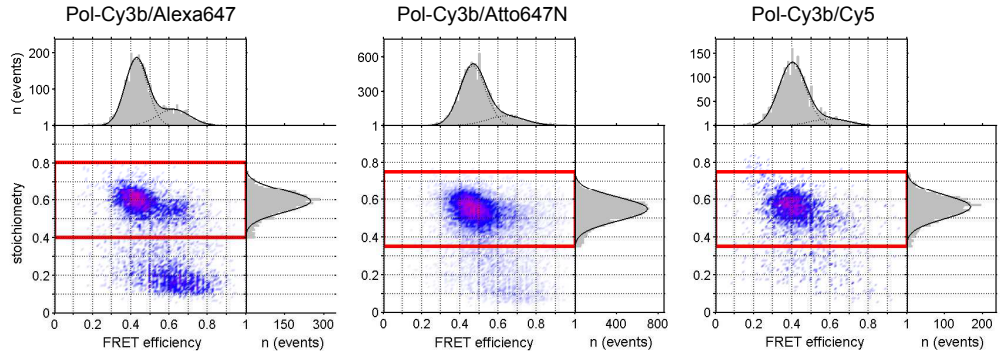


	Green dye	Red dye
Pol-Cy3b/Alexa647	6.5 %	1.6 %
Pol-Cy3b/Atto647N	6.2 %	2.2 %
Pol-Cy3b/Cy5	7.0 %	1.4 %

Figure 8.2: *SDS-PAGE analysis of doubly-labelled Pol samples.* ***(a)*** *Green and* ***(b)*** *red in-gel fluorescence of the three samples. Pol is denoted with an arrowhead, and dye contaminants with an asterisk.* ***(c)*** *Coomassie-staining of the gel, confirming the identity of protein bands.* ***(d)*** *Dye contamination as a percentage of total fluorescence in the sample, quantified from gel images in (a) and (b).*

In addition, we characterized the samples using confocal (ALEX) microscopy *in vitro*, to measure their FRET efficiency distributions. In the presence of DNA, Pol constructs showed two FRET populations, corresponding to the open and closed conformations, with the open conformation being more populated (Figure 8.3a). The mean FRET efficiencies varied from 0.40 to 0.47 for the open state, and from 0.60 to 0.64 for the closed state, likely due to the different linker lengths of the three acceptors and the different Förster radii of the dye pairs (Figure 8.3b). These values serve as references that can be used when evaluating FRET efficiencies observed in cells. In addition, our single-molecule sorting capability allowed us to extract the percentages of the doubly-labelled, donor-only and acceptor-only species. We found that 85% of molecules in the Pol-Cy3b/Atto647N sample were doubly labelled, whereas the percentage was lower for Pol-Cy3b/Alexa647 and Pol-Cy3b/Cy5, at 41% and 43%, respectively (Figure 8.3c).

a



b

	E^*_{open}	E^*_{closed}
Pol-Cy3b/Alexa647	0.43	0.62
Pol-Cy3b/Atto647N	0.47	0.64
Pol-Cy3b/Cy5	0.40	0.60

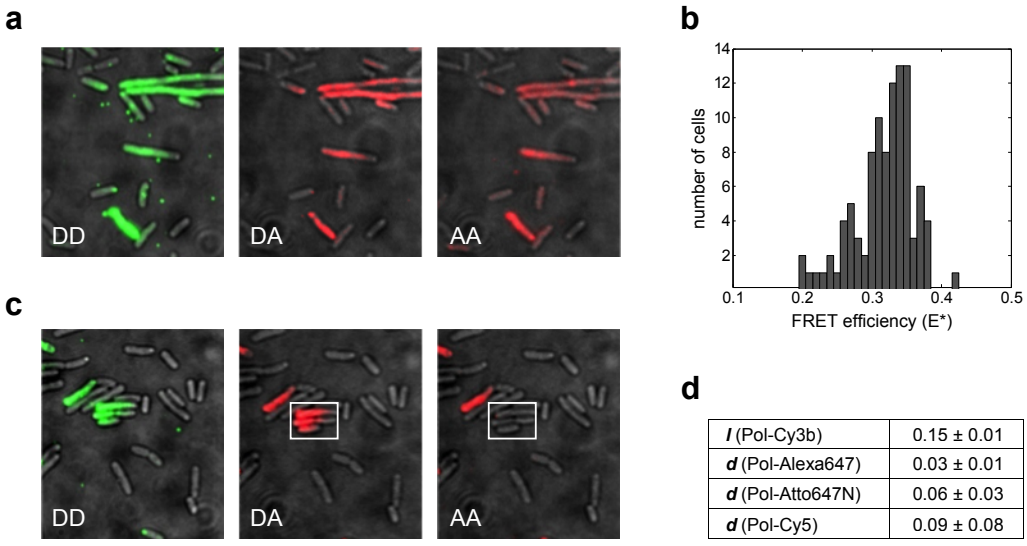
c

	D-only	DA	A-only
Pol-Cy3b/Alexa647	36%	41%	23%
Pol-Cy3b/Atto647N	7%	85%	8%
Pol-Cy3b/Cy5	52%	43%	5%

Figure 8.3: *Confocal analysis of doubly-labelled Pol samples.* ***(a)*** *Uncorrected E/S histograms of Pol-Cy3b/Alexa647, Pol-Cy3b/Atto647N and Pol-Cy3b/Cy5, in the presence of DNA, obtained using a red-burst search.* ***(b)*** *Mean FRET efficiencies of open and closed populations, obtained by Guassian fitting of histograms in (a).* ***(c)*** *Relative amounts of donor-only, doubly-labelled (DA) and acceptor-only species, estimated from E/S histograms obtained using an all-photon burst search.*

8.2.2 Single-cell measurements

We first internalized and imaged the doubly-labelled Pol samples at the single-cell level. ALEX microscopy was used to probe the donor (DD), acceptor (AA) and FRET channels (DA). With Pol-Cy3b/Alexa647, we experienced significant issues with non-internalized fluorescence, whereas cleaner internalizations could be obtained with Pol-Cy3b/Atto647N and Pol-Cy3b/Cy5. We observed relatively low internalization efficiencies, with most cells exhibiting intensities close to the level of autofluorescence, and hence we restricted our analysis to the low number of highly loaded cells. In the case of Pol-Cy3b/Cy5 (Figure 8.4a, b), we obtained a raw FRET efficiency histogram centred at ~0.35 FRET, lower than the mean efficiencies measured *in vitro*, although the sampling was too limited to allow reliable Gaussian fitting.



l (Pol-Cy3b)	0.15 ± 0.01
d (Pol-Alexa647)	0.03 ± 0.01
d (Pol-Atto647N)	0.06 ± 0.03
d (Pol-Cy5)	0.09 ± 0.08

Figure 8.4: *Single-cell FRET measurements of Pol.* ***(a)*** *Pol-Cy3b/Cy5 internalization. A rare example of a field of view with highly loaded cells is shown, with fluorescence from DD, DA and AA channels overlaid on the white-light image.* ***(b)*** *Single-cell FRET histogram for Pol-Cy3b/Cy5, calculated from cell-averaged DD and DA intensities of highly loaded cells.* ***(c)*** *Pol-Cy3b/Atto647N internalization. Cells exhibiting a DA but not an AA signal are shown boxed.* ***(d)*** *Estimated values of leakage and direct-excitation coefficients, obtained from average fluorescence intensities in DD, DA and AA channels, of cells electroporated with singly-labelled Pol. Electroporation with 1 μM protein at 1.4 kV voltage, widefield mode with alternating 532-nm excitation at 3.0 mW and 637-nm excitation at 1.5 mW power, 50-ms exposure.*

Curiously, in the case of Pol-Cy3b/Atto647N, some cells exhibited a strong DA but no AA signal (Figure 8.4c). In order to determine if this effect was due to the contributions of direct excitation of the acceptor, or the donor fluorescence leakage into the acceptor channel, we aimed to measure the leakage coefficient (l) and the direct-excitation coefficient (d) for our Pol constructs in cells. Therefore, we internalized singly-labelled Pol samples and calculated l and d from the ratios of intensities observed in the DD, DA and AA channels (see Section 8.5.4 for details). Both the leakage and the direct-excitation coefficient were similar to what would be expected *in vitro* (Figure 8.4d), and hence could not explain the high intensities observed in the DA channel. We opted not to proceed with the single-cell FRET analysis of Pol-Cy3b/Atto647N until this effect could be understood.

8.2.3 Single-molecule measurements

In order to overcome the limitations of single-cell experiments, we tried to detect single-molecule FRET by internalizing doubly-labelled Pol into *E. coli* at low concentration, and imaging it using continuous illumination or ALEX. We observed a very low number of PSFs in the DA channel, particularly in the case of Pol-Cy3b/Alexa647 and Pol-Cy3b/Cy5, despite the high number of molecules in the DD channel. The reasons for this effect could include the poor signal-to-noise ratio in the FRET channel, as well as the fast photobleaching of the acceptors, and indeed using lower laser powers slightly increased the probability of observing FRET events. Pol-Cy3b/Atto647N displayed higher brightness and photostability than the Alexa647 and Cy5 constructs, and showed more events in the DA channel, but similarly to Pol-Atto647 it was prone to non-specific binding to the cell membrane. Notably, the number of PSFs in the AA channel was also higher than in the DA channel, leading us to reason that the observed low relative number of DA events was likely also due to the presence of singly-labelled and free-dye species in the DD and AA channels.

We initially restricted our FRET analysis to immobile molecules, due to the limitations of our tracking software, and selected for time traces in which the DD and DA intensities showed anti-correlated behavior indicative of smFRET (Figure 8.5). Typically, only single

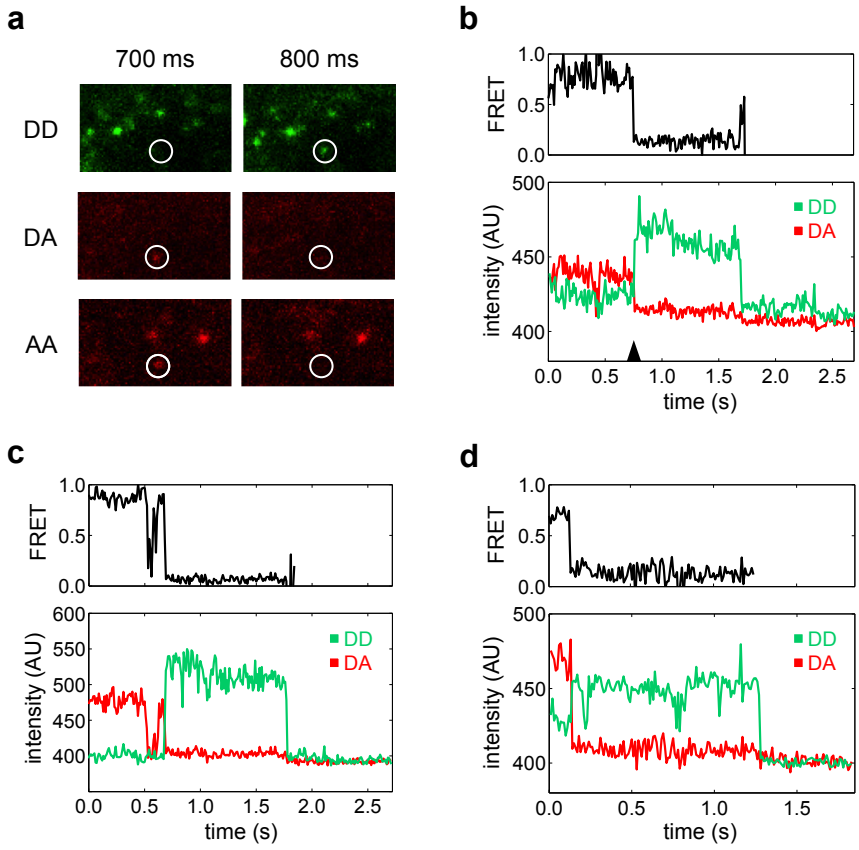


Figure 8.5: *Single-molecule FRET measurements of Pol in cells.* ***(a)*** *Example Pol-Cy3b/Alexa647 molecule undergoing an anti-correlated event. A frame before (700 ms) and a frame after the transition (800 ms) are shown for each of the three fluorescence channels, with the PSF of the molecule circled.* ***(b)*** *DA-intensity (red), AA-intensity (green) and FRET-efficiency (black) time traces corresponding to the event in (a). The time point of the transition is shown with an arrowhead.* ***(c-d)*** *Time traces for example anti-correlated events for (c) Pol-Cy3b/Atto647N and (d) Pol-Cy3b/Cy5. Electroporation with 500 nM protein at 1.0 kV voltage, HILO mode with (a, b) alternating 532-nm excitation at 1.0 mW and 637-nm excitation at 0.5 mW or (c, d) 532-nm excitation at 4.7 mW and 637-nm excitation at 3.0 mW, (a, b) 50-ms or (c, d) 10-ms exposure.*

anti-correlated events were observed in each time trace, suggesting that they were induced by acceptor photobleaching rather than by protein conformational changes. The raw FRET efficiencies differed significantly from molecule to molecule, and showed time-dependent variations, but were usually in the range of 0.7 to 0.9, significantly higher than their corresponding *in vitro* values (up to 0.64 for the closed state of Pol). Unfortunately, the number of observed events was not sufficient to allow reliable FRET histograms to be constructed.

8.2.4 Discussion

Internalization of doubly-labelled Pol has enabled preliminary single-cell and single-molecule FRET experiments *in vivo*. Single-cell experiments with Pol-Cy3b/Cy5 yielded a mean raw FRET efficiency value of ~0.35, lower than the values measured *in vitro* (0.40 and 0.60 for the open and closed states, respectively). This difference could be due to the effect of autofluorescence, and the presence of singly-labelled and free-dye species. SDS-PAGE analysis showed that the green-dye contamination was present at 6-7 %, which would be expected to shift the ensemble FRET efficiency towards lower values. In addition, *in vitro* confocal analysis indicated that ~50 % of the molecules were singly labelled, most likely corresponding to Pol species with an absent or inactive (photobleached) acceptor fluorophore. Although this percentage may be overestimated, due to the fact that photobleaching is partly induced by the confocal experiment itself, the presence of singly-labelled species is probably the main factor accounting for the skew in the FRET-efficiency distribution. It may be possible to account for these effects by estimating their contribution to the in-cell fluorescence, however, obtaining an accurate FRET signature of Pol in this way would be difficult.

The single-cell approach thus faces significant limitations, and an alternative approach is needed to distinguish the molecule of interest from the background of singly-labelled species. On the single-molecule level, we have observed anti-correlated events with all three doubly-labelled Pol samples. The measured raw FRET efficiencies were significantly higher than those measured *in vitro*, even assuming that the observed Pol molecules adopted the closed state. The physiological structure of the polymerase domain of Pol is unlikely to be grossly different from that observed *in vitro*, although the structure of the labelled and electroporated constructs could be affected, and could account for the unusual FRET signature. Alternatively, the observed difference between the *in vitro* and in-cell FRET efficiencies could arise from the effects of the cellular environment on the photophysics of the dyes. The number of FRET events was low, which could partly be explained by the poor signal-to-noise ratio (e.g. due to the lower quantum yield of acceptors in cells [231])

and by acceptor photobleaching. Unfortunately, the brightest and most photostable dye that we tested (Atto647N) is also significantly hydrophobic, and hence suboptimal for in-cell studies. In the future, it would be worth exploring alternative acceptor dyes that are as photostable as Atto647N but less hydrophobic, such as Abberior*635 [227].

Notably, both the single-cell and the single-molecule approaches could benefit from improvements in sample quality. Unfortunately, our FRET experiments were carried out before we realized the importance of dye contamination (Sections 7.3.2 and 7.5.2), and hence our doubly-labelled samples were not subjected to a thorough dye-purification treatment. However, the concentration of singly-labelled species (~50 %) extends significantly beyond that of the contaminating dye (6-7 %), meaning that the double-labelling protocol would also need to be improved to eliminate singly-labelled species. These improvements would allow single-cell experiments, and would also give a more balanced ratio of the numbers of events in the DA and DD channels at the single-molecule level. Pol could then be internalized at a higher concentration to maximize the number of FRET events, without raising the level of DD fluorescence to a level at which it would interfere with single-molecule tracking.

Ultimately, however, achieving a sufficiently high internalization of doubly-labelled Pol by electroporation may be difficult. Although singly-labelled Pol could be internalized at reasonable efficiencies, the efficiency of electroporation-based internalization appears to be affected by the attached dyes and their associated charges (Sections 7.3.1 and 7.3.2), which could be the reason for the poor internalization of doubly-labelled Pol. We therefore reasoned that Pol function would be more easily probed indirectly, by measuring single-molecule FRET within its DNA substrate, as we demonstrate in the next section.

8.3 Gapped-DNA bending

In Chapter 4, we established that Pol binds and significantly bends its gapped-DNA substrate. In addition, we observed a DNA species with an even higher bend angle, corresponding to a ternary complex of the DNA with a Pol dimer. In order to test whether DNA bending by Pol is also observed in the native environment of the living cell, and to examine the physiological relevance of the Pol_2-DNA species, we performed in-cell smFRET experiments. We used electroporation to internalize doubly-labelled DNA substrates into *E. coli*, and measured smFRET of the internalized diffusing molecules, probing their binding and bending by the endogenous full-length Pol (Figure 8.6a).

8.3.1 Probing for Pol-DNA species

We screened our doubly-labelled gapped-DNA library (Figure 4.3a) to find a construct that would give sufficiently different FRET signals when free and when bound by the endogenous polymerase. We selected construct T-12/T+8, which shows *in vitro* accurate FRET efficiencies of 0.46 when unbound, 0.73 in the Pol-DNA complex, and 0.92 in the Pol_2-DNA complex. We internalized this construct into *E. coli* cells in small numbers (1-10 molecules), and imaged the cells under HILO illumination. Both diffusing and immobile DNA molecules were observed in the DD and DA channels (Figure 8.6b). To measure single-molecule FRET, we tracked the molecules in the DA channel and correlated the measured intensities with those in the DD channel. The resulting FRET efficiency histogram across all frames and all molecules showed a clear bimodal distribution, with the two populations centred at E^*=0.40 and E^*=0.83. The lower-FRET population is in good agreement with the unbound DNA species *in vitro*, whereas the higher-FRET population could correspond either to the Pol monomer or the Pol dimer species *in vitro*, or to an unresolved ensemble of the two (Figure 8.6c, d). The bound population was present at 20% of the total number of detected molecules.

To test whether the observed FRET signature of the gapped-DNA construct was biologically relevant, we internalized a control DNA construct with a very similar separation

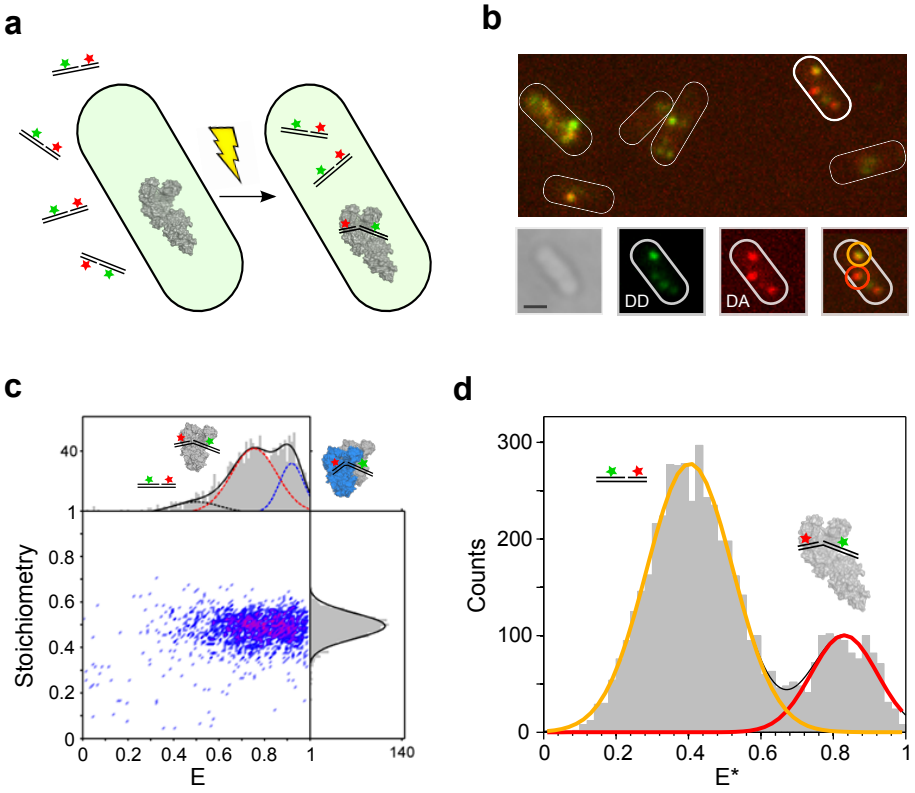


Figure 8.6: *Detection of gapped-DNA binding and bending by Pol.* ***(a)*** *Schematic of experiment that probes the binding of internalized doubly-labelled gapped DNA to endogenous polymerase.* ***(b)*** *Top, example field of view, shown as an overlay of DD (green) and DA (red) channels. Bottom, white-light image, DD- and DA-channel fluorescence, and the overlay shown for one example cell, highlighted in the large field of view above. Two example single molecules are circled.* ***(c)*** *Fully corrected E/S histogram of T-12/T+8 construct in vitro, in the presence of 3 nM Pol. The FRET-efficiency distribution is fitted with three Gaussians, corresponding to unbound DNA (black), Pol-DNA (red) and Pol_2-DNA species (blue), which are depicted schematically.* ***(d)*** *Apparent FRET-efficiency histogram of T-12/T+8 construct in cells. Two Gaussians are fitted, corresponding to unbound DNA (yellow) and full-length Pol-DNA species (red).*

between the labels (B-11/T+8), but with no gap in its non-template strand (Figure 8.7a, b). The unbound duplex DNA gave an accurate FRET efficiency of 0.46 *in vitro*, which was unaffected by the presence of Pol. The apparent FRET efficiency histogram of this construct showed a single population at E*=0.38 in cells, corresponding to the unbound DNA. The absence of a high-FRET population for this construct, which is not a substrate for the polymerase, is consistent with the interpretation that the high-FRET population observed with

the gapped-DNA construct is a result of bending induced by the endogenous full-length Pol. The good agreement between the *in vitro* and in-cell FRET efficiencies for this sample also suggests that the FRET species observed for the DNA constructs in cells can indeed be directly compared to those characterized *in vitro*.

8.3.2 Probing for Pol_2-DNA species

We further explored the existence of the Pol_2-DNA species in cells by selecting a DNA construct that would give sufficiently different FRET signals when bound by a single Pol or a Pol dimer (Figure 8.7c, d). We chose construct T-18/T+15, which gives *in vitro* accurate FRET efficiencies of 0.07, 0.44 and 0.83 for the unbound, the Pol-DNA and the Pol_2-DNA species, respectively. In cells, we observed a heterogeneous population at lower FRET values, which appeared as an ensemble of two species. We fitted this population with two Gaussians and obtained E^* values of 0.18 and 0.36, corresponding to the unbound and the full-length Pol monomer-bound DNA. The bound population was estimated to be present at 43% of all observed DNA molecules. Only very few events were detected in the FRET range corresponding to the Pol dimer species, suggesting that the Pol_2-DNA species may not exist or is of minor importance under physiological conditions.

8.3.3 Discussion

Our single-molecule experiments unequivocally show that gapped-DNA constructs are bent in live *E. coli*, unlike the duplex DNA. The close agreement between the FRET signatures of the bent species observed in cells and *in vitro* suggests that the bending is likely mediated by the endogenous full-length Pol binding, although the effect of other DNA-binding proteins cannot be excluded. To conclusively demonstrate that the observed bending was due to flPol binding, we constructed a Δ*polA* strain that was deficient in full-length Pol; however, the viability of this strain was too compromised to allow internalization by electroporation.

Although we see a good level of agreement between the *in vitro* and in-cell FRET effi-

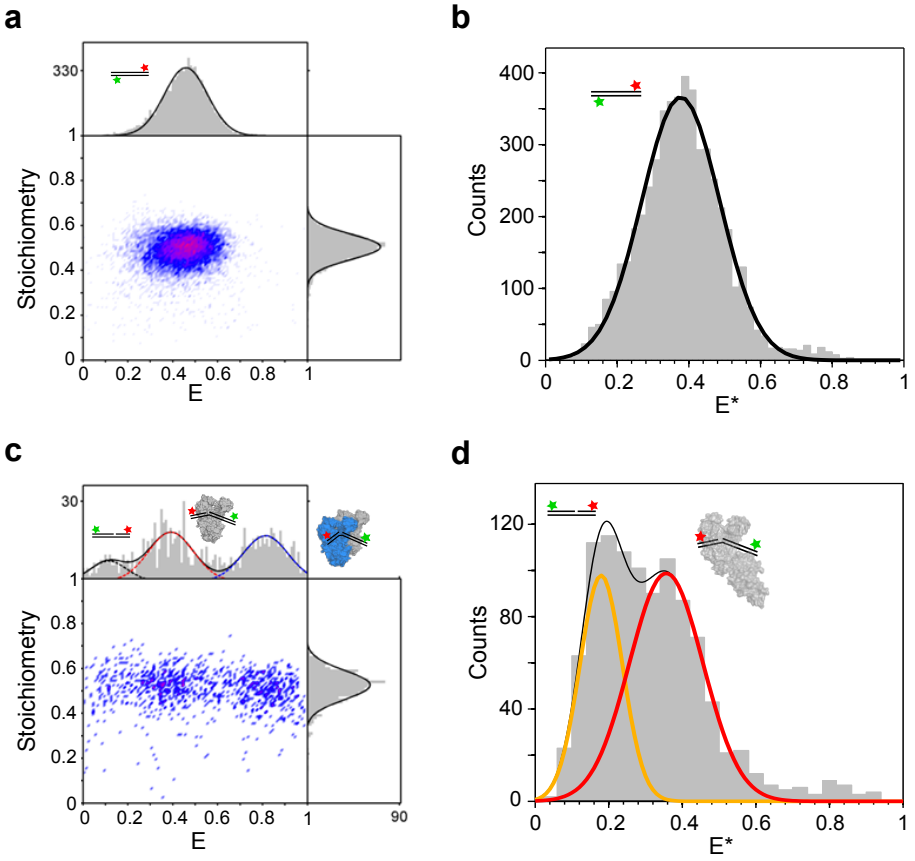


Figure 8.7: *Experiments with duplex-DNA control and gapped-DNA construct that probes for Pol_2-DNA species.* ***(a)*** *Fully corrected E/S histogram of B-11/T+8 duplex-DNA construct in vitro, in the absence of Pol.* ***(b)*** *Apparent FRET-efficiency histogram of the same construct in cells. In both (a) and (b), the only observed population corresponds to unbound DNA.* ***(c)*** *Fully corrected E/S histogram of T-18/T+15 gapped-DNA construct, in the presence of 3 nM Pol. Three Gaussians are fitted, corresponding to unbound DNA, Pol-DNA and Pol_2-DNA species.* ***(d)*** *Apparent FRET-efficiency histogram of the same construct in cells. Two Gaussians are fitted, corresponding to unbound DNA (yellow) and full-length Pol-DNA (red).*

ciencies of the unbound duplex DNA, the resolution of smFRET in cells is not sufficient to establish the identity of the higher-FRET species of construct T-12/T+8 with certainty. Experiments with construct B-18/T+15 do not show any significant population at high FRET, suggesting that only the Pol-DNA species is biologically relevant, and that the Pol_2-DNA species may be an artefact of the *in vitro* experimental conditions. As discussed in Section (4.6.2), Pol dimerization has so far only been reported *in vitro*, and whilst the observed dimer formation cannot be an effect of a high Pol concentration *in vitro*, the non-

physiological nature of the Pol (KF) construct or the *in vitro* buffer conditions could play a role.

Similarly, the higher-FRET populations of both constructs in cells are smaller than what is observed *in vitro*, despite the high estimated concentration of flPol in cells (~400 nM). The true populations are likely to be even smaller, since our in-cell FRET analysis tends to overestimate the relative amounts of higher-FRET species. This error arises from the fact that molecules are detected in the DA channel, and thus low-FRET species are more likely to be missed. The high abundance of low-FRET DNA molecules in cells may be due to the contributions from autofluorescence and the presence of donor-only species, resulting from duplex dehybridization and intracellular DNA degradation. Chemical protection of DNA ends, which has produced promising results in our laboratory (Plochowietz et al., in preparation), could be used to minimize damage to the constructs upon internalization. Other reasons for the apparent difference in the ratio of the bound and unbound DNA *in vitro* and in cells could include the different protein constructs involved (Pol vs. flPol), the effect of intracellular conditions, and the involvement of other proteins that could compete with flPol for gapped-DNA binding.

8.4 Conclusions and future work

Although we have been able to demonstrate both single-cell and single-molecule FRET imaging of Pol in cells, these preliminary experiments have not produced sufficient statistics to conclude whether the *in vitro* structure of Pol is preserved in cells. Direct imaging of Pol structure and its conformational states thus remains challenging, partly due to the molecular-size limit of electroporation-based internalization. Efficient smFRET imaging of smaller proteins has been achieved in our laboratory, suggesting that our imaging capabilities are not limiting smFRET detection of Pol in cells. We anticipate that advancements in the unnatural amino acid technology will soon provide a means of efficient and specific labelling of endogenous Pol, avoiding the need for its internalization and allowing the study of its structure and conformational dynamics in cells.

On the other hand, labelling, internalization and smFRET imaging of DNA constructs are highly efficient, and we show that probing conformational changes in the DNA substrate is an optimal approach for detecting Pol-DNA binding in cells by smFRET. In this way, we have demonstrated both the physiological relevance of gapped-DNA bending by Pol, as well as the likely artefactual nature of the Pol dimer species. These experiments exemplify the importance of *in vivo* studies, as they show that only a subset of molecular behaviours observed *in vitro* may be relevant in the context of the cell. The DNA-based approach for probing flPol binding in cells also has the scope for being extended to full structure determination of the gapped-DNA in complex with flPol. With an implementation of accurate FRET corrections in cells, a number of DNA-DNA distances could be obtained, and a docked structure *in vivo* calculated in the same way as *in vitro*. Similarly, other DNA constructs could be prepared and internalized to probe DNA binding by the 3'-5' endonuclease and the 5' exonuclease domains of flPol. Finally, if sufficient time resolution could be achieved, DNA bending could be used as a probe to measure real-time kinetics of DNA binding by flPol in cells.

8.5 Materials and methods

8.5.1 Sample preparation

Singly- and doubly-labelled Pol samples were previously prepared. Single labelling was done as described in reference [13], and double labelling was performed using the biased protocol in the presence of DNA, as described in Section 6.7.6. DNAs were previously synthesised and labelled (Section 4.8.1), and were annealed in a low-salt annealing buffer (20 mM Tris-HCl pH 8.0, 10-100 mM NaCl, 1 mM EDTA) for the purposes of electroporation.

8.5.2 *In vitro* characterization

In-gel fluorescence of Pol samples was analysed as in Section 7.7.7. Confocal analysis of Pol and DNA samples was done as in Section 4.8.2. Pol samples were analysed at 50-250 pM concentration, in the presence of 100 nM stem-loop DNA. 2-4 datasets of 10 min were recorded for each sample. The percentages of singly-labelled species were estimated by running an all-photon burst search with different green- and red-photon thresholds, and quantifying the resulting populations from E/S histograms. DNA samples were analysed at 100 pM concentration, in the presence or absence of 3 nM Pol (KF). 3-6 datasets of 10 min were recorded for each sample.

8.5.3 Internalization by electroporation

Pol samples were internalized using the protocol described in Section 7.7.2. In the case of DNA internalization, labelled DNA was diluted to 1 μM in water, and added to a 20-μl aliquot of cells to a final concentration of 12.5 nM. EDTA was added to the mix of cells and DNA at 500 μM concentration, to minimize electroporation-induced DNA dehybridization, and electroporation carried out at 1.4 kV. Cells were washed 5-times with PBS only; no salt or detergent treatment was performed.

8.5.4 FRET analysis of Pol

Cells were imaged in the widefield or HILO mode, using alternating 532-nm excitation at 1.0-5.0 mW power and 637-nm excitation at 0.5-3.0 mW power, with 10-50 ms exposure. In the case of single-cell FRET, cell intensities F_{DD}, F_{DA} and F_{AA} were quantified and normalized for the cell area, as before. Apparent FRET efficiency was calculated as $E^* = F_{DA}/(F_{DA} + F_{DD})$. The leakage coefficient (l) was calculated from the F_{DA} and F_{DD} intensities of cells electroporated with Pol-Cy3b, and the direct-excitation coefficient (d) was calculated from the F_{DA} and F_{AA} intensities of cells electroporated with Pol-Alexa647, Pol-Atto647N or Pol-Cy5. In the case of single-molecule FRET, movies were examined for immobile molecules that showed a signal in both the DA and AA channels, and intensities over time quantified in Fiji. Time traces of apparent FRET efficiency were constructed using a custom-written function in MATLAB; the apparent FRET efficiency was calculated as above.

8.5.5 FRET analysis of DNA

Cells were imaged in the HILO mode, using continuous-wave 532-nm excitation at 1.5 mW power, with 20-ms exposure. Single-molecule tracking was done in the DA channel, by adapting the MATLAB script used in Section 7.7.8. The DD channel was correlated with the DA channel using a transformation matrix, generated by mapping fluorescent-bead images recorded in the DD channel to those recorded in the DA channel. The PSFs that co-localized after DD/DA channel correlation were analysed for their intensities in the two channels, and apparent FRET efficiency calculated as above.

8.6 Contributions

- Parts of the introduction have been published in reference [254].
- Doubly-labelled Pol samples were previously prepared by Tim Craggs and Johannes Hohlbein.
- DNA constructs for single-molecule FRET experiments in cells were prepared and characterized *in vitro* by Tim Craggs.
- Single-molecule FRET analysis of internalized DNA was done by Anne Plochowietz.
- Figures 8.6 and 8.7 were prepared based on figure elements provided by Tim Craggs and Anne Plochowietz.

9
Concluding remarks

In this thesis, we have addressed a number of outstanding questions in the structure and function of DNA polymerase I. We have used single-molecule FRET in combination with rigid-body docking to solve the structure of Pol in complex with its gapped-DNA substrate, and determined the previously unknown position of downstream DNA. By further subjecting the model to all-atom molecular dynamics simulations, we have also measured the extent of unwinding in downstream DNA, probed the nature of Pol-DNA interactions and investigated the mechanism of strand-displacement synthesis. Based on the observed dynamics of the complex and that of the gapped-DNA substrate alone, we have proposed a model for structure-specific substrate recognition by Pol. We have also made progress towards understanding the stability determinants in Pol, with preliminary data suggesting that Pol conformational dynamics are highly sensitive to mutations, and that their stability is governed through a complex network of residues far apart in the protein structure. Finally, we have shown that whilst the observed Pol-DNA structure is preserved in cells, Pol dimerization is unlikely to be physiologically relevant.

Many of the biological questions addressed in this thesis are of general relevance. In addition to Pol being an excellent model for our understanding of all DNA polymerases, the mechanisms of sequence-independent substrate recognition likely apply across many families of DNA-processing enzymes, such as those involved in DNA repair and recombin-

ation [19]. Similarly, the use of EDM in elucidating the stability determinants in proteins has previously generated some general conclusions, such as the existence of the 'hotspots' of stabilization [184], and our application of EDM to the study of Pol extends this inquiry to complex, multi-domain proteins. Finally, probing Pol structure and dynamics in cells addresses the general question of whether biomolecular structure and dynamics are preserved *in vivo*.

Efforts to improve or implement existing methodologies for studying Pol structure and function have comprised a significant part of this thesis. Firstly, we have established the protocols for expression, purification and fluorescent labelling of unnatural amino acid-tagged Pol variants, an achievement that will benefit a number of projects relying on orthogonal labelling of Pol. Secondly, we have optimized the electroporation-based protocol for internalization of Pol and of proteins generally. Particularly for small-and medium-sized proteins not prone to aggregation, this development will enable single-molecule tracking at long time scales inaccessible to fluorescent-protein fusions, enabling more detailed studies of protein cellular dynamics and function. Thirdly, we have provided one of the first examples of single-molecule FRET imaging of proteins in cells, which will allow protein structure and conformational dynamics to be probed *in vivo*, particularly when combined with the unnatural amino acid labelling approaches [254].

We have also demonstrated a successful synthesis of *in vitro*, *in silico* and *in vivo* approaches in studying macromolecular structure and conformational dynamics. We believe that these three levels of investigation can complement and benefit each other in a variety of ways, addressing questions not amenable to any one approach alone. Perhaps the best example of this synthesis in the thesis is provided by our method of Pol-DNA structure determination. In this case, *in vitro* data (the crystal structure and the single-molecule measurements) were used for initial model generation *in silico*. The model was used to design new labelling positions and collect additional *in vitro* data, which in turn yielded a significantly improved *in silico* model. Furthermore, subjecting the model to *in silico* (molecular dynamics) simulations provided novel information on the degree of unwinding in downstream DNA, which was subsequently tested *in vitro* by quenchable FRET assays. Ad-

ditionally, we were able to test our *in vitro* / *in silico* model of Pol-DNA structure *in vivo*, thus extending the proposed mechanisms to the full-length protein and to the complex environment of the living cell.

Generally, we have seen good agreement between *in vitro* and *in silico* approaches, and between the specific methods used. The analysis of the gapped-DNA substrate dynamics showed very close agreement between single-molecule FRET data and coarse-grained simulations, and the results of atomistic simulations further corroborate these results. Similarly, the observed DNA unpairing is consistent across all the methods used to probe it, including rigid-body docking, molecular dynamics simulations and quenchable FRET assays. Further, our preliminary single-molecule FRET data on the effect of substitutions on Pol conformational equilibrium have confirmed the predictions from EDM analysis, although additional work is required to discern the effects of mutations in more detail. The *in vitro* and *in vivo* approaches also show good agreement, with Pol diffusion in cells being consistent with previous reports and with what would be expected based on its *in vitro* structure [15], and with the Pol-DNA complex giving very similar FRET signatures *in vitro* and in cells. However, our *in vivo* analysis did not confirm the existence of the Pol-DNA_2 suggesting that this species may not be relevant in the context of the full-length protein and/or in the environment of the living cell. This observation further strengthens the importance of *in vivo* studies, since whilst many *in vitro* behaviours can be reproduced *in vivo*, some are likely to prove irrelevant or artefactual.

In summary, we have addressed some of the gaps in our knowledge of the structure and conformational dynamics of DNA polymerase I. We have also contributed to the development of methodologies that we hope will serve as useful tools in the future studies of Pol and other proteins, particularly in the context of unnatural amino acid labelling and in-cell FRET. We have explored the different ways in which *in vitro*, *in silico* and *in vivo* approaches can be combined to provide an accurate and comprehensive picture of biomolecular structure and function, and we look forward to seeing the on-going synthesis of these approaches in the future.

Bibliography

1. Friedberg, E. C. The eureka enzyme: the discovery of DNA polymerase. *Nat. Rev. Mol. Cell Biol.* **7,** 143–7 (2006).
2. Joyce, C. M. & Steitz, T. A. Function and structure relationships in DNA polymerases. *Annu. Rev. Biochem.* **63,** 777–822 (1994).
3. Patel, P. H., Suzuki, M., Adman, E., Shinkai, A. & Loeb, L. A. Prokaryotic DNA polymerase I: evolution, structure, and "base flipping" mechanism for nucleotide selection. *J. Mol. Biol.* **308,** 823–37 (2001).
4. Kunkel, T. A. & Bebenek, K. DNA Replication Fidelity. *Annu. Rev. Biochem.* **69,** 497–529 (2000).
5. Hohlbein, J., Aigrain, L., Craggs, T. D., Bermek, O., Potapova, O., Shoolizadeh, P., Grindley, N. D. F., Joyce, C. M. & Kapanidis, A. N. Conformational landscapes of DNA polymerase I and mutator derivatives establish fidelity checkpoints for nucleotide insertion. *Nat. Commun.* **4,** 2131 (2013).
6. Evans, G. W., Hohlbein, J., Craggs, T., Aigrain, L. & Kapanidis, A. N. Real-time single-molecule studies of the motions of DNA polymerase fingers illuminate DNA synthesis mechanisms. *Nucleic Acids Res.* **43,** 5998–6008 (2015).
7. Van den Bedem, H. & Fraser, J. S. Integrative, dynamic structural biology at atomic resolution—it's about time. *Nat. Methods* **12,** 307–318 (2015).
8. Johnson, S. J., Taylor, J. S. & Beese, L. S. Processive DNA synthesis observed in a polymerase crystal suggests a mechanism for the prevention of frameshift mutations. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 3895–3900 (2003).
9. Turner, R. M., Grindley, N. D. F. & Joyce, C. M. Interaction of DNA Polymerase I (Klenow Fragment) with the Single-Stranded Template beyond the Site of Synthesis. *Biochemistry* **42,** 2373–2385 (2003).
10. Singh, K., Srivastava, A., Patel, S. S. & Modak, M. J. Participation of the fingers subdomain of Escherichia coli DNA polymerase I in the strand displacement synthesis of DNA. *J. Biol. Chem.* **282,** 10594–604 (2007).
11. Garvie, C. W. & Wolberger, C. Recognition of specific DNA sequences. *Mol. Cell* **8,** 937–46 (2001).
12. Li, Y., Korolev, S. & Waksman, G. Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of Thermus aquaticus DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.* **17,** 7514–7525 (1998).

13. Santoso, Y., Joyce, C. M., Potapova, O., Le Reste, L., Hohlbein, J., Torella, J. P., Grindley, N. D. & Kapanidis, A. N. Conformational transitions in DNA polymerase I revealed by single-molecule FRET. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 715–720 (2010).

14. Xie, X. S., Yu, J. & Yang, W. Y. Living cells as test tubes. *Science* **312,** 228–230 (2006).

15. Uphoff, S., Reyes-Lamothe, R., Garza de Leon, F., Sherratt, D. J. & Kapanidis, A. N. Single-molecule DNA repair in live bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 8063–8068 (2013).

16. Joubert, A. M., Byrd, A. S. & LiCata, V. J. Global conformations, hydrodynamics, and X-ray scattering properties of Taq and Escherichia coli DNA polymerases in solution. *J. Biol. Chem.* **278,** 25341–25347 (2003).

17. Hübscher, U., Spadari, S., Villani, G. & Maga, G. *DNA Polymerases: Discovery, Characterization and Functions in Cellular DNA Transactions* (World Scientific, 2010).

18. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. *Molecular Biology of the Cell* 4th ed. (Garland Science, 2002).

19. Friedberg, E. C., Walker, G. C., Siede, W. & Wood, R. D. *DNA repair and mutagenesis* 2nd ed. (ASM Press, 2005).

20. Lyamichev, V., Brow, M. A. & Dahlberg, J. E. Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science* **260,** 778–83 (1993).

21. Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G. & Steitz, T. A. Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. *Nature* **313,** 762–766 (1985).

22. Freemont, P. S., Friedman, J. M., Beese, L. S., Sanderson, M. R. & Steitz, T. A. Cocrystal structure of an editing complex of Klenow fragment with DNA. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 8924–8928 (1988).

23. Beese, L. S., Derbyshire, V. & Steitz, T. A. Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Science* **260,** 352–355 (1993).

24. Beese, L. S., Friedman, J. M. & Steitz, T. A. Crystal structures of the Klenow fragment of DNA polymerase I complexed with deoxynucleoside triphosphate and pyrophosphate. *Biochemistry* **32,** 14095–101 (1993).

25. Kiefer, J. R., Mao, C., Hansen, C. J., Basehore, S. L., Hogrefe, H. H., Braman, J. C. & Beese, L. S. Crystal structure of a thermostable Bacillus DNA polymerase I large fragment at 2.1 A resolution. *Structure* **5,** 95–108 (1997).

26. Kiefer, J. R., Mao, C., Braman, J. C. & Beese, L. S. Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. *Nature* **391,** 304–307 (1998).

27. Steitz, T. A. DNA polymerases: structural diversity and common mechanisms. *J. Biol. Chem.* **274,** 17395–8 (1999).

28. Kim, Y., Eom, S. H., Wang, J., Lee, D. S., Suh, S. W. & Steitz, T. A. Crystal structure of Thermus aquaticus DNA polymerase. *Nature* **376,** 612–6 (1995).

29. Aliotta, J. M., Pelletier, J. J., Ware, J. L., Moran, L. S., Benner, J. S. & Kong, H. Thermostable Bst DNA polymerase I lacks a 3'–>5' proofreading exonuclease activity. *Genet. Anal. Eng.* **12,** 185–95 (1996).

30. Eom, S. H., Wang, J. & Steitz, T. A. Structure of Taq polymerase with DNA at the polymerase active site. *Nature* **382,** 278–281 (1996).

31. Murali, R., Sharkey, D. J., Daiss, J. L. & Murthy, H. M. K. Crystal structure of Taq DNA polymerase in complex with an inhibitory Fab: The Fab is directed against an intermediate in the helix-coil dynamics of the enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 12562–12567 (1998).

32. Srivastava, A., Singh, K. & Modak, M. J. Phe 771 of Escherichia coli DNA polymerase I (Klenow fragment) is the major site for the interaction with the template overhang and the stabilization of the pre-polymerase ternary complex. *Biochemistry* **42,** 3645–54 (2003).

33. Wu, E. Y. & Beese, L. S. The structure of a high fidelity DNA polymerase bound to a mismatched nucleotide reveals an "ajar" intermediate conformation in the nucleotide selection mechanism. *J. Biol. Chem.* **286,** 19758–19767 (2011).

34. Torella, J. P., Holden, S. J., Santoso, Y., Hohlbein, J. & Kapanidis, A. N. Identifying molecular dynamics in single-molecule FRET experiments with burst variance analysis. *Biophys. J.* **100,** 1568–77 (2011).

35. Golosov, A. A., Warren, J. J., Beese, L. S. & Karplus, M. The mechanism of the translocation step in DNA replication by DNA polymerase I: a computer simulation analysis. *Structure* **18,** 83–93 (2010).

36. Miller, B. R., Parish, C. A. & Wu, E. Y. Molecular dynamics study of the opening mechanism for DNA polymerase I. *PLoS Comput. Biol.* **10,** e1003961 (2014).

37. Minnick, D. T., Astatke, M., Joyce, C. M. & Kunkel, T. A. A thumb subdomain mutant of the large fragment of Escherichia coli DNA polymerase I with reduced DNA binding affinity, processivity, and frameshift fidelity. *J. Biol. Chem.* **271,** 24954–61 (1996).

38. Kunkel, T. A. DNA replication fidelity. *J. Biol. Chem.* **279,** 16895–8 (2004).

39. Minnick, D. T., Bebenek, K., Osheroff, W. P., Turner, R. M., Astatke, M., Liu, L., Kunkel, T. A. & Joyce, C. M. Side chains that influence fidelity at the polymerase active site of Escherichia coli DNA polymerase I (Klenow fragment). *J. Biol. Chem.* **274,** 3067–75 (1999).

40. Petruska, J., Goodman, M. F., Boosalis, M. S., Sowers, L. C., Cheong, C. & Tinoco, I. Comparison between DNA melting thermodynamics and DNA polymerase fidelity. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 6252–6 (1988).

41. Joyce, C. M., Potapova, O., Delucia, A. M., Huang, X., Basu, V. P. & Grindley, N. D. Fingers-closing and other rapid conformational changes in DNA polymerase I (Klenow fragment) and their role in nucleotide selectivity. *Biochemistry* **47,** 6103–6116 (2008).

42. Doublié, S., Sawaya, M. R. & Ellenberger, T. An open and closed case for all polymerases. *Structure* **7,** R31–5 (1999).

43. Brautigam, C. A. & Steitz, T. A. Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes. *Curr. Opin. Struct. Biol.* **8,** 54–63 (1998).

44. Dahlberg, M. E. & Benkovic, S. J. Kinetic mechanism of DNA polymerase I (Klenow fragment): identification of a second conformational change and evaluation of the internal equilibrium constant. *Biochemistry* **30,** 4835–4843 (1991).

45. Maier, B., Bensimon, D. & Croquette, V. Replication by a single DNA polymerase of a stretched single-stranded DNA. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 12002–7 (2000).

46. Schwartz, J. J. & Quake, S. R. Single molecule measurement of the "speed limit" of DNA polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 20294–20299 (2009).

47. Yin, Y. W. & Steitz, T. A. The structural mechanism of translocation and helicase activity in T7 RNA polymerase. *Cell* **116,** 393–404 (2004).

48. Johnson, K. A. Conformational coupling in DNA polymerase fidelity. *Annu. Rev. Biochem.* **62,** 685–713 (1993).

49. De Lucia, P. & Cairns, J. Isolation of an E. coli strain with a mutation affecting DNA polymerase. *Nature* **224,** 1164–6 (1969).

50. Bates, H., Randall, S. K., Rayssiguier, C., Bridges, B. A., Goodman, M. F. & Radman, M. Spontaneous and UV-induced mutations in Escherichia coli K-12 strains with altered or absent DNA polymerase I. *J. Bacteriol.* **171,** 2480–4 (1989).

51. Sharma, R. C. & Smith, K. C. Role of DNA polymerase I in postreplication repair: a reexamination with Escherichia coli delta polA. *J. Bacteriol.* **169,** 4559–64 (1987).

52. Lakowicz, J. R. *Principles of fluorescence spectroscopy* 3rd ed. (Springer, 2006).

53. Huang, B., Babcock, H. & Zhuang, X. Breaking the diffraction barrier: super-resolution imaging of cells. *Cell* **143,** 1047–58 (2010).

54. Lord, S. J., Lee, H. L. & Moerner, W. E. Single-molecule spectroscopy and imaging of biomolecules in living cells. *Anal. Chem.* **82,** 2192–2203 (2010).

55. Xie, X. S., Choi, P. J., Li, G. W., Lee, N. K. & Lia, G. Single-molecule approach to molecular biology in living bacterial cells. *Annu. Rev. Biophys.* **37,** 417–444 (2008).

56. Stracy, M., Uphoff, S., Garza de Leon, F. & Kapanidis, A. N. In vivo single-molecule imaging of bacterial DNA replication, transcription, and repair. *FEBS Lett.* **588,** 3585–94 (2014).

57. Hwang, L. C., Hohlbein, J., Holden, S. J. & Kapanidis, A. N. *Single-Molecule FRET: Methods and Biological Applications.* In *Handbook of Single-Molecule Biophysics* (Springer, 2009).

58. Nadeau, J. *Introduction to Experimental Biophysics: Biological Methods for Physical Scientists* (CRC Press, 2012).

59. Stryer, L. Fluorescence Energy Transfer as a Spectroscopic Ruler. *Annu. Rev. Biochem.* **47,** 819–846 (1978).

60. Clegg, R. M. Fluorescence resonance energy transfer. *Curr. Opin. Biotechnol.* **6,** 103–10 (1995).

61. Clegg, R. M. Fluorescence resonance energy transfer and nucleic acids. *Methods Enzymol.* **211,** 353–388 (1992).

62. Kapanidis, A. N. & Strick, T. Biology, one molecule at a time. *Trends Biochem. Sci.* **34,** 234–243 (2009).

63. Deniz, A. A., Mukhopadhyay, S. & Lemke, E. A. Single-molecule biophysics: at the interface of biology, physics and chemistry. *J. R. Soc. Interface* **5,** 15–45 (2008).

64. Ha, T. Single-molecule methods leap ahead. *Nat. Methods* **11,** 1015–8 (2014).

65. Hohlbein, J., Craggs, T. D. & Cordes, T. Alternating-laser excitation: single-molecule FRET and beyond. *Chem. Soc. Rev.* **43,** 1156–71 (2014).

66. Claxton, N. S., Fellers, T. J. & Davidson, M. W. *Laser scanning confocal microscopy* <http://www.olympusconfocal.com/theory/LSCMIntro.pdf> (2006).

67. Nie, S., Chiu, D. T. & Zare, R. N. Probing individual molecules with confocal fluorescence microscopy. *Science* **266,** 1018–21 (1994).

68. Schneckenburger, H. Total internal reflection fluorescence microscopy: technical innovations and novel applications. *Curr. Opin. Biotechnol.* **16,** 13–18 (2005).

69. Ha, T., Enderle, T., Ogletree, D. F., Chemla, D. S., Selvin, P. R. & Weiss, S. Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 6264–6268 (1996).

70. Kapanidis, A. N., Lee, N. K., Laurence, T. A., Doose, S., Margeat, E. & Weiss, S. Fluorescence-aided molecule sorting: analysis of structure and interactions by alternating laser excitation of single molecules. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 8936–8941 (2004).

71. Lee, N. K., Kapanidis, A. N., Wang, Y., Michalet, X., Mukhopadhyay, J., Ebright, R. H. & Weiss, S. Accurate FRET measurements within single diffusing biomolecules using alternating laser excitation. *Biophys. J.* **88,** 2939–2953 (2005).

72. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450,** 964–72 (2007).

73. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9,** 646–52 (2002).

74. Karplus, M. & Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 6679–85 (2005).

75. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.* **9,** 71 (2011).

76. McGeagh, J. D., Ranaghan, K. E. & Mulholland, A. J. Protein dynamics and enzyme catalysis: Insights from simulations. *BBA Proteins Proteom.* **1814,** 1077–1092 (2011).

77. Stansfeld, P. J. & Sansom, M. S. P. Molecular simulation approaches to membrane proteins. *Structure* **19,** 1562–72 (2011).

78. Ercolessi, F. *A molecular dynamics primer* <http://www.fisica.uniud.it/~ercolessi/md/md> (1997).

79. Hernández, E. R., Zetina, L. M. M., Vega, G. T., Rocha, M. G., Ochoa, L. F. R. & Fernandez, R. L. *Molecular Dynamics: from basic techniques to applications (A Molecular Dynamics Primer).* In *AIP Conf. Proc.* in. **1077** (AIP, 2008), 95–123.

80. Price, N. C., Dwek, R. A., Ratcliffe, R. G. & Wormald, M. *Principles and Problems in Physical Chemistry for Biochemists* 3rd (Oxford University Press, 2001).

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

82. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D. & Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102,** 3586–616 (1998).

83. Oostenbrink, C., Villa, A., Mark, A. E. & van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **25,** 1656–76 (2004).

84. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118,** 11225–11236 (1996).

85. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17,** 490–519 (1996).

86. Baker, C. M. Polarizable force fields for molecular dynamics simulations of biomolecules. *WIREs Comput. Mol. Sci.* **5,** 241–254 (2015).

87. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65,** 712–25 (2006).

88. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O. & Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78,** 1950–8 (2010).

89. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **26,** 1668–88 (2005).

90. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4,** 435–447 (2008).

91. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L. & Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26,** 1781–802 (2005).

92. Jensen, F. *Introduction to Computational Chemistry* (Wiley, 2007).

93. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089 (1993).

94. Jorgensen, W. L. & Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 6665–70 (2005).

95. Berendsen, H., Postma, J., van Gunsteren, W. & Hermans, J. *Interaction Models for Water in Relation to Protein Hydration.* In *Intermolecular Forces* (Springer Netherlands, 1981).

96. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91,** 6269–6271 (1987).

97. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926 (1983).

98. Jorgensen, W. L. & Madura, J. D. Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Mol. Phys.* **56,** 1381–1392 (1985).

99. Yu, H. & van Gunsteren, W. F. Accounting for polarization in molecular simulation. *Comput. Phys. Commun.* **172,** 69–85 (2005).

100. Chen, J., Brooks, C. L. & Khandogin, J. Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.* **18,** 140–8 (2008).

101. Feig, M. & Brooks, C. L. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **14,** 217–24 (2004).

102. Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **15,** 144–50 (2005).

103. Saunders, M. G. & Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **42,** 73–93 (2013).

104. Riniker, S., Allison, J. R. & van Gunsteren, W. F. On developing coarse-grained models for biomolecular simulation: a review. *Phys. Chem. Chem. Phys.* **14,** 12423–30 (2012).

105. Hockney, R. W. & Eastwood, J. W. *Computer Simulation Using Particles* (McGraw-Hill, 1981).

106. Standard, J. M. *Energy Minimization Methods* <`http://chemistry.illinoisstate.edu/standard/che38037/handouts/380.37emin.pdf`> (2015).

107. Walton, E. B. & Vanvliet, K. J. Equilibration of experimentally determined protein structures for molecular dynamics simulation. *Phys. Rev. E* **74,** 061901 (2006).

108. Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **72,** 2384 (1980).

109. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684 (1984).

110. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **52,** 255–268 (2006).

111. Parrinello, M. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182 (1981).

112. Schlick, T. Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules. *F1000 Biol. Rep.* **1,** 51 (2009).

113. Schlitter, J., Engels, M. & Krüger, P. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graph.* **12,** 84–9 (1994).

114. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120,** 11919–29 (2004).

115. Sawaya, M. R., Prasad, R., Wilson, S. H., Kraut, J. & Pelletier, H. Crystal structures of human DNA polymerase beta complexed with gapped and nicked DNA: evidence for an induced fit mechanism. *Biochemistry* **36,** 11205–15 (1997).

116. Beard, W. A. & Wilson, S. H. Structure and Mechanism of DNA Polymerase β. *Biochemistry* **53,** 2768–2780 (2014).

117. Craggs, T. D. & Kapanidis, A. N. Six steps closer to FRET-driven structural biology. *Nat. Methods* **9,** 1157–1158 (2012).

118. Muschielok, A., Andrecka, J., Jawhari, A., Brückner, F., Cramer, P. & Michaelis, J. A nano-positioning system for macromolecular structural analysis. *Nat. Methods* **5,** 965–971 (2008).

119. Brunger, A. T., Strop, P., Vrljic, M., Chu, S. & Weninger, K. R. Three-dimensional molecular modeling with single molecule FRET. *J. Struct. Biol.* **173,** 497–505 (2011).

120. Kalinin, S., Peulen, T., Sindbert, S., Rothwell, P. J., Berger, S., Restle, T., Goody, R. S., Gohlke, H. & Seidel, C. A. A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nat. Methods* **9,** 1218–1225 (2012).

121. Sabir, T., Schröder, G. F., Toulmin, A., McGlynn, P. & Magennis, S. W. Global structure of forked DNA in solution revealed by high-resolution single-molecule FRET. *J. Am. Chem. Soc.* **133,** 1188–1191 (2011).

122. Choi, U. B., Strop, P., Vrljic, M., Chu, S., Brunger, A. T. & Weninger, K. R. Single-molecule FRET-derived model of the synaptotagmin 1-SNARE fusion complex. *Nat. Struct. Mol. Biol.* **17,** 318–324 (2010).

123. Sisamakis, E., Valeri, A., Kalinin, S., Rothwell, P. J. & Seidel, C. A. M. Accurate single-molecule FRET studies using multiparameter fluorescence detection. *Methods Enzymol.* **475,** 455–514 (2010).

124. Antonik, M., Felekyan, S., Gaiduk, A. & Seidel, C. A. M. Separating structural heterogeneities from stochastic variations in fluorescence resonance energy transfer distributions via photon distribution analysis. *J. Phys. Chem. B* **110,** 6970–8 (2006).

125. Sindbert, S., Kalinin, S., Nguyen, H., Kienzler, A., Clima, L., Bannwarth, W., Appel, B., Müller, S. & Seidel, C. A. M. Accurate distance determination of nucleic acids via Förster resonance energy transfer: implications of dye linker length and rigidity. *J. Am. Chem. Soc.* **133,** 2463–80 (2011).

126. Ouldridge, T. E., Louis, A. A. & Doye, J. P. K. Structural, mechanical, and thermodynamic properties of a coarse-grained DNA model. *J. Chem. Phys.* **134** (2011).

127. Doye, J. P. K., Ouldridge, T. E., Louis, A. A., Romano, F., Šulc, P., Matek, C., Snodin, B. E. K., Rovigatti, L., Schreck, J. S., Harrison, R. M. & Smith, W. P. J. Coarse-graining DNA for simulations of DNA nanotechnology. *Phys. Chem. Chem. Phys.* **15,** 20395–414 (2013).

128. Šulc, P., Romano, F., Ouldridge, T. E., Rovigatti, L., Doye, J. P. K. & Louis, A. A. Sequence-dependent thermodynamics of a coarse-grained DNA model. *J. Chem. Phys.* **137,** 135101 (2012).

129. Snodin, B. E. K., Randisi, F., Mosayebi, M., Šulc, P., Schreck, J. S., Romano, F., Ouldridge, T. E., Tsukanov, R., Nir, E., Louis, A. A. & Doye, J. P. K. Introducing improved structural properties and salt dependence into a coarse-grained model of DNA. *J. Chem. Phys.* **142,** 234901 (2015).

130. Markiewicz, R. P., Vrtis, K. B., Rueda, D. & Romano, L. J. Single-molecule microscopy reveals new insights into nucleotide selection by DNA polymerase I. *Nucleic Acids Res.* **40,** 7975–84 (2012).

131. Yuan, Y. C., Whitson, R. H., Liu, Q., Itakura, K. & Chen, Y. A novel DNA-binding motif shares structural homology to DNA replication and repair nucleases and polymerases. *Nat. Struct. Biol.* **5,** 959–964 (1998).

132. Maitra, M., Gudzelak, A., Li, S.-X., Matsumoto, Y., Eckert, K. A., Jager, J. & Sweasy, J. B. Threonine 79 is a hinge residue that governs the fidelity of DNA polymerase beta by helping to position the DNA within the active site. *J. Biol. Chem.* **277,** 35550–60 (2002).

133. Knight, J. L., Mekler, V., Mukhopadhyay, J., Ebright, R. H. & Levy, R. M. Distance-restrained docking of rifampicin and rifamycin SV to RNA polymerase using systematic FRET measurements: developing benchmarks of model quality and reliability. *Biophys. J.* **88,** 925–938 (2005).

134. Bailey, M. F., Van Der Schans, E. J. C. & Millar, D. P. Dimerization of the Klenow fragment of Escherichia coli DNA polymerase I is linked to its mode of DNA binding. *Biochemistry* **46,** 8085–8099 (2007).

135. Tang, K.-H. & Tsai, M.-D. Structure and function of 2:1 DNA polymerase-DNA complexes. *J. Cell. Physiol.* **216,** 315–320 (2008).

136. Junker, H.-D., Hoehn, S. T., Bunt, R. C., Marathius, V., Chen, J., Turner, C. J. & Stubbe, J. Synthesis, characterization and solution structure of tethered oligonucleotides containing an internal 3'-phosphoglycolate, 5'-phosphate gapped lesion. *Nucleic Acids Res.* **30,** 5497–508 (2002).

137. Roll, C., Ketterlé, C., Faibis, V., Fazakerley, G. V. & Boulard, Y. Conformations of nicked and gapped DNA structures by NMR and molecular dynamic simulations in water. *Biochemistry* **37,** 4059–70 (1998).

138. Sharma, M., Predeus, A. V., Mukherjee, S. & Feig, M. DNA bending propensity in the presence of base mismatches: Implications for DNA repair. *J. Phys. Chem. B* **117,** 6194–6205 (2013).

139. Xu, Y., Grindley, N. D. & Joyce, C. M. Coordination between the polymerase and 5'-nuclease components of DNA polymerase I of Escherichia coli. *J. Biol. Chem.* **275,** 20949–55 (2000).

140. Doose, S., Heilemann, M., Michalet, X., Weiss, S. & Kapanidis, A. N. Periodic acceptor excitation spectroscopy of single molecules. *Eur. Biophys. J.* **36,** 669–674 (2007).

141. Selvin, P. R. & Ha, T. *Single-molecule techniques : a laboratory manual* (Cold Spring Harbor Laboratory Press, 2008).

142. Würth, C., Grabolle, M., Pauli, J., Spieles, M. & Resch-Genger, U. Relative and absolute determination of fluorescence quantum yields of transparent samples. *Nat. Protoc.* **8,** 1535–50 (2013).

143. Magde, D., Wong, R. & Seybold, P. G. Fluorescence quantum yields and their relation to lifetimes of rhodamine 6G and fluorescein in nine solvents: improved absolute standards for quantum yields. *Photochem. Photobiol.* **75,** 327–334 (2002).

144. *Fluorescent labels and dyes* <`http://www.atto-tec.com`> (2015).

145. Van Dijk, M. & Bonvin, A. M. J. J. 3D-DART: a DNA structure modelling server. *Nucleic Acids Res.* **37,** W235–9 (2009).

146. Lee, J. & Kim, S. H. PDB Editor: A user-friendly Java-based Protein Data Bank file editor with a GUI. *Acta Crystallogr. D* **65,** 399–402 (2009).

147. Schreck, J. S., Ouldridge, T. E., Romano, F., Louis, A. A. & Doye, J. P. K. Characterizing the bending and flexibility induced by bulges in DNA duplexes. *J. Chem. Phys.* **142,** 165101 (2015).

148. Cheatham, T. E. & Kollman, P. A. Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* **51,** 435–71 (2000).

149. Maffeo, C., Yoo, J., Comer, J., Wells, D. B., Luan, B. & Aksimentiev, A. Close encounters with DNA. *J. Phys. Condens. Matter* **26,** 413101 (2014).

150. Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A. & Orozco, M. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92,** 3817–29 (2007).

151. Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M. & Mackerell, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J. Chem. Theory Comput.* **8,** 3257–3273 (2012).

152. Pasi, M., Maddocks, J. H., Beveridge, D., Bishop, T. C., Case, D. A., Cheatham, T., Dans, P. D., Jayaram, B, Lankas, F., Laughton, C., Mitchell, J., Osman, R., Orozco, M., Pérez, A., Petkevičiūtė, D., Spackova, N., Sponer, J., Zakrzewska, K. & Lavery, R. μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* **42,** 12272–83 (2014).

153. Guy, A. T., Piggot, T. J. & Khalid, S. Single-stranded DNA within nanopores: conformational dynamics and implications for sequencing; a molecular dynamics simulation study. *Biophys. J.* **103,** 1028–36 (2012).

154. Cheatham, T. E. & Case, D. A. Twenty-five years of nucleic acid simulations. *Biopolymers* **99,** 969–77 (2013).

155. Norberg, J. & Nilsson, L. On the truncation of long-range electrostatic interactions in DNA. *Biophys. J.* **79,** 1537–53 (2000).

156. Cheatham, T. E. I., Miller, J. L., Fox, T., Darden, T. A. & Kollman, P. A. Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins. *J. Am. Chem. Soc.* **117,** 4193–4194 (1995).

157. Auffinger, P., Cheatham, T. E. & Vaiana, A. C. Spontaneous Formation of KCl Aggregates in Biomolecular Simulations: A Force Field Issue? *J. Chem. Theory Comput.* **3,** 1851–1859 (2007).

158. Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112,** 9020–41 (2008).

159. Luo, Y. & Roux, B. Simulation of Osmotic Pressure in Concentrated Aqueous Salt Solutions. *J. Phys. Chem. Lett.* **1,** 183–189 (2010).

160. Jiao, D., King, C., Grossfield, A., Darden, T. A. & Ren, P. Simulation of Ca2+ and Mg2+ solvation using polarizable atomic multipole potential. *J. Phys. Chem. B* **110,** 18553–9 (2006).

161. Callahan, K. M., Casillas-Ituarte, N. N., Roeselová, M., Allen, H. C. & Tobias, D. J. Solvation of magnesium dication: molecular dynamics simulation and vibrational spectroscopic study of magnesium chloride in aqueous solutions. *J. Phys. Chem. A* **114,** 5141–8 (2010).

162. Allnér, O., Nilsson, L. & Villa, A. Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations. *J. Chem. Theory Comput.* **8,** 1493–1502 (2012).

163. Chocholousová, J. & Feig, M. Implicit solvent simulations of DNA and DNA-protein complexes: agreement with explicit solvent vs experiment. *J. Phys. Chem. B* **110,** 17240–51 (2006).

164. Villa, E., Balaeff, A., Mahadevan, L. & Schulten, K. Multiscale Method for Simulating Protein-DNA Complexes. *Multiscale Model. Simul.* **2,** 527–553 (2006).

165. Eriksson, M. A., Härd, T. & Nilsson, L. Molecular dynamics simulations of the glucocorticoid receptor DNA-binding domain in complex with DNA and free in solution. *Biophys. J.* **68,** 402–26 (1995).

166. Pan, Y. & Nussinov, R. Structural basis for p53 binding-induced DNA bending. *J. Biol. Chem.* **282,** 691–9 (2007).

167. Fuxreiter, M., Mezei, M., Simon, I. & Osman, R. Interfacial water as a "hydration fingerprint" in the noncognate complex of BamHI. *Biophys. J.* **89,** 903–11 (2005).

168. Furini, S., Barbini, P. & Domene, C. DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. *Nucleic Acids Res.* **41,** 3963–72 (2013).

169. Habtemariam, B., Anisimov, V. M. & MacKerell, A. D. Cooperative binding of DNA and CBFbeta to the Runt domain of the CBFalpha studied via MD simulations. *Nucleic Acids Res.* **33,** 4212–22 (2005).

170. Arora, K. & Schlick, T. In silico evidence for DNA polymerase-beta's substrate-induced conformational change. *Biophys. J.* **87,** 3088–99 (2004).

171. Sampoli Benítez, B. A., Arora, K. & Schlick, T. In silico studies of the African swine fever virus DNA polymerase X support an induced-fit mechanism. *Biophys. J.* **90,** 42–56 (2006).

172. Li, Y. & Schlick, T. "Gate-keeper" residues and active-site rearrangements in DNA polymerase μ help discriminate non-cognate nucleotides. *PLoS Comput. Biol.* **9,** e1003074 (2013).

173. Li, Y. & Schlick, T. Modeling DNA Polymerase μ Motions: Subtle Transitions before Chemistry. *Biophys. J.* **99,** 3463–3472 (2010).

174. Florián, J., Goodman, M. F. & Warshel, A. Theoretical Investigation of the Binding Free Energies and Key Substrate-Recognition Components of the Replication Fidelity of Human DNA Polymerase β. *J. Phys. Chem. B* **106,** 5739–5753 (2002).

175. Florián, J., Warshel, A. & Goodman, M. F. Molecular Dynamics Free-Energy Simulations of the Binding Contribution to the Fidelity of T7 DNA Polymerase. *J. Phys. Chem. B* **106,** 5754–5760 (2002).

176. Oelschlaeger, P., Klahn, M., Beard, W. A., Wilson, S. H. & Warshel, A. Magnesium-cationic dummy atom molecules enhance representation of DNA polymerase beta in molecular dynamics simulations: improved accuracy in studies of structural features and mutational effects. *J. Mol. Biol.* **366,** 687–701 (2007).

177. Rittenhouse, R. C., Apostoluk, W. K., Miller, J. H. & Straatsma, T. P. Characterization of the active site of DNA polymerase beta by molecular dynamics and quantum chemical calculation. *Proteins* **53,** 667–82 (2003).

178. Thompson, E. H. Z., Bailey, M. F., van der Schans, E. J. C., Joyce, C. M. & Millar, D. P. Determinants of DNA mismatch recognition within the polymerase domain of the Klenow fragment. *Biochemistry* **41,** 713–22 (2002).

179. Vriend, G. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8,** 52–6, 29 (1990).

180. Chinea, G., Padron, G., Hooft, R. W., Sander, C. & Vriend, G. The use of position-specific rotamers in model building by homology. *Proteins* **23,** 415–21 (1995).

181. Bryce, R. A. *AMBER parameter database* <`http://www.pharmacy.manchester.ac.uk/bryce/amber`> (2015).

182. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 014101 (2007).

183. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–8, 27–8 (1996).

184. Tiana, G., Simona, F., De Mori, G. M. S., Broglia, R. A. & Colombo, G. Understanding the determinants of stability and folding of small globular proteins from their energetics. *Protein Sci.* **13,** 113–124 (2004).

185. Dill, K. A. Dominant forces in protein folding. *Biochemistry* **29,** 7133–7155 (1990).

186. Pace, C. N. Polar Group Burial Contributes More to Protein Stability than Nonpolar Group Burial. *Biochemistry* **40,** 310–313 (2001).

187. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222,** 67–88 (1991).

188. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J. Mol. Biol.* **240,** 421–33 (1994).

189. Genoni, A., Morra, G. & Colombo, G. Identification of domains in protein structures from the analysis of intramolecular interactions. *J. Phys. Chem. B* **116,** 3331–3343 (2012).

190. Ragona, L., Colombo, G., Catalano, M. & Molinari, H. Determinants of protein stability and folding: comparative analysis of beta-lactoglobulins and liver basic fatty acid binding protein. *Proteins* **61,** 366–76 (2005).

191. Colacino, S., Tiana, G. & Colombo, G. Similar folds with different stabilization mechanisms: the cases of Prion and Doppel proteins. *BMC Struct. Biol.* **6,** 17 (2006).

192. Morra, G. & Colombo, G. Relationship between energy distribution and fold stability: Insights from molecular dynamics simulations of native and mutant proteins. *Proteins* **72,** 660–72 (2008).

193. Genoni, A., Morra, G., Merz, K. M. & Colombo, G. Computational study of the resistance shown by the subtype B/HIV-1 protease to currently known inhibitors. *Biochemistry* **49,** 4283–4295 (2010).

194. Ratner, V., Kahana, E., Eichler, M. & Haas, E. A general strategy for site-specific double labeling of globular proteins for kinetic FRET studies. *Bioconjugate Chem.* **13,** 1163–1170 (2002).

195. Puljung, M. C. & Zagotta, W. N. Labeling of specific cysteines in proteins using reversible metal protection. *Biophys. J.* **100,** 2513–2521 (2011).

196. Wang, L., Xie, J. & Schultz, P. G. Expanding the genetic code. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 225–49 (2006).

197. Wang, L. Expanding the genetic code. *Science* **302,** 584–5 (2003).

198. Liu, C. C. & Schultz, P. G. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* **79,** 413–444 (2010).

199. Lajoie, M. J., Rovner, A. J., Goodman, D. B., Aerni, H.-R., Haimovich, A. D., Kuznetsov, G., Mercer, J. A., Wang, H. H., Carr, P. A., Mosberg, J. A., Rohland, N., Schultz, P. G., Jacobson, J. M., Rinehart, J., Church, G. M. & Isaacs, F. J. Genomically recoded organisms expand biological functions. *Science* **342,** 357–60 (2013).

200. Lang, K. & Chin, J. W. Cellular incorporation of unnatural amino acids and bioorthogonal labeling of proteins. *Chem. Rev.* **114,** 4764–806 (2014).

201. Agard, N. J., Prescher, J. A. & Bertozzi, C. R. A strain-promoted [3 + 2] azide-alkyne cycloaddition for covalent modification of biomolecules in living systems. *J. Am. Chem. Soc.* **126,** 15046–7 (2004).

202. Plass, T., Milles, S., Koehler, C., Schultz, C. & Lemke, E. A. Genetically encoded copper-free click chemistry. *Angew. Chem. Int. Ed.* **50,** 3878–81 (2011).

203. Reddington, S. C., Tippmann, E. M. & Jones, D. D. Residue choice defines efficiency and influence of bioorthogonal protein modification via genetically encoded strain promoted Click chemistry. *Chem. Commun.* **48,** 8419–8421 (2012).

204. Papworth, C., Bauer, J. C. & Braman, J. Site-directed mutagenesis in one day with >80% efficiency. *Strategies* **9,** 3–4 (1996).

205. Brustad, E. M., Lemke, E. A., Schultz, P. G. & Deniz, A. A. A general and efficient method for the site-specific dual-labeling of proteins for single molecule fluorescence resonance energy transfer. *J. Am. Chem. Soc.* **130,** 17664–17665 (2008).

206. Kim, J., Seo, M. H., Lee, S., Cho, K., Yang, A., Woo, K., Kim, H. S. & Park, H. S. Simple and efficient strategy for site-specific dual labeling of proteins for single-molecule fluorescence resonance energy transfer analysis. *Anal. Chem.* **85,** 1468–1474 (2013).

207. Guzman, L. M., Belin, D, Carson, M. J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177,** 4121–30 (1995).

208. *AMBER Tools 12 manual* <http://ambermd.org/doc12/AmberTools12.pdf> (2015).

209. Pierce, L. C. T., Salomon-Ferrer, R., Augusto F de Oliveira, C., McCammon, J. A. & Walker, R. C. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *J. Chem. Theory Comput.* **8,** 2997–3002 (2012).

210. Meli, M., Pagano, K., Ragona, L. & Colombo, G. Investigating the dynamic aspects of drug-protein recognition through a combination of MD and NMR analyses: implications for the development of protein-protein interaction inhibitors. *PLoS One* **9,** e97153 (2014).

211. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–8 (2007).

212. Joyce, C. M. & Derbyshire, V. Purification of Escherichia coli DNA polymerase I and Klenow fragment. *Methods Enzymol.* **262,** 3–13 (1995).

213. *QuikChange Primer Design* <http://www.genomics.agilent.com/primerDesignProgram.jsp> (2015).

214. *GE Healthcare Life Sciences* <http://www.gelifesciences.com> (2015).

215. Ellis, R. J. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.* **26,** 597–604 (2001).

216. Zhou, H. X., Rivas, G. & Minton, A. P. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* **37,** 375–397 (2008).

217. Tsien, R. Y. The green fluorescent protein. *Annu. Rev. Biochem.* **67,** 509–544 (1998).

218. Shaner, N. C., Steinbach, P. A. & Tsien, R. Y. A guide to choosing fluorescent proteins. *Nat. Methods* **2,** 905–909 (2005).

219. Betzig, E., Patterson, G. H., Sougrat, R., Lindwasser, O. W., Olenych, S., Bonifacino, J. S., Davidson, M. W., Lippincott-Schwartz, J. & Hess, H. F. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313,** 1642–5 (2006).

220. Billinton, N. & Knight, A. W. Seeing the wood through the trees: a review of techniques for distinguishing green fluorescent protein from endogenous autofluorescence. *Anal. Biochem.* **291,** 175–197 (2001).

221. Yang, L., Zhou, Y., Zhu, S., Huang, T., Wu, L. & Yan, X. Detection and quantification of bacterial autofluorescence at the single-cell level by a laboratory-built high-sensitivity flow cytometer. *Anal. Chem.* **84,** 1526–1532 (2012).

222. Dixit, R. & Cyr, R. Cell damage and reactive oxygen species production induced by fluorescence microscopy: effect on mitosis and guidelines for non-invasive fluorescence microscopy. *Plant J.* **36,** 280–290 (2003).

223. Dempsey, G. T., Vaughan, J. C., Chen, K. H., Bates, M. & Zhuang, X. Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. *Nat. Methods* **8,** 1027–1036 (2011).

224. Plochowietz, A., Crawford, R. & Kapanidis, A. N. Characterization of organic fluorophores for in vivo FRET studies based on electroporated molecules. *Phys. Chem. Chem. Phys.* **16,** 12688–12694 (2014).

225. Gust, A., Zander, A., Gietl, A., Holzmeister, P., Schulz, S., Lalkens, B., Tinnefeld, P. & Grohmann, D. A starting point for fluorescence-based single-molecule measurements in biomolecular research. *Molecules* **19,** 15824–15865 (2014).

226. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3,** 793–5 (2006).

227. Wurm, C. A., Kolmakov, K., Göttfert, F., Ta, H., Bossi, M., Schill, H., Berning, S., Jakobs, S., Donnert, G., Belov, V. N. & Hell, S. W. Novel red fluorophores with superior performance in STED microscopy. *Opt. Nanoscopy* **1,** 7 (2012).

228. Royant, A. & Noirclerc-Savoye, M. Stabilizing role of glutamic acid 222 in the structure of Enhanced Green Fluorescent Protein. *J. Struct. Biol.* **174,** 385–390 (2011).

229. Crawford, R., Torella, J. P., Aigrain, L., Plochowietz, A., Gryte, K., Uphoff, S. & Kapanidis, A. N. Long-lived intracellular single-molecule fluorescence using electroporated molecules. *Biophys. J.* **105,** 2439–2450 (2013).

230. Hinner, M. J. & Johnsson, K. How to obtain labeled proteins and what to do with them. *Curr. Opin. Biotechnol.* **21,** 766–776 (2010).

231. Fessl, T., Adamec, F., Polívka, T., Foldynová-Trantírková, S., Vácha, F. & Trantírek, L. Towards characterization of DNA structure under physiological conditions in vivo at the single-molecule level using single-pair FRET. *Nucleic Acids Res.* **40,** e121 (2012).

232. Dower, W. J., Miller, J. F. & Ragsdale, C. W. High efficiency transformation of E. coli by high voltage electroporation. *Nucleic Acids Res.* **16,** 6127–6145 (1988).

233. Taylor, D. L. & Wang, Y. L. Molecular cytochemistry: incorporation of fluorescently labeled actin into living cells. *Proc. Natl. Acad. Sci. U. S. A.* **75,** 857–861 (1978).

234. Clarke, M. S. & McNeil, P. L. Syringe loading introduces macromolecules into living mammalian cell cytosol. *J. Cell Sci.* **102,** 533–541 (1992).

235. McNeil, P. L., Murphy, R. F., Lanni, F. & Taylor, D. L. A method for incorporating macromolecules into adherent cells. *J. Cell Biol.* **98,** 1556–1564 (1984).

236. Plass, T., Milles, S., Koehler, C., Szymański, J., Mueller, R., Wiessler, M., Schultz, C. & Lemke, E. A. Amino acids for Diels-Alder reactions in living cells. *Angew. Chem. Int. Ed.* **51,** 4166–70 (2012).

237. Lang, K., Davis, L., Torres-Kolbus, J., Chou, C., Deiters, A. & Chin, J. W. Genetically encoded norbornene directs site-specific cellular protein labelling via a rapid bioorthogonal reaction. *Nat. Chem.* **4,** 298–304 (2012).

238. Lang, K., Davis, L., Wallace, S., Mahesh, M., Cox, D. J., Blackman, M. L., Fox, J. M. & Chin, J. W. Genetic Encoding of bicyclononynes and trans-cyclooctenes for site-specific protein labeling in vitro and in live mammalian cells via rapid fluorogenic Diels-Alder reactions. *J. Am. Chem. Soc.* **134,** 10317–20 (2012).

239. Carlson, J. C. T., Meimetis, L. G., Hilderbrand, S. A. & Weissleder, R. BODIPY-Tetrazine Derivatives as Superbright Bioorthogonal Turn-on Probes. *Angew. Chem. Int. Ed.* **52,** 6917–6920 (2013).

240. Shieh, P., Hangauer, M. J. & Bertozzi, C. R. Fluorogenic azidofluoresceins for biological imaging. *J. Am. Chem. Soc.* **134,** 17428–31 (2012).

241. Tokunaga, M., Imamoto, N. & Sakata-Sogawa, K. Highly inclined thin illumination enables clear single-molecule imaging in cells. *Nat. Methods* **5,** 159–161 (2008).

242. Berezin, M. Y. & Achilefu, S. Fluorescence lifetime measurements and biological imaging. *Chem. Rev.* **110,** 2641–2684 (2010).

243. López-Mirabal, H. R. & Winther, J. R. Redox characteristics of the eukaryotic cytosol. *Biochim. Biophys. Acta* **1783,** 629–640 (2008).

244. Weaver, J. C. & Chizmadzhev, Y. Theory of electroporation: A review. *Bioelectroch. Bioenerg.* **41,** 135–160 (1996).

245. Hillger, F., Nettels, D., Dorsch, S. & Schuler, B. Detection and analysis of protein aggregation with confocal single molecule fluorescence spectroscopy. *J. Fluoresc.* **17,** 759–765 (2007).

246. Puchalla, J., Krantz, K., Austin, R. & Rye, H. Burst analysis spectroscopy: a versatile single-particle approach for studying distributions of protein aggregates and fluorescent assemblies. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 14400–14405 (2008).

247. Kooijmans, S. A. A., Stremersch, S., Braeckmans, K., de Smedt, S. C., Hendrix, A., Wood, M. J. A., Schiffelers, R. M., Raemdonck, K. & Vader, P. Electroporation-induced siRNA precipitation obscures the efficiency of siRNA loading into extracellular vesicles. *J. Control. Release* **172,** 229–38 (2013).

248. Kim, Y., Ho, S. O., Gassman, N. R., Korlann, Y., Landorf, E. V., Collart, F. R. & Weiss, S. Efficient site-specific labeling of proteins via cysteines. *Bioconjugate Chem.* **19,** 786–791 (2008).

249. Aigrain, L., Sustarsic, M., Crawford, R., Plochowietz, A. & Kapanidis, A. N. Internalization and observation of fluorescent biomolecules in living microorganisms via electroporation. *J. Vis. Exp.* (2015).

250. Young, J. W., Locke, J. C., Altinok, A., Rosenfeld, N, Bacarian, T., Swain, P. S., Mjolsness, E. & Elowitz, M. B. Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nat. Protoc.* **7,** 80–88 (2012).

251. Holden, S. J., Uphoff, S., Hohlbein, J., Yadin, D., Le Reste, L., Britton, O. J. & Kapanidis, A. N. Defining the Limits of Single-Molecule FRET Resolution in TIRF Microscopy. *Biophys. J.* **99,** 3102–3111 (2010).

252. Crocker, J. C. & Grier, D. G. Methods of Digital Video Microscopy for Colloidal Studies. *J. Colloid Interface Sci.* **179,** 298–310 (1996).

253. Michalet, X. & Berglund, A. J. Optimal diffusion coefficient estimation in single-particle tracking. *Phys. Rev. E* **85** (2012).

254. Sustarsic, M. & Kapanidis, A. N. Taking the ruler to the jungle: single-molecule FRET for understanding biomolecular structure and dynamics in live cells. *Curr. Opin. Struct. Biol.* **34,** 52–59 (2015).

255. Sustarsic, M., Plochowietz, A., Aigrain, L., Yuzenkova, Y., Zenkin, N. & Kapanidis, A. Optimized delivery of fluorescently labeled proteins in live bacteria using electroporation. *Histochem. Cell Biol.* **142,** 113–124 (2014).

256. Miyawaki, A. Development of probes for cellular functions using fluorescent proteins and fluorescence resonance energy transfer. *Annu. Rev. Biochem.* **80,** 357–373 (2011).

257. Lam, A. J., St-Pierre, F., Gong, Y., Marshall, J. D., Cranfill, P. J., Baird, M. A., McKeown, M. R., Wiedenmann, J., Davidson, M. W., Schnitzer, M. J., Tsien, R. Y. & Lin, M. Z. Improving FRET dynamic range with bright green and red fluorescent proteins. *Nat. Methods* **9,** 1005–12 (2012).

258. Grünberg, R., Burnier, J. V., Ferrar, T., Beltran-Sastre, V., Stricher, F., van der Sloot, A. M., Garcia-Olivas, R., Mallabiabarrena, A., Sanjuan, X., Zimmermann, T. & Serrano, L. Engineering of weak helper interactions for high-efficiency FRET probes. *Nat. Methods* **10,** 1021–7 (2013).

259. Welch, C. M., Elliott, H., Danuser, G. & Hahn, K. M. Imaging the coordination of multiple signalling activities in living cells. *Nat. Rev. Mol. Cell Biol.* **12,** 749–756 (2011).

260. Frommer, W. B., Davidson, M. W. & Campbell, R. E. Genetically encoded biosensors based on engineered fluorescent proteins. *Chem. Soc. Rev.* **38,** 2833–2841 (2009).

261. Sakon, J. J. & Weninger, K. R. Detecting the conformation of individual proteins in live cells. *Nat. Methods* **7,** 203–205 (2010).

262. König, I., Zarrine-Afsar, A., Aznauryan, M., Soranno, A., Wunderlich, B., Dingfelder, F., Stüber, J. C., Plückthun, A., Nettels, D. & Schuler, B. Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells. *Nat. Methods* **12,** 773–9 (2015).

263. Sako, Y., Minoghchi, S. & Yanagida, T. Single-molecule imaging of EGFR signalling on the surface of living cells. *Nat. Cell Biol.* **2,** 168–172 (2000).

264. Murakoshi, H., Iino, R., Kobayashi, T., Fujiwara, T., Ohshima, C., Yoshimura, A. & Kusumi, A. Single-molecule imaging analysis of Ras activation in living cells. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 7317–7322 (2004).

265. Westhof, E. The amazing world of bacterial structured RNAs. *Genome Biol.* **11,** 108 (2010).