# Structural Studies on Flap Endonuclease complexes

T5 5'-Flap Exonuclease (T5FEN)
Human Flap Endonuclease (hFEN)
*Trypanosoma brucei* Flap Endonuclease (*T. brucei* FEN)

**Faizah Ahmed AlMalki**

Thesis Submitted for the Degree of Doctor of Philosophy (PhD)

Molecular Biology and Biotechnology

September-2014

# Abstract

Flap endonucleases (FENs) are structure-specific enzymes that play critical roles in DNA replication and repair. Three members of the FEN family have been investigated during this project in complexes with DNA substrates and metal ions: bacteriophage T5FEN, hFEN and *Trypanosoma brucei* FEN. T5FEN wild type and two catalytically inactive versions, D153K and D155K were successfully crystallized in complexes with DNA substrates containing 5' or 3' overhangs. The crystal structure for T5FEN-D153K in complex with a duplex containing 5' overhangs at each end and two Mg2+ ions was solved. The structure of T5FEN-D155K was solved in complex with a duplex containing 3' overhangs and a Ca2+ ion in the active site. In addition, wild type T5FEN was also crystallized with the same 3'-overhang substrate in the absence of metal ions. These structures revealed that the single strand-5' overhang in T5FEN-D153K pushed electrostatically and looped-up before threading through an arch-like structure composed of two helixes located over the active site of the enzyme. This arch is fully ordered in all of these structures. In the active sites of the variant complexes the Lys-153 and Lys155 are visible. The lysine's long side chain allows its ε amino group to occupy similar positions to the metal ion sites in one of the two active site subsites known as Cat1. The ε amino of Lys-153 directly coordinates the scissile phosphate of the DNA substrate. Important residues concerned with the 5' overhang threading are also determined. His-36 rotates by 180° to allow the duplex DNA movement before threading while Tyr-90 and Phe-105 form a gate-like structure after the 5' overhang has threaded. The conserved Arg-86 plays a critical role in 5' overhang transmission during threading process. Lys-83 and Arg-125 are found to interact symmetrically with the DNA backbone in the T5FEN-D155K:DNA complex. A new trans-arch/distal phosphate-binding site composed of Gly-70 and Lys-71 has been determined in the far side of the arch. These structures of T5FEN also have a binding site for the potassium ion within the H3TH motif coordinated by the main chain carbonyl oxygens of three residues and directly interacted with the DNA phosphate group.

# Acknowledgement

I would like to thanks my supervisor Prof.Peter Artymiuk for his support, guidance and explanation of crystallography work in the first years of my PhD journey. Many thanks to my second supervisor Dr.John Rafferty for his effort with me either by explanation or reading through my thesis chapters.

The grateful and thanks are continues to my third supervisor and collaborator Prof. Jon Sayers who supported me from the first time of my working with FEN enzymes. This thesis cannot be present in this way without him. His comments, additions and suggestions when he read through the chapters improved the quality of this thesis. When Prof.Sayers talk about FEN proteins he actually describe part of his life story with his FEN children in very amazing way. The most important thing in working with Prof.Sayers is his love for sharing the knowledge and experiences with other people.

I am also very grateful to Jing Zhang and Sarah Oates from the medical school whose collaborated with me during this project and provided me with the FEN proteins and constructs. A lot of thanks words to crystallography group in Sheffield University especially to Fiona Rodgers with her natural personality and sweet heart.

Many thanks to my sponsor Taif University and Saudi government with a special thanks to Saudi embassy and Saudi cultural bureau in London. Finally, I send a special gratitude to my parents who lived this dream with me day by day and my family for their love, encourage and support specifically my brother Abdullah AlMalki who supported me during my study.

# List of Contents

# List of Figures

# List of Tables

# Abbreviations and Symbols

### Species Abbreviations

| | |
|---|---|
| ***Af*** | *Archaeoglobus fulgidus* |
| ***B. subtilis*** | *Bacillus subtilis* |
| ***E. coli*** | *Escherichia coli* |
| ***Mj*** | *Methanococcus janaschii* |
| **T4** | Bacteriophage T4 |
| **T5** | Bacteriophage T5 |
| ***Taq*** | *Thermophilus aquaticus* |
| ***T.brucei*** | *Trypanosoma brucei* |

### Chemical Abbreviations

| | |
|---|---|
| **APS** | Ammonium persulphate |
| **$Ca^{2+}$** | Calcium ion |
| **DTT** | Dithiothreitol |
| **EDTA** | Ethylenediaminetetraacitic acid |
| **HEPES** | 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid |
| **IPTG** | Isopropyl-1-thio-$\beta$-d-galactopyranoside |
| **$K^{+}$** | Potassium ion |
| **MES** | 2-(N-morpholino)ethanesulfonic acid |
| **$Mg^{2+}$** | Magnesium ion |
| **$Mn^{2+}$** | Manganese ion |
| **$Na^{+}$** | Sodium ion |
| **PEG** | Poly ethylene glycol |
| **SDS** | Sodium dodecylsulphate |
| **$Sm^{3+}$** | Samarium ion |
| **$Zn^{2+}$** | Zinc ion |

### Miscellaneous Abbreviations

| | |
|---|---|
| **3D** | Three dimensional |
| **BLAST** | Basic Local Alignment Search Tool |
| **Cat1** | Catalytic metal ion binding site 1 |

| | |
|---|---|
| **Cat2** | Catalytic metal ion binding site 2 |
| **C$\alpha$** | Alpha carbon of amino acid |
| **CCP4** | Collaborative Computational Project No. 4 |
| **CCD** | Charge Coupled Device |
| **CFE** | Cell Free Extract |
| **DLS** | Diamond Light Source, Oxfordshire, UK |
| **DNA** | Deoxyribonucleic acid |
| **dA** | Deoxyadenosine |
| **dC** | Deoxycytidine |
| **dG** | Deoxyguanosine |
| **ds** | Double stranded |
| **dT** | Deoxythymidine |
| **ExoIX** | Exonuclease IX from *Escherichia coli* |
| **FEN** | Flap endonuclease |
| **GhG** | Glycine-hydrophobic-Glycine motif |
| **H3TH** | Helix-three-Turn-Helix |
| **hEXO1** | Exonuclease1 from human |
| **hFEN1** | Flap endonuclease from human |
| **HhH** | Helix-hairpin-Helix |
| **HTH** | Helix-Turn-Helix |
| **Ile** | Isoleucine |
| **JCSG** | Joint Centre for Structural Genomics |
| **LB** | Luria-Bertani media |
| **PAGE** | Polyacrylamide gel electrophoresis |
| **PCNA** | Proliferating Cell Nuclear Antigen |
| **PDB** | Protein DataBank |
| **RMSD** | Root mean square deviation |
| **RPA** | Replication Protein A |
| **WT** | Wild type |

**Units**

| | |
|---|---|
| **°** | Degree |
| **°C** | Degree Celcius |

| | |
|---|---|
| **C** | centimetre |
| **µ** | Micro |
| **Å** | Ångstrom |
| **Da** | Dalton |
| **g** | gram |
| **K** | Kelvin |
| **M** | Molar |
| **m** | Milli |
| **MW** | Molecular weight |
| **$OD_{600}$** | Optical density at 600 nm |
| **rpm** | Rotations per minute |
| **s** | Second |
| **S** | Siemens |
| **v/v** | Volume to volume |
| **v/w** | Volume to weight |
| **xg** | Acceleration of gravity |

**Crystallographic Terms**

| | |
|---|---|
| **a, b, c** | Real space unit cell dimensions |
| $\boldsymbol{\alpha, \beta, \gamma}$ | Real space unit cell angle |
| **AU** | Asymmetric unit |
| **B-factor** | Crystallographic temperature factor |
| **F** | Structure factor |
| **\|F\|** | Structure factor amplitude |
| $\mathbf{F_c}$ | Calculated structure factor |
| $\mathbf{F_o}$ | Observed structure factor |
| ***f*** | Atomic scattering factor |
| $\mathbf{F_{calc}}$ | Calculated structure factor |
| $\mathbf{\lvert F_{calc}\rvert}$ | Calculated structure factor amplitudes |
| **F(*hkl*)** | Structure factor for reflection *hkl* |
| **\| F(*hkl*)\|** | Structure factor amplitude for reflection *hkl* |
| ***h, k, l*** | Reciprocal lattice points |
| **I** | Diffraction intensity |

| | |
|---|---|
| $\boldsymbol{\lambda}$ | Wavelength |
| **LLG** | Log-likelihood gain |
| **MR** | Molecular replacement |
| $\boldsymbol{\rho(r)}$ | Electron density at position r |
| $\boldsymbol{\rho(xyz)}$ | Electron density at position xyz |
| $\mathbf{P}(\boldsymbol{uvw})$ | Patterson function |
| **R** | Crystallographic R-factor |
| $\mathbf{R_{free}}$ | R-factor for free set of reflection |
| **RFZ** | Rotation function Z-score |
| $\mathbf{R_{merge}}$ | R-factor relating agreement between symmetry related reflections |
| $\mathbf{R_{p.i.m}}$ | Multiplicity weighted merging R-factor |
| $\mathbf{R_{work}}$ | R-factor for working set of reflection |
| **TFZ** | Translation function Z-score |
| **u, v, w** | Patterson space coordinates |
| **V** | Volume of the unit cell |
| **x, y, z** | Real space coordinates |

**Dedicate to .....**

**my home**

**Kingdome of Saudi Arabia**

# Chapter 1: Introduction

In every living organisms DNA plays a fundamental role in determining the structure and function of the cell because hereditary information is stored as the sequence of bases along the DNA molecule. To transfer the genetic information from one generation to another it is necessary to replicate DNA with high fidelity and very few errors. In this chapter, the DNA replication mechanism from the model system *E. coli* will be introduced before dealing with the role of Flap Endonucleases (FENs) in DNA replication and repair processes, finally focusing on one member of this family from T5 phage.

## 1.1 DNA Replication

DNA replication is a very complex process that was first studied in bacteria and *E. coli* is the most studied organism (Kornberg & Baker, 1992). Deoxyribonucleic acid (DNA) is the genetic material that is composed of a deoxyribose sugar-phosphate backbone and millions of copies of four specific nitrogen-rich bases that are arranged in a double helical structure. The genetic information lies within the sequence of these bases. Organism's DNA exists in the form of chromosomes that need to be unwound and opened to expose the base sequence for replication and translation processes. Each single-strand then acts as a template to generate two identical copies of the original molecule. DNA strands have directionality and are antiparallel which means the two strands are synthesized in opposite directions.

### 1.1.1 Replication Machinery

In *E. coli* DNA presents as a short circular chromosome that is about 4.5 million base pairs long. The replication process starts by assembly of the initiator protein DnaA at specific locations called origins (*ori*C) found along the strands of the DNA molecule (Kolter & Helinski, 1979) and composed of 245 bp long of specific sequences (Oka et al, 1980). The binding of DnaA protein to the *ori*C site leads to melting of the duplex DNA to expose the

ssDNA followed by loading of hexameric helicase DanB from the complex with the helicase loader DnaC. This process results in the replisome activation through assembling of its components which include seven subunits, $\tau_2\gamma\delta\delta'\chi\psi$, that is known as DnaX complex. The complex binds to the helicase DnaB and Pol III core using $\tau$ subunit domains (Figure 1.1). These interactions are important for the replisome subunits combination and to join the leading and the lagging strands together during DNA duplication process which allow the helicase and the polymerase to move together in the same direction. The interaction of DnaB with the DnaG (DNA primase) direct the later one, DnaG, to the replication fork and continuously synthesises short, 10-20 nt of RNA primers which are deposited on the lagging strand and then extended by Pol III to form Okazaki fragments (Figure 1.1). In addition to the critical role of DnaX in the DNA duplication it is also responsible for positioning of the $\beta$ clamps around the primed DNA template which are acting as platforms for Pol III to elongate a new DNA fragment (Yeeles, 2014). Two or three Pol III core can be present in the replication fork (Figure 1.2) depending on the number of $\tau$ subunit in the DnaX complex (McInerney et al, 2007). Pol III molecules can replicate the two strands, leading and lagging, in the same time and more than one can work to extend Okazaki fragments on the lagging strand (Figure 1.2)

The nature of the replication process in *vitro* has confirmed seen as the semi-discontinuous model (Figure 1.3a) while in *vivo* is still as an ambiguous and elusive matter in how does this process happening? Is it following the semi-discontinuous or discontinuous model (Figure 1.3a)? Does priming accrue in both leading and lagging strand if the replication has been done discontinuous? A lot of experiments have been done in *vivo* to investigate this issue and look for persuasive answers for these questions. According to these studies the researchers divided into two groups. The first team support the semi-discontinuous replication in which the leading strand synthesises continuous while the lagging strand polymeriz in a form of short pieces ~ 1-2 kb. The second team is going with discontinuous replication and in this model both DNA strand synthesized discontinuous as short fragments which

ligated together to give new strands. More discussion and argument are available in (Yeeles, 2014) and (Amado & Kuzminov, 2013).

### 1.1.2 DNA Polymerases

The chromosomal replication mechanism in live cells requires enzymes to carry out the reactions which are known as polymerases. Multiple DNA polymerases are found in single cell. A human cell for example encodes at least 15 DNA-template-dependent DNA polymerases (Menezes & Sweasy, 2012) while the yeast cell has eight. *E. coli* produces five forms of DNA polymerases: Pol I, Pol II, Pol III, Pol IV and Pol V (Garcia-Diaz & Bebenek, 2007; Hastings et al, 2010). The major replicase enzyme in the DNA replication procedure is pol III while Pol I has a fidelity role in processing Okazaki fragments.

#### *1.1.2.1 DNA Polymerase III (Pol III)*

Replicative machines in bacteria are tripartite composed of a polymerase (pol III), a $\beta_2$ processivity factor and a DnaX complex which is known as replisome (Figure 1.3b) (McHenry, 2011). These three parts comprise the DNA Pol III holoenzyme (HE) which is a multisubunit protein complex that consists of 17 subunits with an overall composition $(\alpha\varepsilon\theta)_2(\beta_2)_2\tau_2\gamma\delta\delta'\chi\psi$ (O'Donnell, 2006). Pol III can be characterized by its very high processivity, catalytic potency and fidelity. This enzyme can bind to a DNA template and synthesize over 50 kilo bases per binding event (Yao et al, 2009) and up to 1000 nucleotides per second. In *E. coli* the Pol III core enzyme is an asymmetric dimer consisting of two $(\alpha\varepsilon\theta)$ subunits (Figure 1.3b). The $\alpha$ subunit is the DNA polymerase, the $\varepsilon$ subunit is the 3'→5' proofreading exonuclease and the $\theta$ subunit is an $\varepsilon$-stabilizing factor.

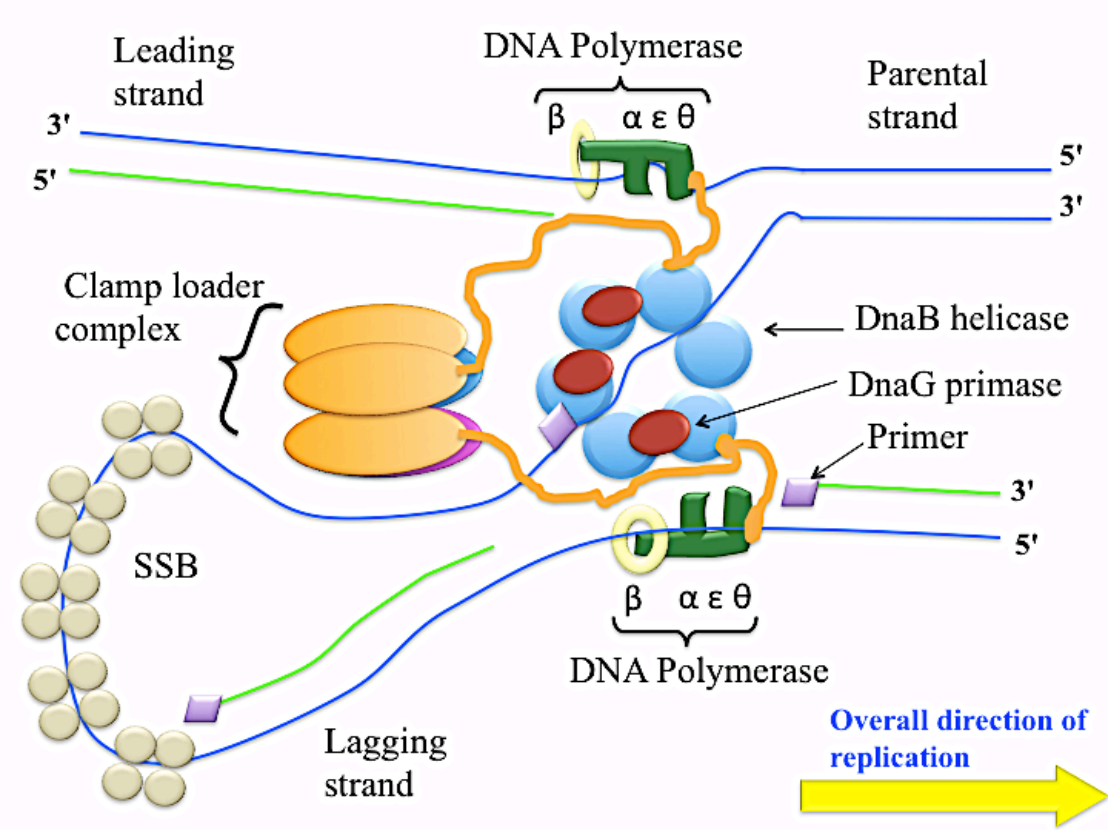



**Figure 1.1:** Diagram for DNA replication process.
It shows the generation of new strands from the parental strands and the DNA replication with the enzymes involved in this process as well as the replication fork. The figure was modified from (Tanner et al, 2008).

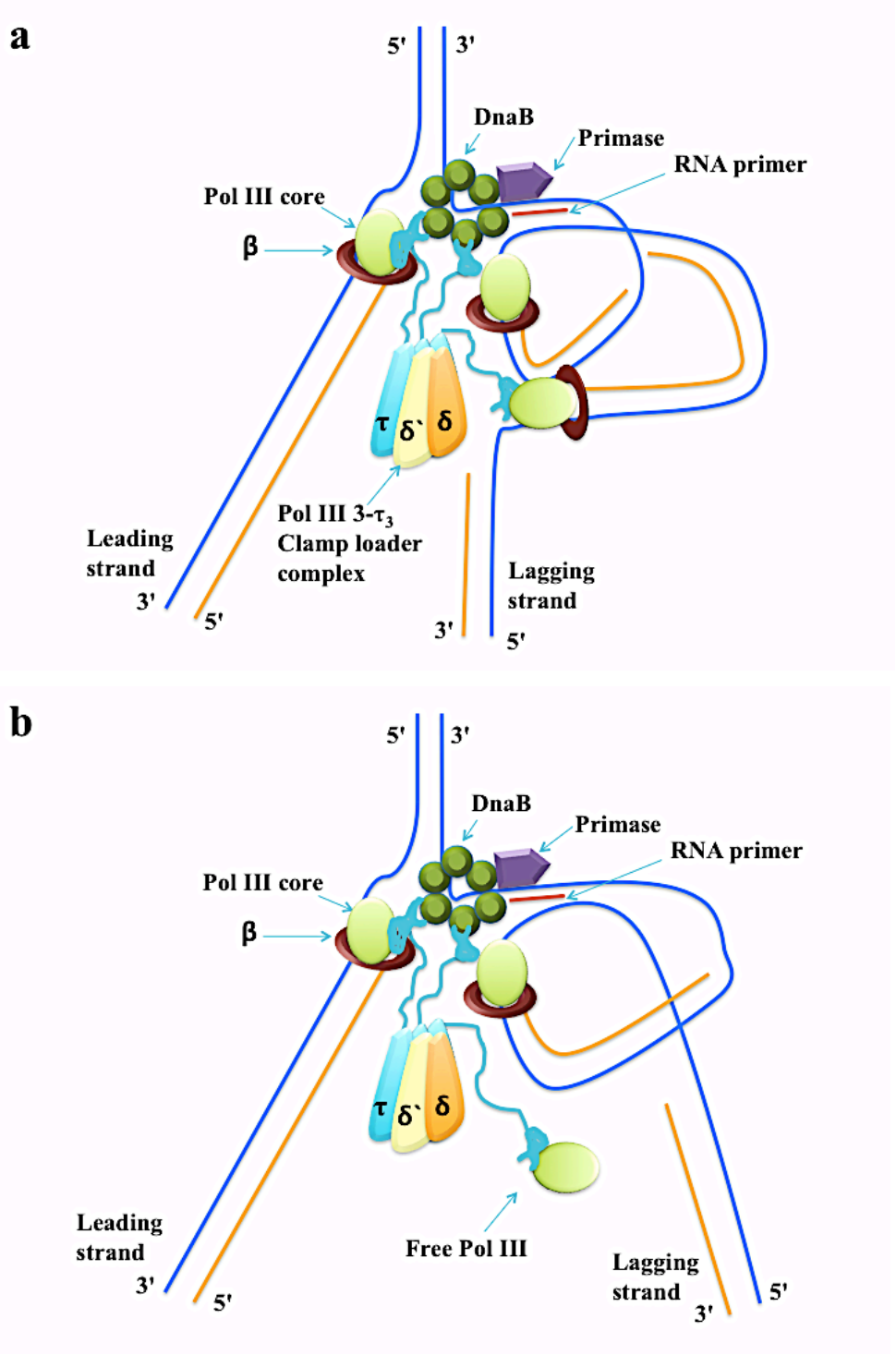



**Figure 1.2:** Illustrates trimetric replicas at the *E. coli* replication fork.
**(a)** Shows the replication fork containing three Pol III cores bound to the same clamp loader while **(b)** Shows the recycled polymerase from completely grown Okazaki fragment to extend another one.
The figure was modified from (McInerney et al, 2007).

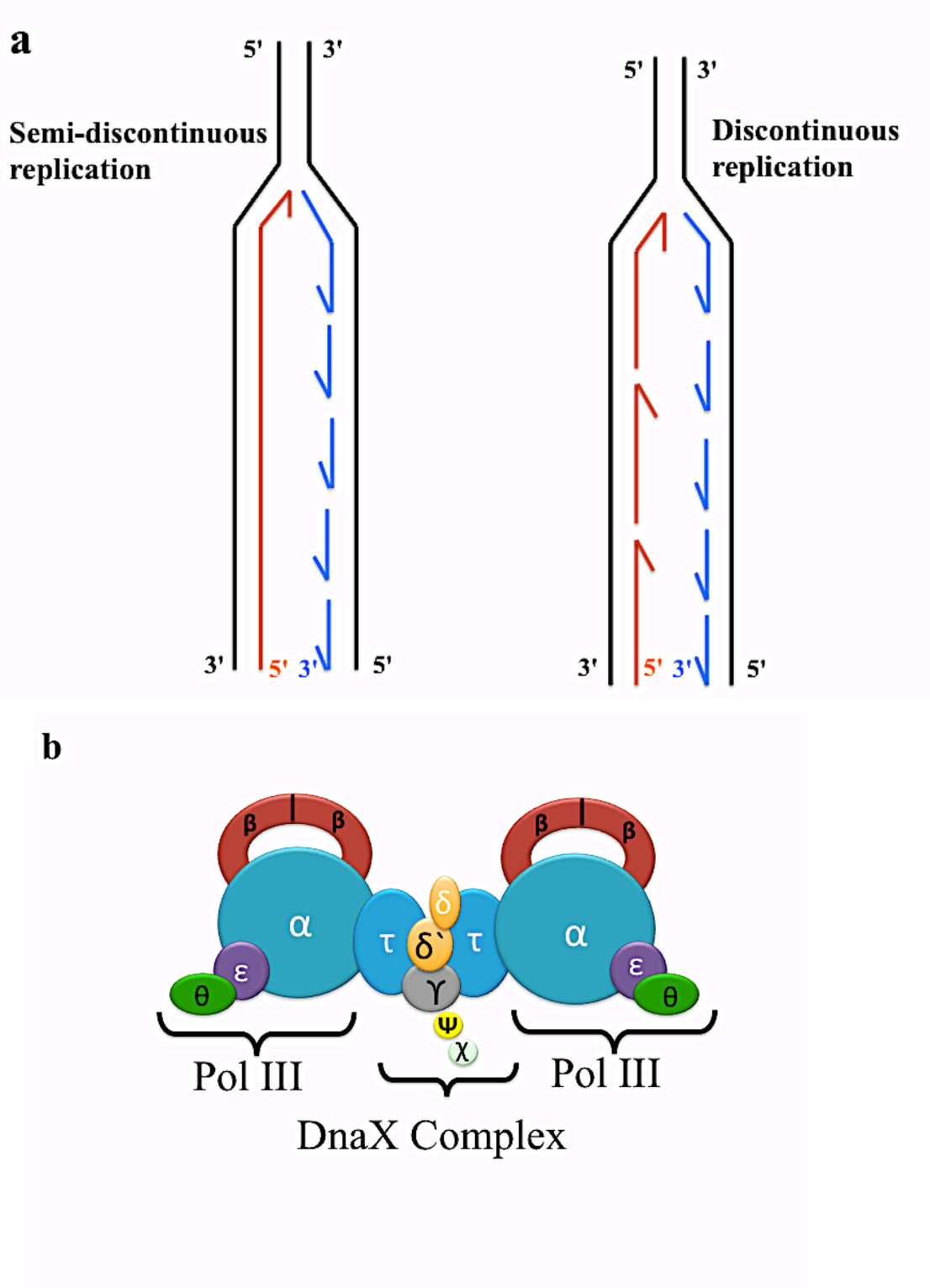



**Figure 1.3:** Replication and the replicative machine in *E. coli*.
**(a)** Illustrates the semi-discontinuous and discontinuous replication models. **(b)** Shows the tripartite replicative machine in *E. coli* components: a polymerase (Pol III), a $\boldsymbol{\beta}_2$ processivity factor and a DnaX complex.
The figures were adapted from (Yeeles, 2014) and (Fijalkowska et al, 2012) respectively.

The asymmetric nature of Pol III is essential as it performs replication on both the leading and the lagging strands simultaneously (Amado & Kuzminov, 2013; Chikova & Schaaper, 2005; Kirby et al, 2006; Taft-Benz & Schaaper, 2004). Each core is joined to one DNA template by a ring-shaped sliding clamp $\beta_2$ homodimer factor (Kong et al, 1992). The loading and unloading of the $\beta 2$ ring is performed by the action of the $\delta$ and $\delta'$ subunits (Jeruzalmi et al, 2001) which is particularly important for cycling the polymerase on and off during Okazaki fragment synthesis (McHenry, 2011). Additionally, the $\tau_2$ dimer joins the two Pol III $\alpha$ subunits generating the dimeric polymerase unit which can synthesize both leading and lagging strands (Jeruzalmi et al, 2001). Due to the anti-parallel nature of the DNA and the replication direction from 5' → 3' the asymmetric dimer allows Pol III to twist the lagging strand in a way that the direction of replication process in the lagging strand is the same as the leading strand (Zubay, 1993) (Figure 1.2).

Recently, new investigations discovered one more $\tau$ subunit in the DnaX complex instead of the $\gamma$ subunit so the complex is composed of ($\tau_3\delta\delta'\chi\psi$). As a result a new unexpected form of DNA Pol III has been found which consists of three Pol III cores instead of two to replicate the two DNA strands (Figure 1.2) (Georgescu et al, 2012; Indiani & O'Donnell, 2013; McInerney et al, 2007; Reyes-Lamothe et al, 2010). The two Pol III cores from the tri-Pol can extend two Okazaki fragments at the same time producing two lagging strand loops. After an Okazaki fragment is fully extended the polymerase leaves the old one to start synthesis of new Okazaki fragment (McInerney et al, 2007). The presence of tri-Pol III in the replisome has many advantages over di-Pol III which include increasing processivity and efficiency in lagging strand synthesis (Georgescu et al, 2012).

#### *1.1.2.2 DNA Polymerase I (Pol I)*

DNA polymerase I was the first polymerase isolated and identified and it is encoded by the *polA* gene (Delucia & Cairns, 1969; Kornberg et al, 1956). It

was also the first polymerase whose structure was solved (Ollis et al, 1985). Pol I is a protein with a molecular mass of ~ 103 kDa that plays a vital role in chromosomal replication through its three-enzymatic activities: a 5' → 3' polymerase, a 5' → 3' exonuclease and a 3' → 5' proofreading exonuclease (Joyce & Grindley, 1984). Due to its enzymatic activity, DNA Pol I is able to remove RNA primers and fill the resulting gap during the maturation of Okazaki fragments in the lagging strand (Okazaki et al, 1971). Some studies divided this protein into two parts: a Klenow fragment (consisting of a polymerase domain and a 3' → 5' exonuclease domain) and an N-terminal 5' → 3' exonuclease domain (Setlow et al, 1972; Setlow & Kornberg, 1972). Compared to the intracellular concentration of Pol III which is about 10-20 molecules per cell, Pol I can be considered as the most abundant DNA polymerase in *E. coli* with about 400 molecules per cell (Kornberg & Baker, 1992).

### 1.1.3 Okazaki Fragment Processing

RNA primers must be removed from Okazaki fragments so they can be ligated together. There are two models in prokaryotes for processing these fragments and digesting RNA primers. In the first model (Figure 1.4 right) RNase H endonuclease starts degradation of all of the RNA primer and leaves the last ribonucleotide joined to the DNA template (Ishimi et al, 1988; Kornberg & Baker, 1992). The single ribonucleotide is removed later by the 5' → 3' exonuclease activity of DNA Pol I followed by gap filling by its polymerase activity and is then ligated by DNA ligase. This system is known as the RNase H dependent pathway (Ogawa & Okazaki, 1984). The RNase H independent pathway is the second model (Figure 1.4 left) reported by Ogawa (1984) for removing oligoribonucleotide primers from Okazaki fragments. Here DNA Pol I continues synthesizing the new strand, displacing the RNA primer into a 5' flap structure. Later, the 5' flap is cleaved endonucleolytically at the flap duplex junction by the 5' → 3' exonuclease domain in DNA Pol I, generating a ligatable nick that DNA ligase then seals.

Mutations in *Streptococcus pneumonia polA* gene that abolished the polymerase activity did not affect the cell viability while deletion of the 5' → 3' exonuclease domain was lethal for these cells (Diaz et al, 1992). Another evidence clarified the importance of the 5' → 3' exonuclease activity for bacterial cells growth was determined by Fukushima et al. (2007). Complete deletion of *polA* gene in *E. coli* and *Bacillus subtilis* strains was not essential for viability but the 5' → 3' exonuclease reaction appeared important for these strains as well for cyanobacterium *Synechococcus elongates*. This study was also demonstrated that deficiency of *rnh* gene which codes for RNase H protein in *B. subtilis* was not essential for cell life and its absence can be compensated by the Pol I 5' → 3' activity. Among 250 bacterial genome searches for the 5' → 3' exonuclease domain the authors found all eubacteria have this domain (Fukushima et al, 2007).

According to Ishimi and his colleagues (1988) and (Lieber, 1997) eukaryotes and phage do not have the 5' → 3' exonuclease domain included in their DNA Pol I. Instead the 5' flap endonuclease (FEN) proteins are presented as a separate gene product but still play a critical role required for removal of RNA primers from DNA lagging strand. Interestingly, some *in vitro* genetic studies showed the null mutations of RNase H (Sommers et al, 1995) and or/ FEN1 (Reagan et al, 1995) did not affect the yeast cell viability. These investigations shed a light around protein complexes role in eukaryote replication system and suggested the involvement of another enzyme activity in RNA primers digestion. In *Saccharomyces cerevisiae*, Dna2 was shown to be involved in RNA primer removal from Okazaki fragments (Bae et al, 2001) and its deletion is lethal (Lee et al, 2000). Bae et al. (2000) found biochemical collaboration between Dna2 and FEN1 to facilitate Okazaki fragments in *Saccharomyces cerevisiae* (Bae & Seo, 2000) which also include a critical role for replication protein A (RPA) in this process.

The DNA Pol $\delta$ continues the nascent DNA polymerization and displaces the RNA primer to a 5' flap which is cleaved by FEN1 enzyme through Dna2-dependent pathway or Dna2-independent pathway. In the Dna2-

dependent pathway (Figure 1.5 right) the RPA protein binds to the ss-5' flap which recruits Dna2 and blocks FEN1 from access to the template strand. This complex allows Dna2 to cleave the ss-5' flap producing a short flap which then is processed by FEN1 or Exo1 or by the action of the 3'→5' exonuclease activity of Pol $\delta$. In Dna2-independent pathway (Figure 1.5 left) the ss-5' flap is too short which prevent the binding of RPA and generating the RPA-Dna2 complex. The short flap then cleaved by FEN1 enzyme only. In order to protect the strand template FEN1 activity inhibited by RPA protein (Bae et al, 2001) while the action of Dna2 can be repressor by the single-strand binding proteins (Kao et al, 2004).

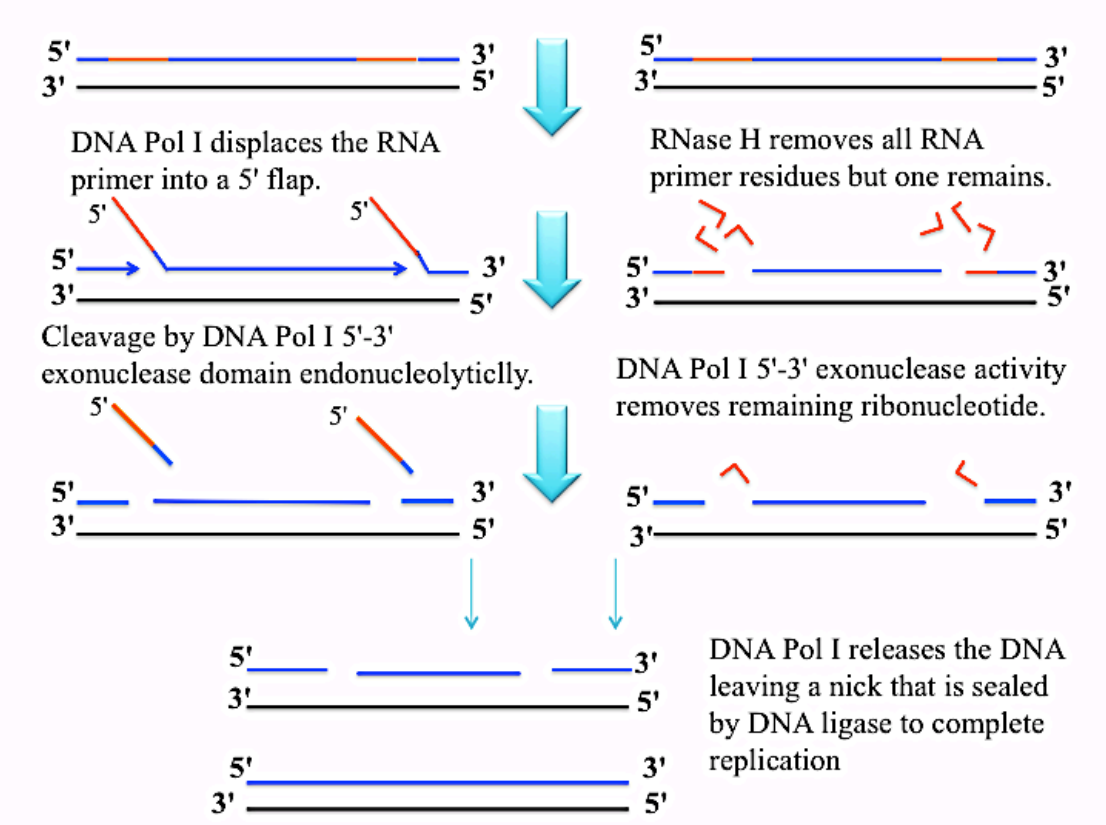



**Figure 1.4:** Okazaki fragment processing in prokaryotes.
The figure shows two models suggested for this process: the Rnase H-dependent pathway is on right while the Rnase H-independent pathway is on left.
The Figure was modified from (Hemsworth, 2009)

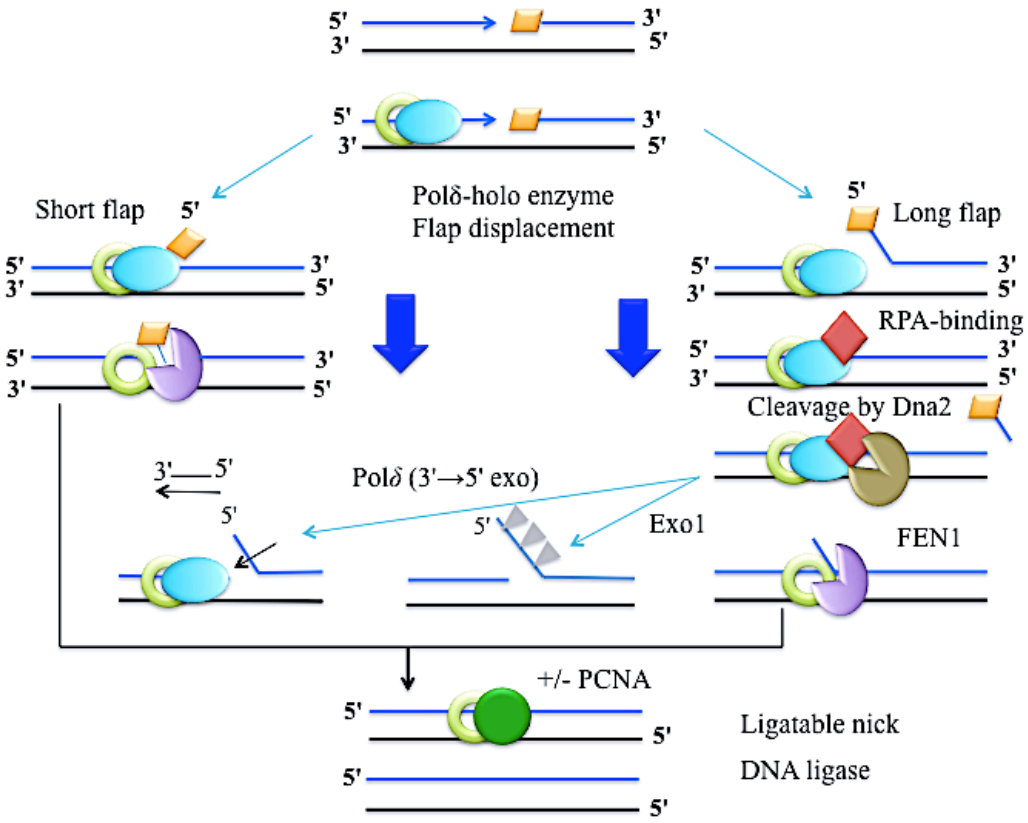



**Figure 1.5:** Okazaki fragment processing in eukaryotes.
Two models are proposed: Dna2 dependent pathway on right while the Dna2 independent pathway on left. RNA primer (orange), PCNA (light green), Pol***δ*** (cyan), RPA (red), Dna2 (sand), FEN1 (purple) and DNA ligase (green).
The figure was modified from (Henneke et al, 2003).

## 1.2 5' Flap Endonucleases

As mentioned previously bacteriophage and eukaryotes lack the 5' → 3' exonuclease domain that exists in bacteria and prokaryote Pol I. In these organisms the removal of RNA primers carries out by the action of flap endonuclease enzymes (FEN1). FENs are a group of structure-specific nucleases that require at least two metal ion as co-factors and possess both endo- and exo- nucleolytic activities (Garforth et al, 2001). In 1994 Harrington and Lieber cloned a mouse nuclease that could cut a "DNA flap" structure. Their flap was composed of both a duplex DNA and a 5'−3′ displaced single-stranded DNA arm. The enzyme cleaved the DNA both endonucleolytically and acted as a 5'−3′ exonuclease and hence named the protein a five prime endo/exonuclease (FEN-1) (Harrington & Lieber, 1994b). FEN family members have since been identified in many organisms from different kingdoms of life on the basis of similarity in function, protein structure and sequence (Harrington & Lieber, 1994a; Shen et al, 1998). In *Saccharomyces cerevisiae,* where the FEN1 enzyme is known as RAD27, null mutations result in temperature sensitive viability, increased susceptibility to mutagens and UV light, as well as genome instability (Moreau et al, 2001; Reagan et al, 1995). In contrast, introducing of the 5' → 3' exonuclease domain of *E. coli* DNA Pol I into a RAD27 deletion mutant resulted in suppression of the phenotypes caused by this mutation (Sun et al, 2002). RAD2 is a homolog of RAD27 which was found in *Saccharomyces pombe* and named by Sommers and colleagues (Sommers et al, 1995).

The importance of FEN1 is further demonstrated by studies that suggest it is as a potential cancer susceptibility gene (Henneke et al, 2003). Studies on higher eukaryotes confirmed that homozygous deletion of the FEN1 gene in mice results in embryonic lethality whereas lymphomas rabiddly developed in heterozygotes and thymic abnormality was diagnosed (Kucherlapati et al, 2002). In human cancer patients, a group of *FEN1* mutations and single-nucleotide polymorphisms (SNP) that depress *FEN1* activities have been identified (Wu et al, 2012). Most of these somatic mutations abolished FEN1 exonuclease (EXO) and gap-endouclease (GEN) activities and related

to FEN1-E160D mutation. To study the effect of this mutation, a mimic mouse line that carries an FEN1-E160D point mutation system was generated. The mutant mice were predisposed to autoimmunity, chronic inflammation and cancer. Additionally, incomplete digested DNA fragments were observed in apoptotic cells (Zheng et al, 2007). A recent study reported a link between the development of lung cancer in humans and mutations in *FEN1* caused by Benzo[$\alpha$]pyrene (B[$\alpha$]P) the major DNA damaging compound found in tobacco smoke (Wu et al, 2012). FEN1 is involved in multiple pathways that are seemingly antithetical, for instance, genome maintenance versus apoptotic DNA fragmentation. These functions of FEN1 may be regulated *via* three mechanisms; formation of complexes with different proteins, cellular compartmentalization and post-translational modifications. To date about 34 proteins involved in different DNA metabolic pathways have been reported as interacting with FEN1 to stimulate its nuclease activity or to guide the enzyme to different DNA processes such as replication, repair or degradation (Zheng et al, 2011). In damaged DNA, FEN1 was shown to localize in the nucleus (Qiu et al, 2001) while a more recent study found FEN1 super-accumulates in the nucleolus where it may stabilize tandem repeats of ribosomal DNA (Guo et al, 2008). In addition, due to the presence of FEN1 in the mitochondrion it is thought to play a critical role in mitochondrial DNA (mDNA) replication and repair as well (Liu et al, 2008). The ability of FEN1 enzyme to be acetylated, phosphorylated and methylated shed a light around the post-translation modifications that may be important for the regulation of its activities (Guo et al, 2012; Hasan et al, 2001).

Several structures of FEN family members have been determined, some of them in complexes with DNA substrates or products including the 5' exonuclease domain of *Thermus aquaticus* (*Taq*) DNA polymerase (Kim et al, 1995), bacteriophage T5 D15 exonuclease (Ceska et al, 1996), bacteriophage T4 RNase H (Mueser et al, 1996), *Methanococcos jannaschii* FEN-1 (Hwang et al, 1998), *Pyrococcus furiosus* FEN-1 (Hosfield et al, 1998), *Pyrococcus horikoshii* FEN (Matsui et al, 2002), *Archaeoglobus fulgidus* FEN-1 (Chapados et al, 2004), human FEN-1 (Sakurai et al, 2005),

*Escherichia coli* ExoIX (Anstey-Gilbert, 2006) and hEXO-1 (Orans et al, 2011) (Figure 1.6). These FEN members share a common structure consisting of a central *beta* sheet surrounded by *alpha* helical domains (Figure 1.7a). The highly conserved active site is located in the protein centre and contains two metal ion-binding sites referred to as Cat1 and Cat2 (Anstey-Gilbert et al, 2013) that bind divalent metal ions either directly by acidic residue or through water molecules. Above the active site a helical arch or a flexible loop is located which is presumed to play an important role in substrate recognition and binding (Ceska et al, 1996; Sayers & Artymiuk, 1998). FEN enzymes are able to form complexes with several DNA substrates including Pseudo-Y, double flap, 5'-overhang and 3'-overhang (Figure 1.7b).

## 1.2.1 FENs Structures and Conserved Motifs

### *1.2.1.1 The Motifs Interact with DNA Duplex Regions*

#### 1.2.1.1.1 The H3TH:$K^+$ Motif

In many DNA binding proteins such as DNA polymerases, S13 ribosomal proteins, and DNA glycosylases a common non-specific DNA binding motif is found and known as helix-hairpin-helix (HhH) motif (Doherty et al, 1996; Thayer et al, 1995). This motif binds to DNA through hydrogen bonds formed between protein backbone nitrogens and DNA phosphate groups. Structurally, the motif is composed of two antiparallel $\alpha$ helices joined by a hairpin-like loop that interacts with DNA and contains a consensus sequence motif of glycine-hydrophobic residue-glycine (GhG) (Doherty et al, 1996). In all FEN structures a similar motif found in the C-terminal domain of these enzymes, is involved in DNA binding and is known as the helix-three-turn-helix (H3TH). It shares a similar structure with HhH motif except for the addition of amino acids in the loop region before the GhG sequence (Hosfield et al, 1998).

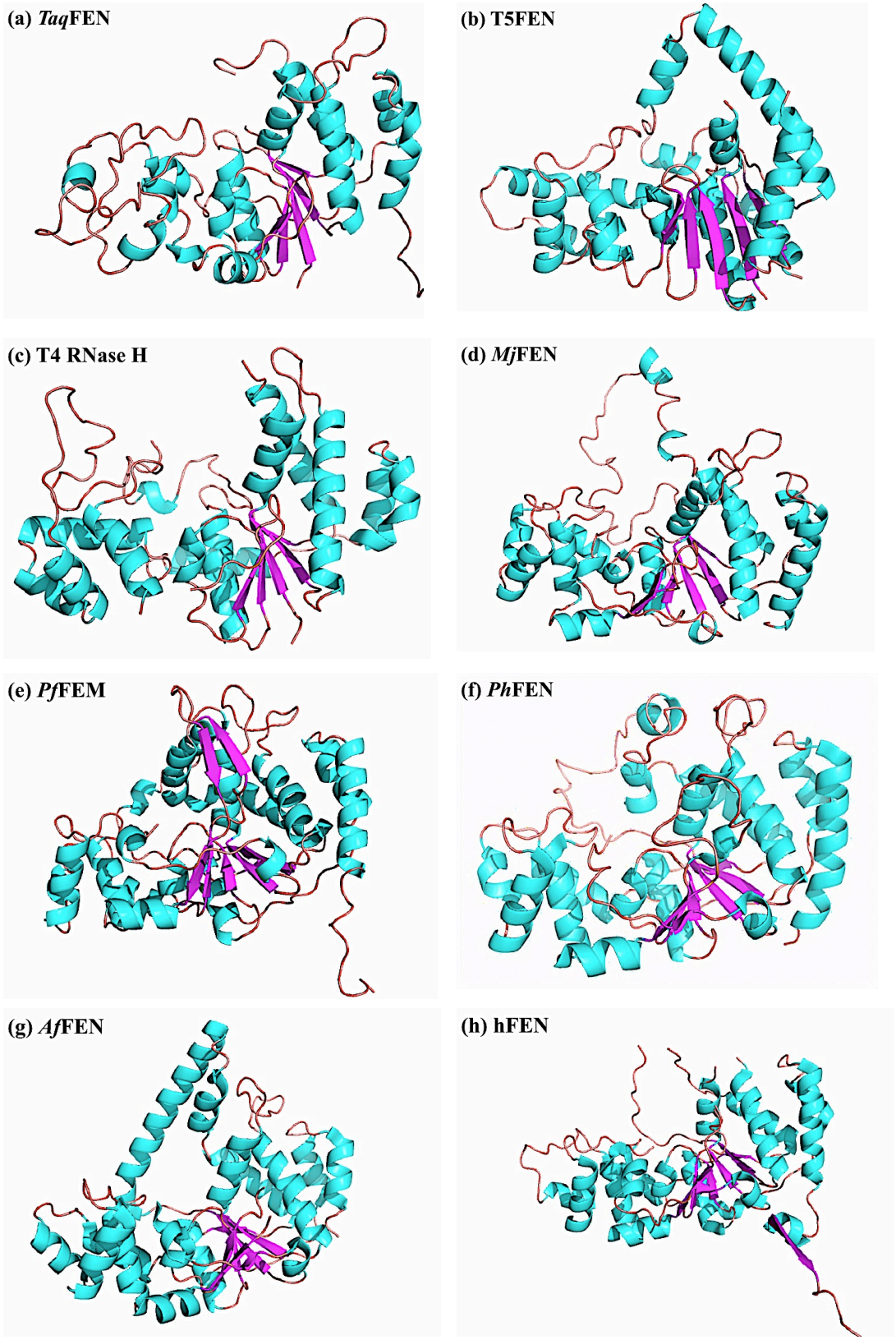



**Figure 1.6:** Cartoon diagrams for the three-dimensional crystal structures of some FEN family members.

$\alpha$-helices colored cyan, $\beta$-sheets magenta and the loops in orange. Pdb IDs **(a)** 1TAQ, **(b)** 1EXN, **(c)** 1TFR, **(d)** 1A76, **(e)** 1B43, **(f)** 1MC8, **(g)** 1RXV, **(h)** 1UL1.

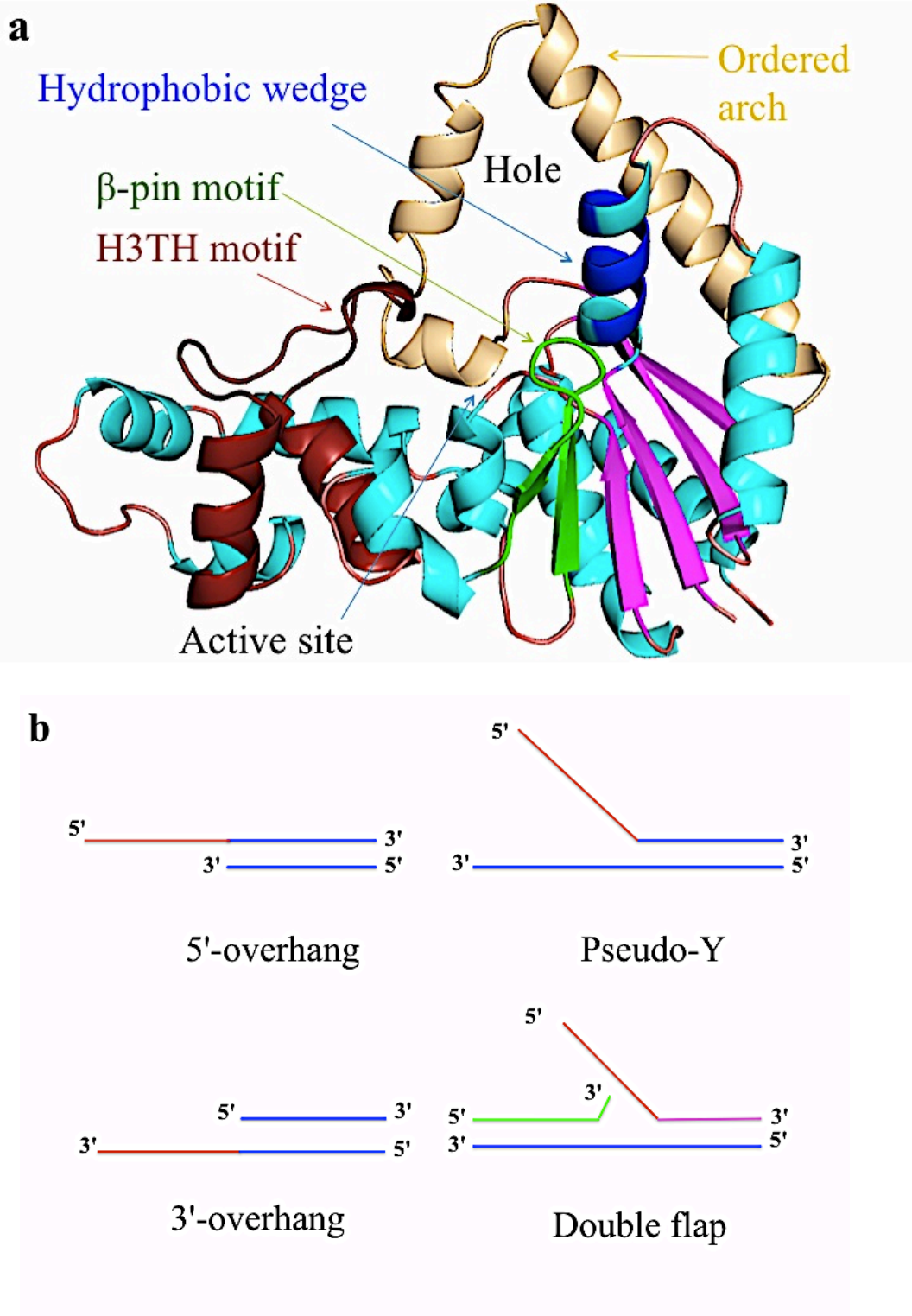



**Figure 1.7:** FEN family member common fold and their DNA substrates.
**(a)** Crystal structure for a FEN (T5FEN) shows the common fold for these proteins. ***β***-strands (magenta) are surrounded with ***α***-helices (cyan) and the conserved motifs are colored. **(b)** DNA substrates that can form complexes with FEN family members.

The extended loop contains some acidic residues such as aspartate residues (D201 and D204) in T5FEN which extend and engage in the enzyme active site formation (Artymiuk et al, 1997). Furthermore, the H3TH motif is thought to interact with the DNA backbone through the basic lysine and arginine residues that are found in the motif structure (Hosfield et al, 1998) and may play a role in the substrate movement into the appropriate position for enzyme activity. When these basic residues were substituted by alanine in the T5FEN-H3TH motif the enzyme:DNA became destabilized. The Kd's increased from 4 nM for the WT to 48 nM and 500 nM respectively for the K215A and R216A (Dervan et al, 2002). According to recent biochemical studies, FEN proteins recognize double stranded DNA (dsDNA) first and then place the scissile phosphate in the active site by threading or clamping the single strand DNA (ssDNA) (Finger et al, 2009; Stewart et al, 2009). A theoretical model for the *Pfu*FEN-1 enzyme in a complex with DNA has been proposed on the basis of the knowledge of the enzyme 3D crystal structure and biochemical and molecular dynamics (MD) data (Allawi et al, 2003). In this model hydrogen bonds are formed between the backbone nitrogen atoms of the Gln-244, Gln-246, Lys-248 and Lys-249 residues and the phosphate backbone of the duplex DNA. The (GhG) consensus sequence was present within H3TH motif.

In an experimental structure of T4 RNase H complexed with a pseudo Y DNA the interactions that are formed between the H3TH motif and the duplex DNA were observed (Figure 1.8a) (Devos et al, 2007). The backbone amide groups of Thr-221, Met-224 and Thr-226 in the H3TH loop form hydrogen bonds with the phosphate groups of the dsDNA but the GhG consensus sequence is absent in the T4 RNase H protein. Other interactions were observed: salt bridges were formed between the side chains of Lys-195, Lys-198 and Arg-215 and the phosphate backbone (Figure 1.8a). In the minor groove of the DNA, Arg-220 residue interacts with oxygen atoms in the sugar rings of guanosine and thymidine nucleotides as well as with the carbonyl group of the thymidine base. Recently, in 2011 Tsutakawa and co-workers have obtained a structure for a complex of hFEN-1 with double-flap DNA and at the same time (2011) Orans with his team determined the

crystal structure for hEXO-1 in a complex with 3' flap DNA substrate. In hFEN-1 (Tsutakawa et al, 2011) and hExo-1 (Orans et al, 2011) H3TH motif presents with shorter loop which is called the helix-two-turn-helix motif (H2TH) instead. In both enzymes this motif provides significant interactions with DNA substrates. The interactions between hFEN-1 and the substrate can be separated into two regions. The first binding region through hydrogen bonds formed by Gly-240, Gly-242, Lys-244 and Arg-245 main chain amide nitrogens to DNA phosphate groups (Figure 1.8b). The second binding region between the protein and the DNA template strand occurs indirectly *via* a $K^+$ ion which is coordinated by Ile-238 and Ile-241 carbonyl oxygens and Ser-237 hydroxyl group in the H2TH loop region (Figure 1.9a). The GhG sequence is presented clearly in hFEN-1 enzyme and interacts with the DNA backbone by Gly-240, Ile-241 and Gly-242 residues.

A few contacts between hEXO-1 and DNA substrate have been determined (Orans et al, 2011). These interactions are similar to those formed between hFEN-1 and DNA in the H2TH motif region. The backbone amide groups of Arg-231, Gln-232, Gln-234 and Lys-237 in the H2TH motif form hydrogen bonds with the phosphate oxygen atoms of the duplex DNA (Figure 1.8c). Further interactions are formed from the phosphate oxygen backbone of the DNA indirectly through the $K^+$-binding site which has a $K^+$ ion coordinated by Ser-222 and Ile-233 backbone carbonyl groups, Ser-229 hydroxyl in H2TH loop and two water molecules (Figure 1.9b). The GhG motif is also present in this protein (Gln-232, Ile-233 and Gln-234) and interacts with the substrate backbone.

Anstey-Gilbert and his colleagues have determined another structure for other member of the FEN family, *E. coli* ExoIX in a complex with a DNA product (Anstey-Gilbert et al, 2013). In this structure the enzyme contacts the DNA phosphate backbone through 6 hydrogen bonds in the H3TH motif (Figure 1.8d).

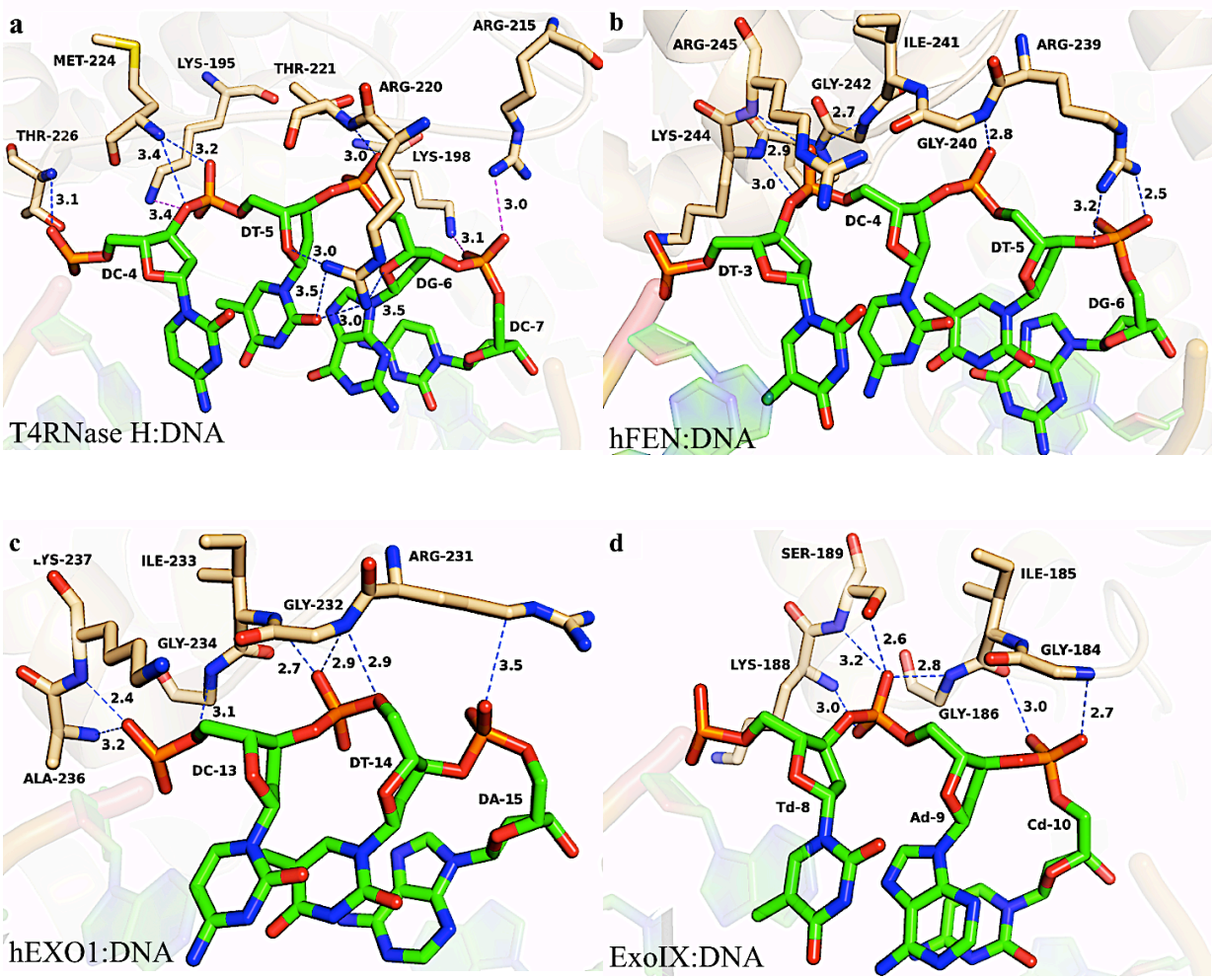



**Figure 1.8:** The H3/2TH motif in some FEN members in complexes with DNA substrates.

**(a), (b), (c)** and **(d)** show the interaction between the DNA and H3/2TH motifs in T4RNase H (2IHN), hFEN:DNA (3Q8K), hEXO-1:DNA (3QEA) and *E. coli* ExoIX:DNA (3ZDC) respectively. These interactions are formed by hydrogen bonds (blue dashes) and salt bridges (violet dashes).

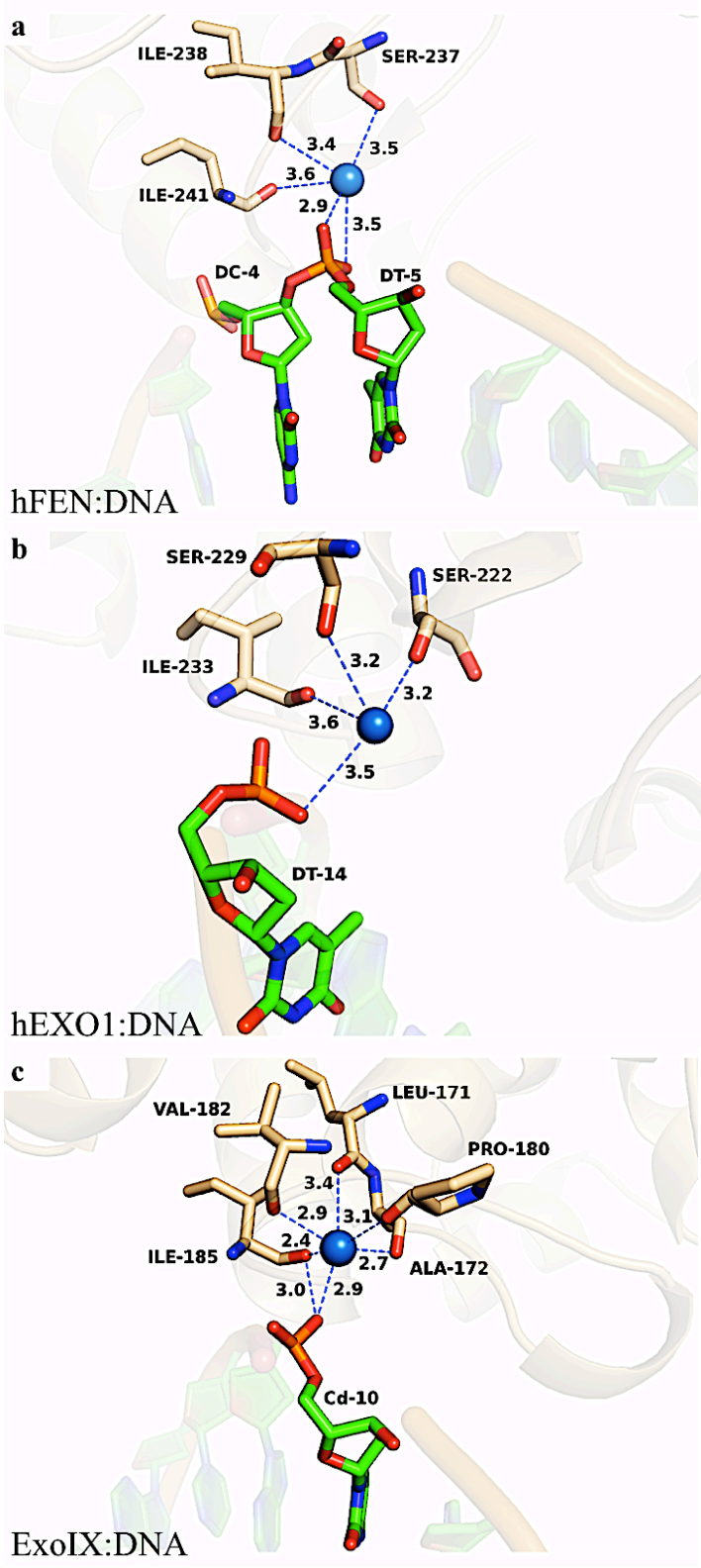



**Figure 1.9:** Diagram for the potassium ion in its binding site in FEN enzyme complexes with DNA substrates.
**(a), (b)** and **(c)** Show $K^+$ ion (blue spheres) in hFEN:DNA (3Q8K), hEXO-1:DNA (3QEA) and *E. coli* ExoIX:DNA (3ZDC) respectively. The potassium ion forms indirect interactions between the DNA backbone and FEN proteins.

The main chain amide groups of Gly-184, Gly-186 and Lys-188, the main chain carbonyl of Ile-185 and the side chain of Ser-189 residues in H3TH motif make these interactions. The first two residues are the glycines present in the GhG consensus sequences within the HhH motif but absent in H3TH motif in T4 RNase H (Figure 1.8d). Further interactions were made through $K^+$ ion which is coordinated by five residues backbone carbonyl groups (Figure 1.9c). The $K^+$ binding site present in the presence of DNA substrates in hFEN-1 (Tsutakawa et al, 2011), hEXO-1 (Orans et al, 2011) and in *E.coli* ExoIX (Anstey-Gilbert et al, 2013) is similar to the one observed in human DNA polymerase $\beta$ (hPol $\beta$) (Pelletier & Sawaya, 1996). The role of this site in hPol $\beta$ is thought to facilitate movement of the protein along the DNA substrate backbone (Pelletier & Sawaya, 1996) which could be the same for the other FENs.

#### 1.2.1.1.2 The Hydrophobic Wedge and $\beta$-pin

According to the latest investigations (Chapados et al, 2004; Orans et al, 2011; Tsutakawa et al, 2011) and reviews (Grasby et al, 2012) another motif is identified in FEN family members in complexes with DNA substrates. This motif interacts with the duplex region of the DNA substrate from the other end (in respect to the DNA interaction region with H3/2TH motif) and is known as "hydrophobic wedge". The interactions with the helix wedge in FEN superfamily force the DNA substrate to bend and interact with another small motif called the $\beta$-pin which is composed of two antiparallel strands connected with a flexible loop (Grasby et al, 2012; Tsutakawa et al, 2011). In 2004, the first structure of a FEN:DNA complex was determined by (Chapados et al, 2004) for *Archaeoglobus fulgidus* (*Af* FEN) bound to a short branch of duplex DNA with a single nucleotide 3' overhang. A hydrophobic wedge stacked against the DNA duplex which is thought to play a role in stabilizing of the upstream duplex and in opening the DNA at the flap junction region (Figure 1.10a). Additionally, the duplex DNA in T4RNase H complex interacts with the side chains of hydrophobic residues Ile-29 and Val-30 and with Met-173 and His-174 extended from a loop occupies a same position of $\beta$-pin motif loop in FEN superfamily members

(Figure 1.10b). The interactions between the DNA and this motif were not observed in *E. coli* ExoIX because the duplex region of the DNA substrate did not reach this part of the protein. The hydrophobic wedge region in the hFEN:DNA complex is formed by the $\alpha 2$ helix and the $\alpha 2-\alpha 3$ loop which was disordered in the DNA-free enzyme (Sakurai et al, 2005) so it forms through a disorder-to-order transition in the presence of the DNA substrate. It was stacked against the duplex DNA and made some interactions by Arg-47, Gln-48 and Met-65 with the substrate bases (Figure 1.10c). The $\beta$ pin region is formed by $\beta 6$ and $\beta 7$ and the loop between them. This motif lies closer to the active site in the DNA-free enzyme while in the complex structure it moves and interacts with the substrate DNA. The movement in this motif residues are ranging from 7.3 Å for Ala-196 to 2.0 Å for Lys-201 in C$\alpha$s.

#### *1.2.1.2 The 3' flap Binding Site*

Biochemical experiments on FEN enzymes in complexes with DNA have shown that a 5' flap substrate with a single-nucleotide 3' flap is the better DNA substrate for these family members. The presence of the 3' flap in the substrate has been shown to enhance the specificity of 5' flap cleavage and recognition by FENs (Kaiser et al, 1999). Resent studies in FEN:DNA complexes determined the presence of a specific pocket formed by the wedge and a C-terminal loop which is comprised of a helix-hairpin-helix that is occupied by a 3' single nucleotide flap. The C-terminal which is a part from the 3' flap binding site or specific pocket is identified in eukaryotic FEN proteins such as *Af* FEN:DNA (Chapados et al, 2004) and hFEN:DNA (Tsutakawa et al, 2011) but absent from hEXO1 (Orans et al, 2011), *E.coli* ExoIX (Anstey-Gilbert et al, 2013), T4RNase H (Devos et al, 2007) and T5FEN (Ceska et al, 1996) which suggests that the 3' single nucleotide pocket is also absent in these proteins (Figure 1.11) . Mutations in the amino acids that formed the 3' flap pocket site in *Af* FEN structure decreases the enzyme activity by 100 fold compared to the wild type and disrupted the interactions between the protein and the duplex DNA phosphate backbone.

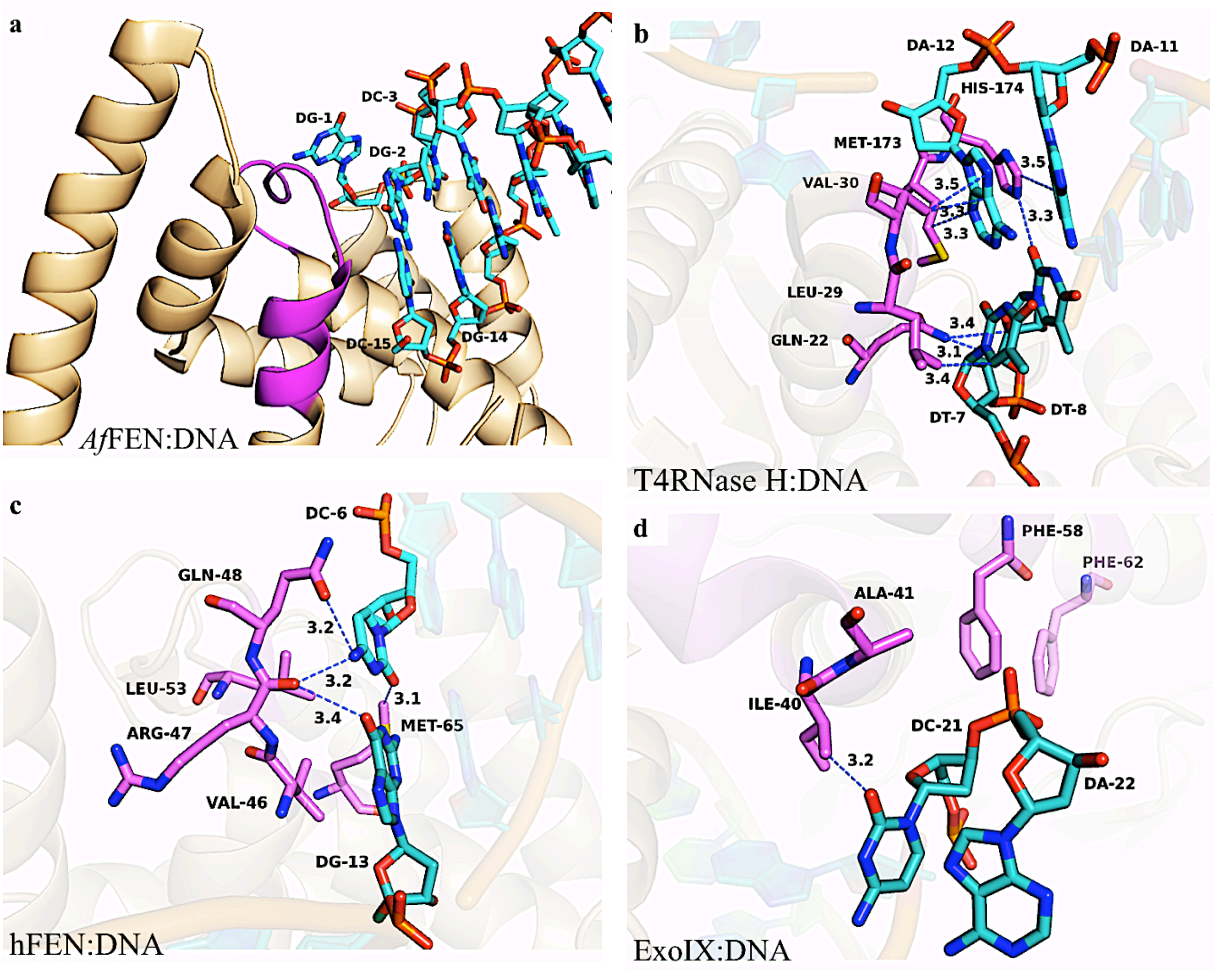



**Figure 1.10:** The hydrophobic wedge and its interactions with the DNA substrates in some FEN proteins.
**(a)** Shows the hydrophobic helix (magenta cartoon) in *Af* FEN (1RXW). **(b), (c)** and **(d)** Illustrates the hydrophobic (magenta sticks) interactions in T4RNase H (2IHN), hFEN:DNA (3Q8K) and hEXO-1:DNA (3QEA) respectively.

The role of the 3' single nucleotide flap in increasing the enzyme activity is thought to be by ordering the helical arch region allowing cleavage of the 5' flap to occur (Chapados et al, 2004). In *Af* FEN:DNA crystal structure, the 3' single nucleotide was inserted into this pocket and coordinated with Thr-55 side chain and His-308 main chain carbonyl (Figure 1.12a) and in hFEN:DNA complex (Tsutakawa et al, 2011) a 3' single base was also inserted into the pocket and interacted with Lys-314 main chain carbonyl group, Gln-54 main chain amide nitrogen and Thr-61 side chain (Figure 1.12b). As mentioned previously, T4RNase H protein lacks the 3' single nucleotide pocket but some interactions between the DNA and the protein were noticed. Lys-52, Asn-303 and Ser-301 side chains formed hydrogen bonds to deoxycytosine-15 and deoxyguanosine-16 and dG-17 while Asn-300 formed interactions by its main and side chains with dG-17 (Figure 1.12c). Additionally, *E.coli* ExoIX lacks the 3' flap in its DNA substrate and so it is hard to identify any interactions in this part of the protein (Anstey-Gilbert et al, 2013) whilst hEXO1 has 3' arm in its DNA complex but does not have the 3' single-nucleotide pocket. This 3' arm is too short (about two bases) and cannot extend beyond the hydrophobic wedge. The sequence alignment analyses confirm the conservation of the amino acid residues in the 3' flap-binding pocket from archaeal to human FEN's (Friedrich-Heineken et al, 2003). However, there are also some variability between them giving different specificity.

#### *1.2.1.3 The Helical Arch and 5' Flap Interactions*

Another motif found in all FEN family members as ordered or disordered structure and is called helical, clamp arch or intermediate domain. This motif is thought to play a critical role in DNA 5' flap interactions either by threading, tracking or clamping the ssDNA 5' flap. In the 5'→3' exonuclease domain of *Taq* DNA polymerase (Kim et al, 1995), the crystal structure of T4 RNase H (Mueser et al, 1996) and *Escherichia coli* ExoIX (Anstey-Gilbert, 2006) the helical arch region is partially disordered and is appeared as a flexible loop (Figure 1.13a, b & c).

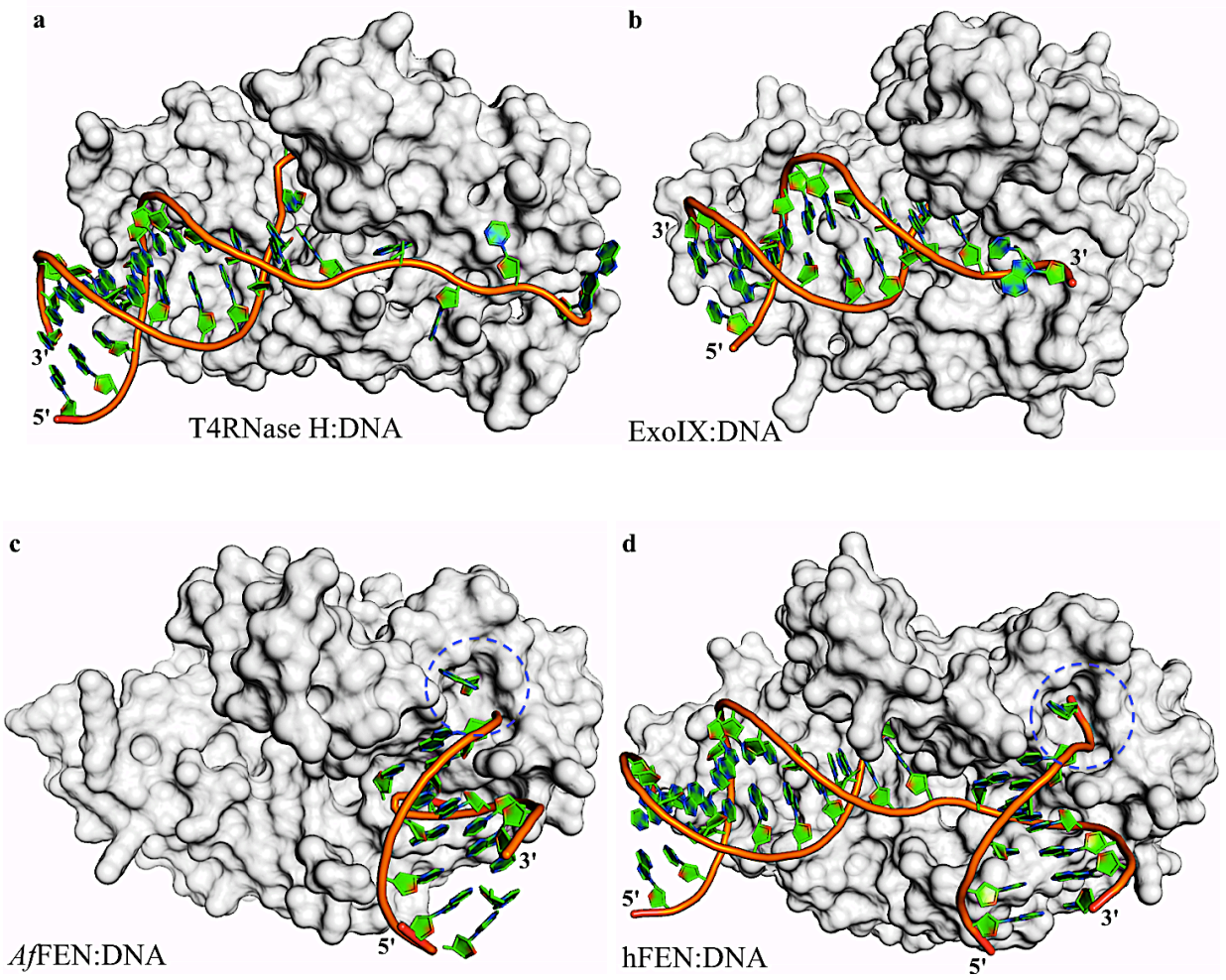



**Figure 1.11:** Cartoons and surface representations for four FEN enzymes in complexes with DNA substrates illustrate the absence or presence of the 3'-single nucleotide-binding pocket (overhead view).
**(a)** and **(b)** show surface representations for T4 RNase H:DNA (2IHN) and hEXO-1:DNA (3QEA) and the 3' flap binding pocket is not found in ether of them.
**(c)** and **(d)** show surface representations for *Af*FEN:DNA (1RXW) and hFEN-1:DNA (3Q8K) and the 3' flap binding pocket can be observed (blue dashed circle) in both of them. The 3' single nucleotide is inserted into the pocket.

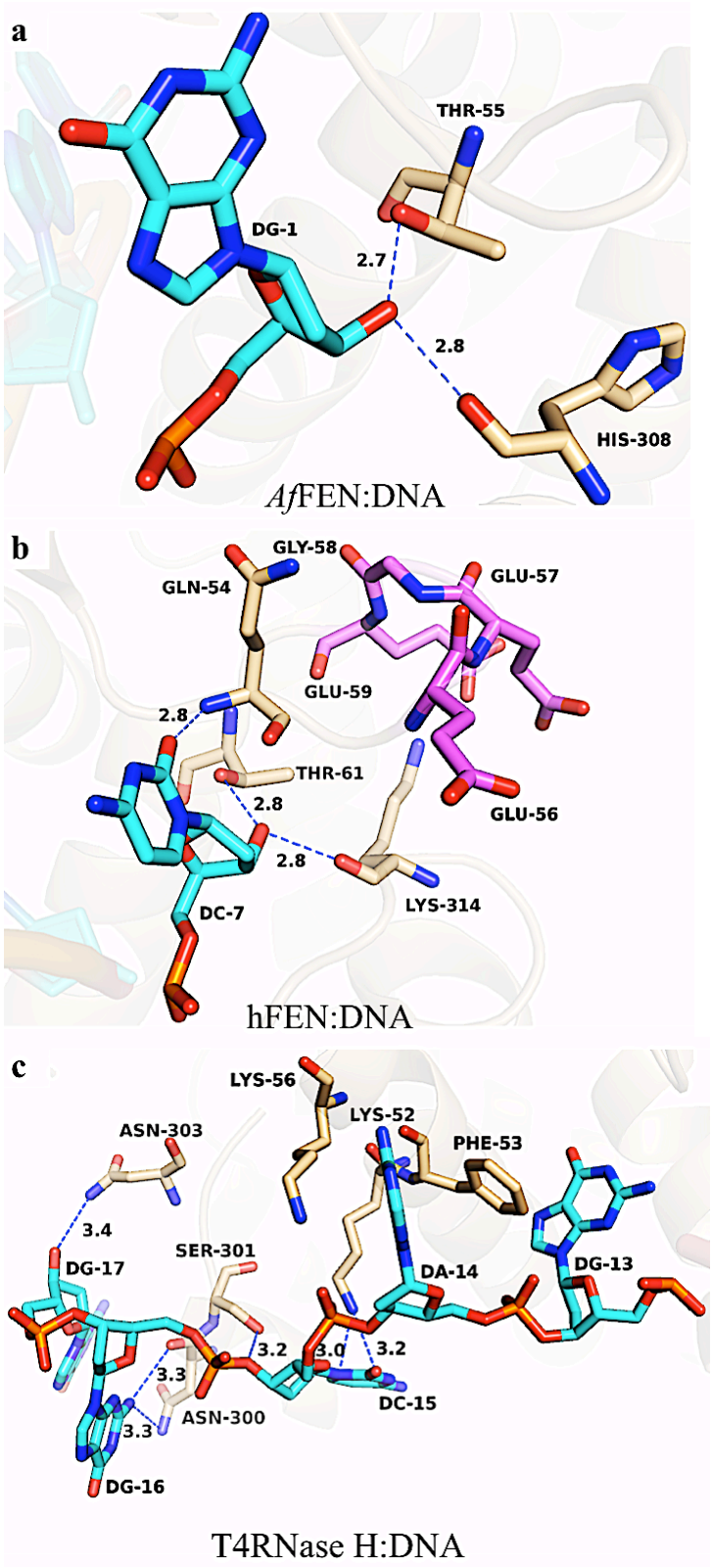



**Figure 1.12:** The interactions between some FEN members and the DNA 3' flap **(a)** and **(b)** show the 3' single nucleotide inserted in a specific pocket made by the hydrophobic wedge loop and the C-terminal in *Af*FEN:DNA (1RXW) and hFEN:DNA (3Q8K) respectively. **(c)** Shows the interactions with the 3' arm in T4RNase H complex with fork-DNA substrate (2IHN).

T5FEN has an ordered arch in DNA-free enzyme (Figure 1.13d) (Ceska et al, 1996; Feng et al, 2004) while hFEN has disordered arch (Figure 1.13e) in the protein complexed with PCNA but without DNA substrate (Sakurai et al, 2005) and a fully ordered arch in a complex with DNA double flap (Tsutakawa et al, 2011) (Figure 1.13f). A lot of suggestions and models hypotheses have been published for the helical arch and its role in the interactions between the FEN enzymes and the DNA 5' flap.

Threading, tracking, clamping or disorder-to-order mechanisms have been proposed, so do FEN enzymes use more than mechanism? The threading model in which the 5' flap is pushed through a hole in the protein large enough to accommodate ssDNA but not dsDNA and then comes to rest at the ss/dsDNA junction (Ceska et al, 1996; Lyamichev et al, 1993; Patel et al, 2012; Tsutakawa et al, 2011) (Figure 1.14a). This model is supported by the structure of T5FEN, which contains a helical arch located above the active site, and forms an 18 Å diameter hole lined with basic lysine and arginine residues that are thought to interact with the DNA phosphate backbone (Figure 1.13d) (Ceska et al, 1996; Garforth et al, 1999). Another FEN structure that supports the threading model is *Methanococcus jannaschii* FEN-1 (*Mj*FEN) (Figure 1.6d) (Hwang et al, 1998) that has a loop termed L1 in front of the active site which forms a hole sharing the similar characteristics with the arch in T5FEN. Additionally, by deleting this loop and assaying the protein activity (Hwang et al, 1998) the results showed that the enzyme was not active and only a small amount of non-specific cleavage was observed. Based on this result, Hwang and coworker concluded that the L1 loop is involved in specific recognition of the junction area between the single stranded flap and the duplex DNA.

On the other hand, there are a number of studies that are apparently at odds with the threading model. FEN enzymes can cleave circular single-stranded DNA substrates (Sayers & Eckstein, 1991) and double-stranded plasmid DNA under forcing conditions (Garforth et al, 2001). Additionally, in 1999 Bornarth and his team found that hFEN-1 could bind and cleave a modified flap substrate by addition of large chemical adducts which are too big to

thread through the protein arch (Bornarth et al, 1999). Recently, hFEN-1 was shown to possess gap dependent endonuclease (GEN) activity where it can bind gapped substrates and cleave without a free 5' terminus to interact with (Liu, 2006; Zheng, 2005).

The other model suggested for FEN proteins is the tracking model in which the enzymes do not thread the 5' flap but interacts with the DNA bases and then slide down the ssDNA to interact with the ssDNA/dsDNA junction at a flap (Figure 1.14a) (Gloor et al, 2010). Recent crystallographic studies of hFEN-1 complexed with a double flap DNA substrate (Tsutakawa et al, 2011) and hEXO-1 complexed to a 3' flap DNA substrate (Orans et al, 2011) showed that the two enzymes share a common strategy to bind and cleave the DNA, as mentioned above in sections 1.2.1.1 and 1.2.1.2. There is a disagreement between the two teams in the way that the 5' flap binding in hFEN-1 and hEXO-1 structures and each of them propose a different mechanism. Beese and co-worker did not observe a 5' flap directly in the hEXO-1:DNA structure because they used a 3' flap DNA (Orans et al, 2011). In addition, they raised some questions against threading model such as the energy source that is used to push the single flap into the archway in the absence of ATP and how can other substrates thread through the arch while they do not have free 5' termini. An alternative model involved with clamping by the arch helices would be more acceptable to them (Orans et al, 2011). Based on the structure of DNA-free hFEN-1 (Sakurai et al, 2005) and hFEN-1:DNA structure analyses, Tainer and colleagues suggested a disorder-to-order transition mechanism of the arch helices during a threading process when it was observed that the arch region was disordered in the DNA-free structure (Sakurai et al, 2005) while it appeared completely ordered upon DNA binding (Tsutakawa et al, 2011) (Figure 1.13e & f). According to this mechanism, the helical arch can accommodate different substrates including circular and gapped DNA. However, the arch motif is still disordered in RNase H (Devos et al, 2007) and ExoIX (Anstey-Gilbert et al, 2013) proteins complexed with DNA substrate (Figure 1.15).

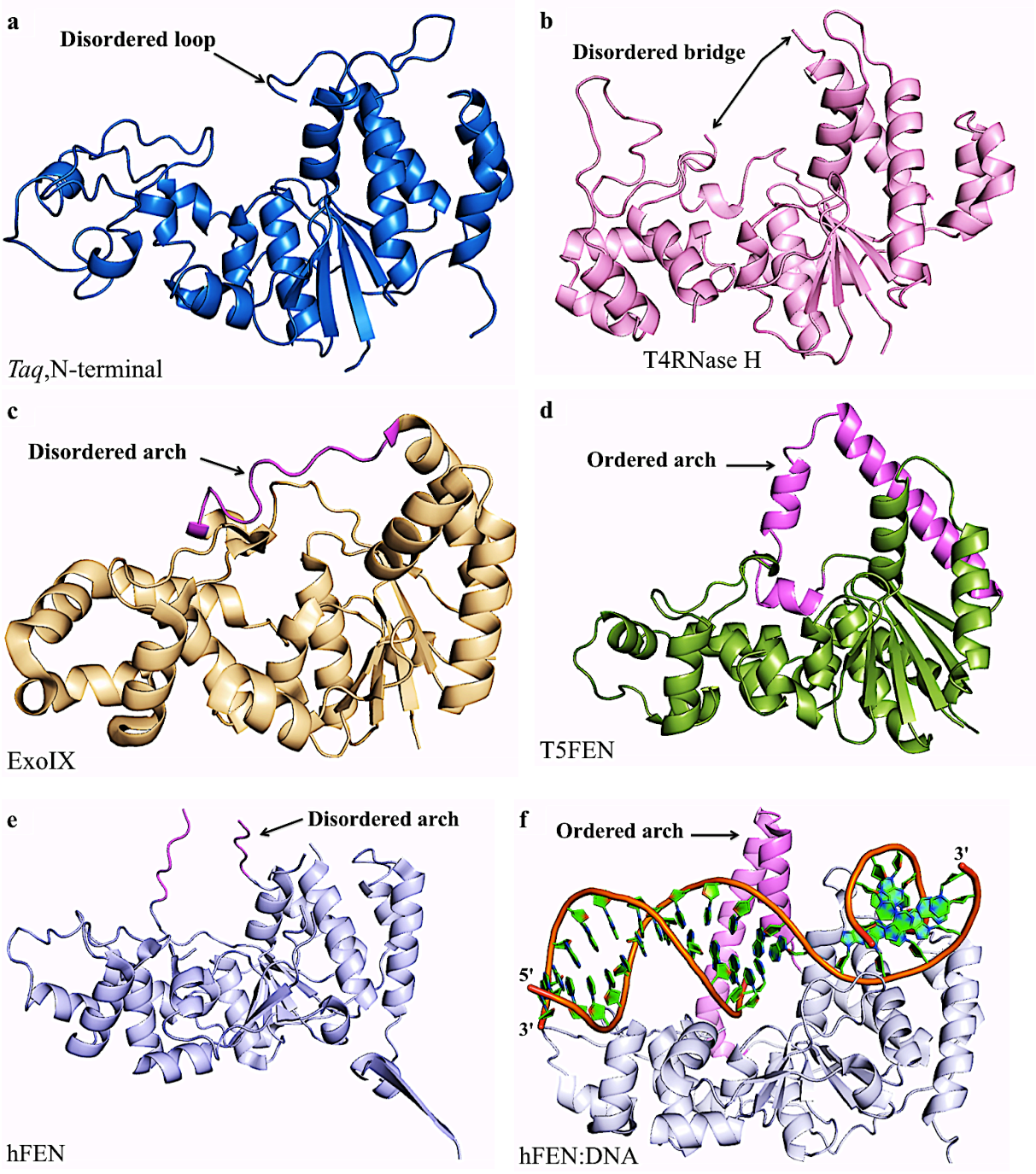



**Figure 1.13:** Cartoons for some FEN members illustrate disordered and ordered arch-like structure.
**(a)** *Taq*-DNA Pol. (1TAQ), **(b)** DNA-free T4RNase H (1TFR), **(c)** DNA-free *E. coli* ExoIX (3ZDE) show disordered arch while **(d)** DNA-free T5FEN (1EXN) illustrates an ordered arch. **(e)** DNA-free hFEN (1UL1) and **(f)** hFEN complex with double-flap DNA (3Q8K) show disordered and ordered arch respectively.

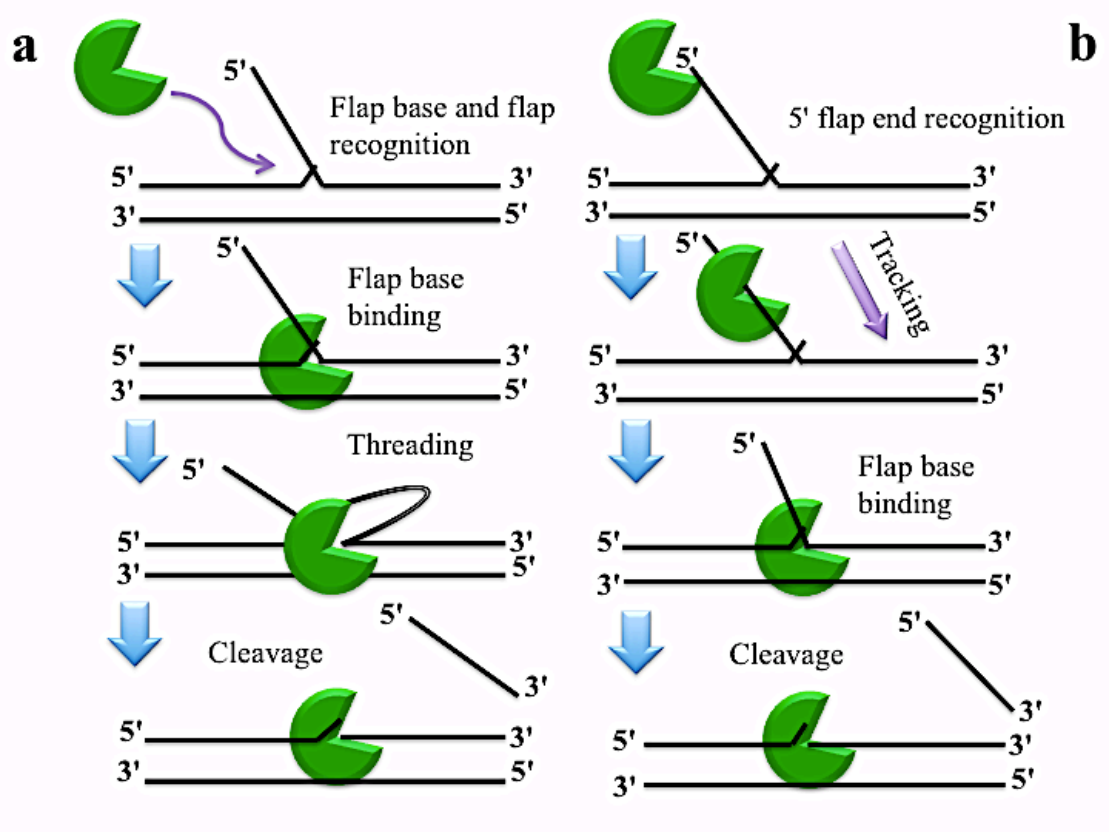



**Figure 1.14:** Diagram explains the threading and tracking models.
**(a)** Shows the threading model in which the enzyme recognized the ss/dsDNA junction first and interacts with it. The 5' flap then pushed through the arch hole to be cleaved. **(b)** Illustrates the tracking model in which the FEN protein interactions with the 5' flap DNA firstly and then slide down the ss/dsDNA area to cleave the flap structure.
The figure was modified from (Gloor et al, 2010).

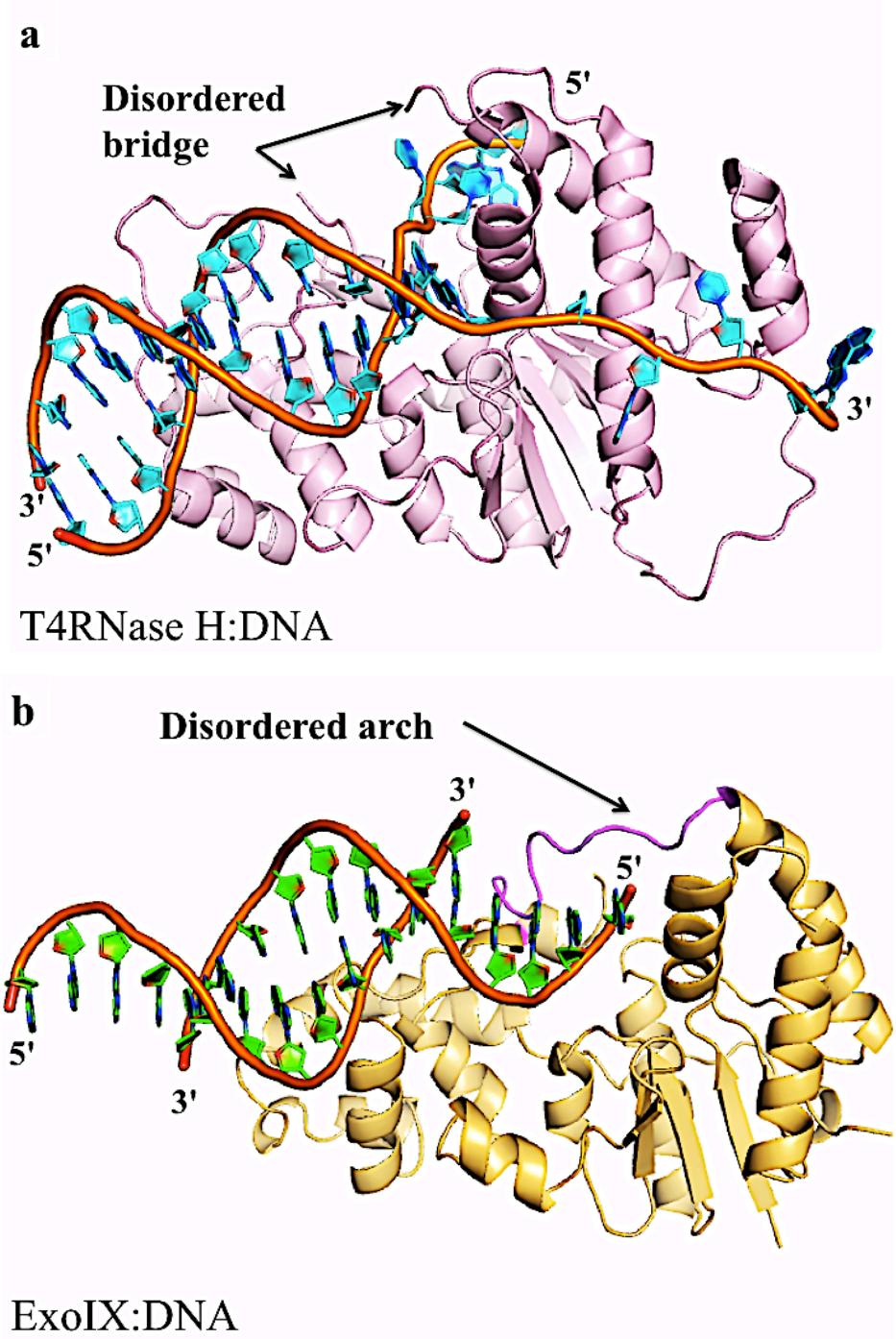



**Figure 1.15:** Cartoon diagrams show the disordered arch in some FEN family members in complexes with DNA substrates.
**(a)** Shows crystal structure for T4RNase H complex with pseudo-Y DNA (2IHN). **(b)** Shows crystal structure for *E. coli* ExoIX complex with 5ov4 DNA (5' overhang DNA, 3ZDC).

More recently in 2012 Grasby and her team discriminated between threading and clamping models biochemically using T5FEN and hFEN-1 and a 53 kDa mass molecular weight streptavidin tetramer protein. Streptavidin was bound to the 5' flap substrate before being mixed with the enzymes and the reaction catalysis was measured. The enzyme activity was decreased for both FENs compared with the activity when streptavidin was added to the preformed enzyme-substrate complex. They argued that the inhibition effect resulted from prevention of the flap from threading through the protein arch as it was blocked with streptavidin. In contrast, in the clamping model the 5' streptavidin should not conflict with the reaction. Furthermore, mixing streptavidin with the 3' flap DNA substrate and then adding the T5FEN does not affect the enzyme activity. According to their findings they conclude that there results and data analysis support the threading model for the 5' flap and disagree with the clamping one. Surprisingly, they demonstrated that the substrates with a short gapped 5' duplex have a similar behave to a 5' flap and must be passed through the arch (Patel et al, 2012). However, it is clear that more investigations are needed in order to determine the exact role of the helical arch region during the FEN enzyme's reaction.

As mentioned previously, complexes with DNA substrates for some FEN family members were observed in order to investigate the interactions between FEN proteins and their DNA substrate which could lead to more understanding of this family member activity mechanisms. *Af*FEN was complexed with a short duplex DNA with a single 3' nucleotide and so it does not have the 5' flap. Interactions between the 5' DNA and *Af*FEN have been observed. The side chains of Arg-64, Asn-67 and Lys-317 form hydrogen bonds with the DNA backbone while Tyr-63 and Arg-64 generate stacking interactions with sugars in the DNA minor groove (Figure 1.16a). In T4RNase H:DNA complex the 5' arm passed under a completely disordered bridge and interacted with Trp-101, Phe-105 and His-109 side chains (Figure 1.16b). Further interactions were observed between hFEN and its double-flap substrate. In this complex the 5' flap was cleaved and presented in the active site which makes it hard to identify which mechanism

of either threading, tracking or clamping is used by this enzyme. However, Tsutakawa et al. (2011) proposed a disorder-to-order mechanism to process 5' flaps. In this complex the monophosphate product interacted with the conserved residues Arg-100 and Lys-93 side chain through hydrogen bonds and with the two $Sm^{3+}$ metal ions that are present in the active site of this complex (Figure 1.16c). In the substrate structure the5' flap was far away from these residues and no any significant interactions were observed (Figure 1.16d). There is no published structure for hEXO-1 without DNA but it has a complex structure with 3' flap DNA and duplex 5' arm (Orans et al, 2011). Some interactions were observed between the protein and the first nucleotide in the 5' arm DNA substrate and product. Arg-96 and Tyr-32 side chain formed hydrogen bonds to thymidine-1 phosphate group and sugar ring oxygen in substrate complex (Figure 1.16e) while Arg-92 and Tyr-32 side chains interacted with the thymidine-1 base in the product structure and Lys-85 side chain formed hydrogen bonds with the dT-1 phosphate oxygen (Figure 1.16f). Two $Mn^{2+}$ ion are present in the active site and interacted with the dT-1 phosphate group (Figure 1.16f).

#### *1.2.1.4 FEN Active Site*

The active site of FEN family members is located in the centre of the enzyme in front of the helical arch motif. It is composed of a cluster of conserved acidic residues that coordinate two divalent metal ions (Anstey-Gilbert et al, 2013; Ceska et al, 1996; Chapados et al, 2004; Hosfield et al, 1998; Hwang et al, 1998; Mueser et al, 1996; Orans et al, 2011; Sakurai et al, 2005). Based on biochemical studies most nuclease enzymes require divalent metal ions to catalyze their reaction and FEN members are an example of these nucleases. Most FENs possess two metal binding sites referred to as M1 and M2. The position of the M1 metal-binding site is highly conserved in the FEN structures while M2 is present in variable positions in each structure as observed by superposition of the active site for three members of the FEN family: T5FEN (Ceska et al, 1996), T4 RNase H (Mueser et al, 1996), *Mj*FEN (Hwang et al, 1998) and hFEN (Sakurai et al, 2005) (Figure 1.17a) (Syson et al, 2008).

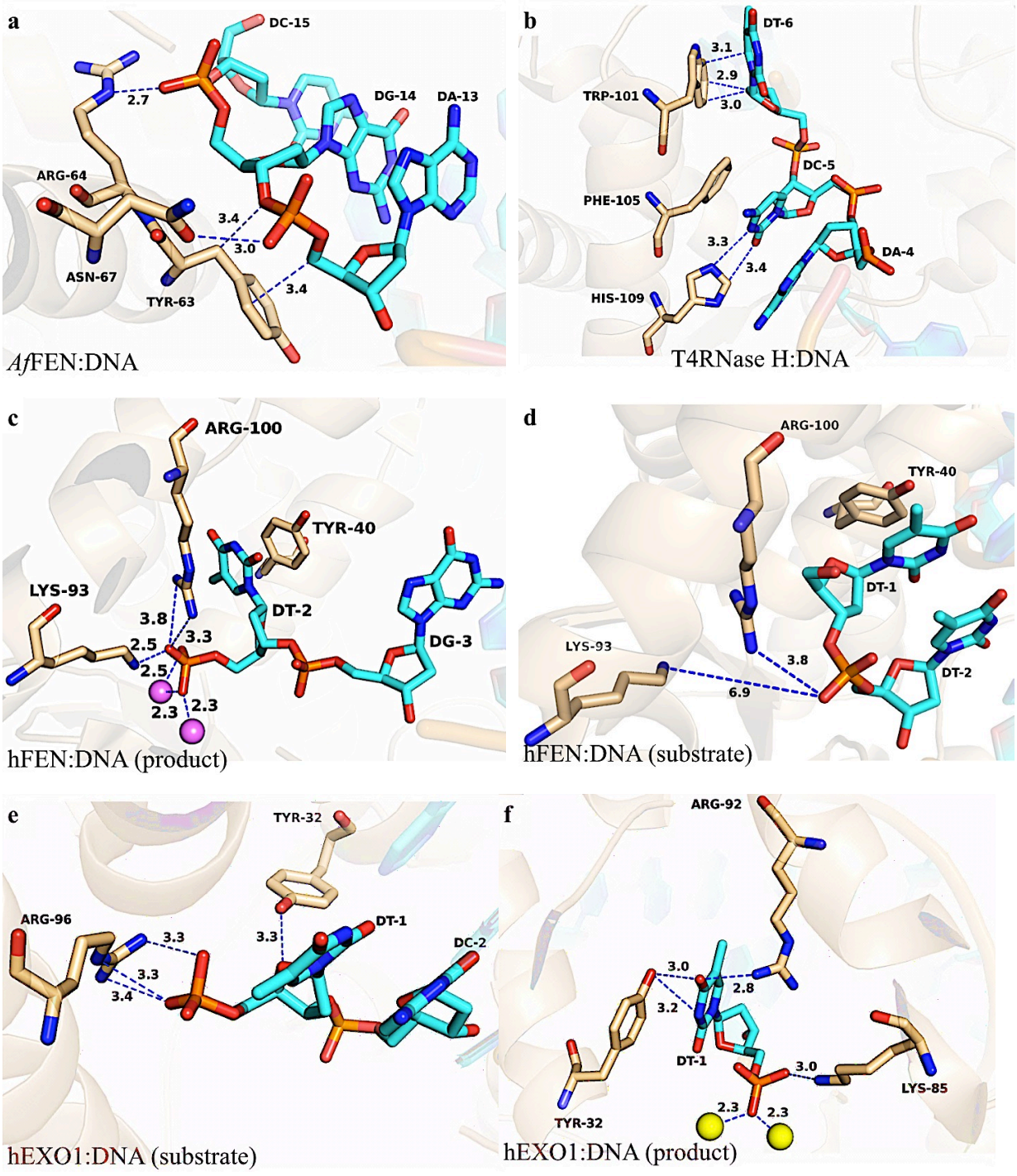



**Figure 1.16:** The 5' flap interactions in some FEN proteins in complexes with DNA substrates.

**(a),** and **(b)** illustrate the interactions in *Af*FEN:DNA (1RXW), T4RNase H:DNA (2IHN). **(c)** and **(d)** show the 5' flap interactions in hFEN:DNA substrate (3Q8L) and product respectively. **(e)** and **(f)** show the interactions between hEXO-1 and the 5' base in DNA substrate (3QEA) and product (3QEB) respectively.

The exact mechanism by which this group of enzymes processes the DNA substrates was controversial for long time. Based on the results of several biochemical and biophysical studies, a two metal ion mechanism was proposed for FEN enzymes similar to that suggested by Beese and Steitz in 1991 for the 3'→5' exonuclease domain of the Klenow fragment of *E. coli* DNA Pol I. In the three-dimensional structure of the Klenow fragment complexed with a part of DNA molecule, the interactions between the two metal ion of the 3'→5' exonuclease domain and the target phosphodiester bond are observed (Figure 1.17b). The metal ion in site I would act as a lewis acid to promote the formation of the hydroxide ion in opposite of the phosphorous in the target phosphodiester bond by attacking a water molecule. Hydroxyl ion lone pair of electrons is oriented through interactions with Glu-357 and Tyr-497 residues *via* hydrogen bonds allowing them towards the phosphorous of scissile phosphodiester bond to be cleaved. The site II metal ion does not have a catalytic role but it is stabilized the pentavalent transition state reaction intermediate (Beese & Steitz, 1991).

On the other hand, some issues are raised against this mechanism for FEN activity. Firstly, the two metal ion mechanism requires less than 4Å distance between these metal ions but all FEN crystal structures determined in complex with divalent metal ions, the two ions are separated with greater than 4Å (Figuer 1.17a) except for hFEN in complex with $Mg^{2+}$. Secondly, there are several proofs suggested that metal ion in M2 site does not play a catalytic role preferring a one-metal ion mechanism (Amblar et al, 2001; Bhagwat et al, 1997; Tomlinson et al, 2011; Zheng et al, 2002). Recently, the prokaryote FENs active site divided to two catalytic sites called Cat1 and Cat2. Cat1 has two metal-binding sites referred to as M1 and M2 while Cat2 has only one metal-binding site known as M3 (Anstey-Gilbert et al, 2013). A crystal structure for T5FEN in a complex with magnesium ions has been determined by Flemming (2011). In this structure the T5FEN-WT binds to five $Mg^{2+}$ ions presented in two molecules, two of these metal ions occupied the Cat1 site and separated by 3.6 Å in the first T5FEN molecule.

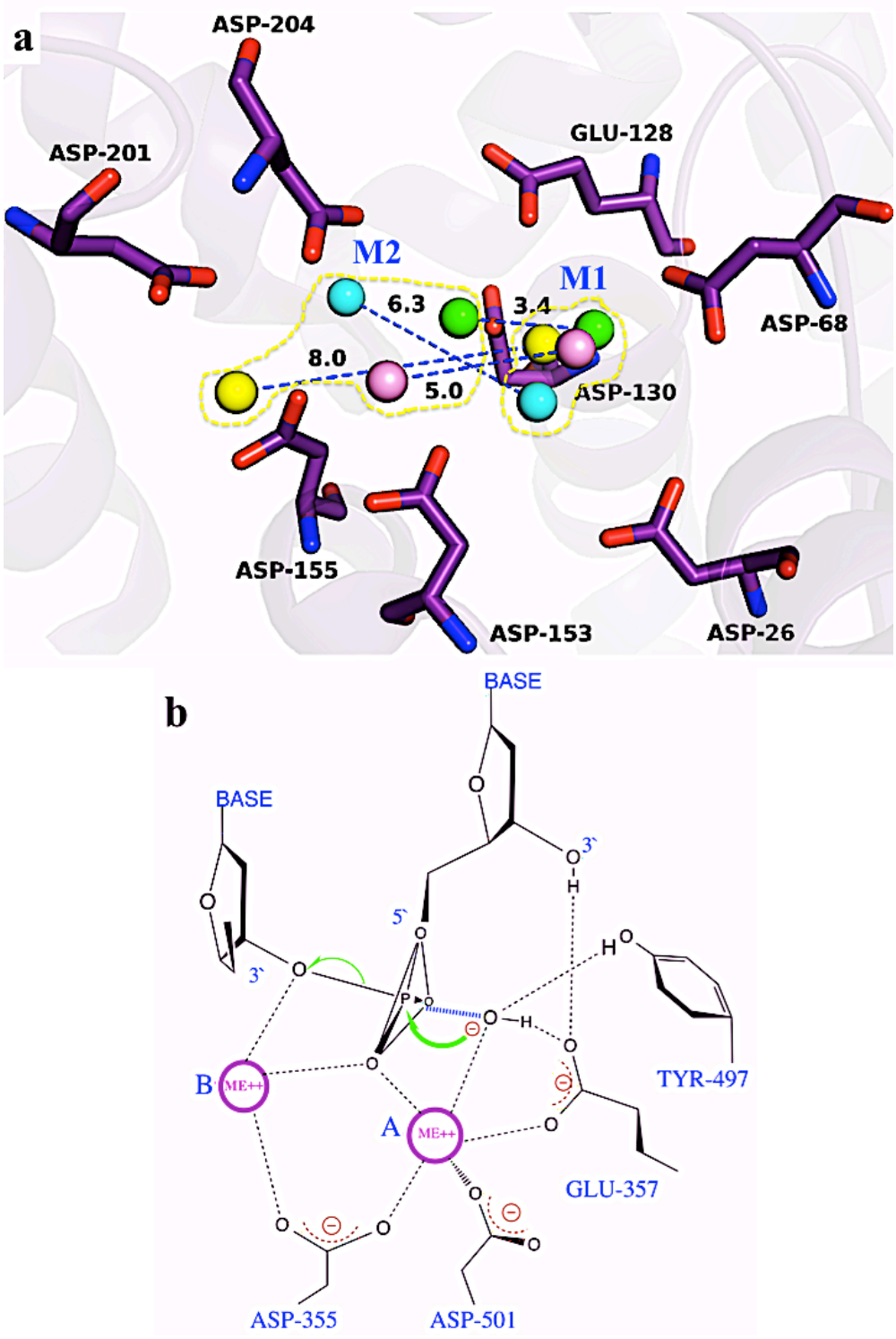



**Figure 1.17:** Diagrams for FEN active site and the two metal ion mechanism.
**(a)** Shows superposition of the active site metal ions in M1 and M2 of four FEN members, T5FEN (IUT5, light blue), T4RNase H (1TFR, cyan), *MjFEN* (1A77, rose) and hFEN1 (1UL1x, green). **(b)** Explanation of the two metal ion mechanism as described for 3' → 5' exonuclease activity in klenow fragment.
The figures were adapted from (Syson et al, 2008) and (Beese & Steitz, 1991) respectively.

The second molecule has three $Mg^{2+}$ ions one of the occupied M1 in Cat1 and the second is present in the Cat2, M3, site while the third $Mg^{2+}$ ion is lied in a place close to a partially disordered arch (Figure 1.18). Superposing of these two molecules together (Figure 1.18e) can give a clear picture for the possible sites of metal ions in T5FEN (Flemming, 2011). In hFEN:DNA (Tsutakawa et al, 2011) and hEXO1:DNA (Orans et al, 2011) substrates structures the metal ions in the active sites were separated by 4.6 Å and 4.4 Å respectively (Figure 1.19a) while in the product complexes the metal ions seems to be moved inside the active site toward the catalytic reaction position and were present within 4 Å distances (Figure 1.19b). Anstey-gilbert and his colleagues (2013) were reported a small distance, 2.5 Å between two $Mg^{2+}$ in the active site of *E. coli* ExoIX in complex with DNA product (Figure 1.19b).

## 1.3 T5 5' Flap Endonuclease (T5FEN)

### 1.3.1 The Discovery and Characterisation

When *E. coli* was infected with bacteriophage T5 a new deoxyribonuclease was observed differing from those present in the uninfected host cell. The enzyme was purified and described as requiring pH 9.3, $Mg^{2+}$ or $Mn^{2+}$ and having both endo- and exonuclease activity (Moyer & Rothe, 1977; Paul & Lehman, 1966). Furthermore, it is hydrolyzed single or double stranded DNA specifically from the 5' end producing oligonucleotides of different lengths (Frenkel & Richards.Cc, 1971). The gene, D15, was cloned and sequenced by Kaliman and his team in 1986 but they could not express it or identify any sequence typical to *E. coli* promoter. In 1990 Sayers and Eckstein used sequence alignments to compare the bacteriophage T5 exonuclease with the N-terminal domain of DNA Pol I from some prokaryotes including *E. coli, Thermus aquaticus* and *Streptococcus pneumoniae.* They found significant sequence homology and four areas of conserved residues were identified between these proteins (Sayers & Eckstein, 1990).

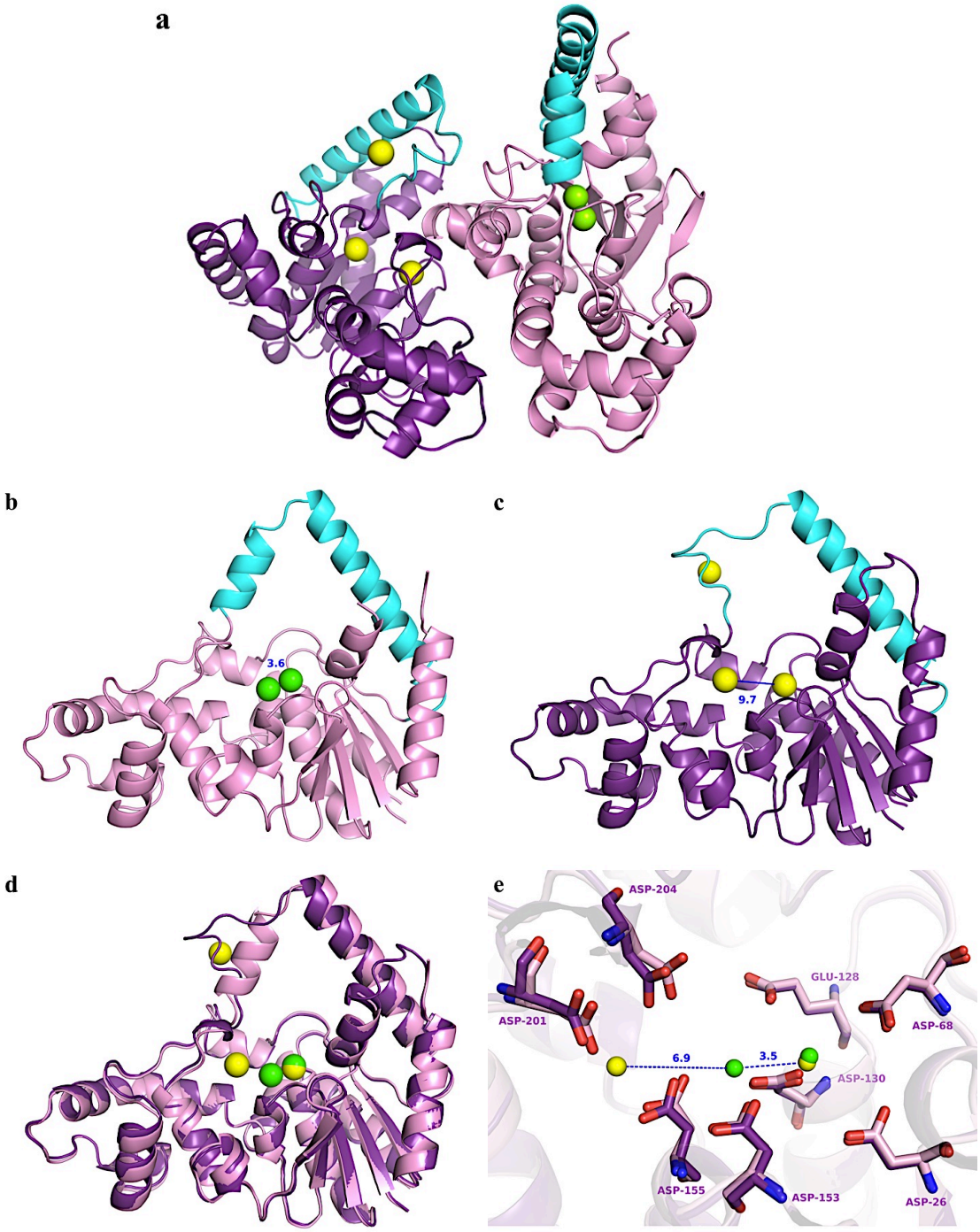



**Figure 1.18:** T5FEN:$Mg^{2+}$ native crystal structure.
**(a)** Shows the two molecules in the asymmetric unit. The arch is shown in cyan. **(b)** Illustrates the first molecule (pink) with two $Mg^{2+}$ ions (green spheres) in its active site. **(c)** The second molecule (purple) with three $Mg^{2+}$ ions (yellow spheres). **(d)** Shows superposition of the two molecules while **(e)** illustrates the superimposed active sites with four $Mg^{2+}$ ions. Two magnesium ions (green spheres) occupy M1 and M2 in Cat1 site while the other (yellow spheres) lie in M2 of Cat1 and M3 of Cat2 sites.

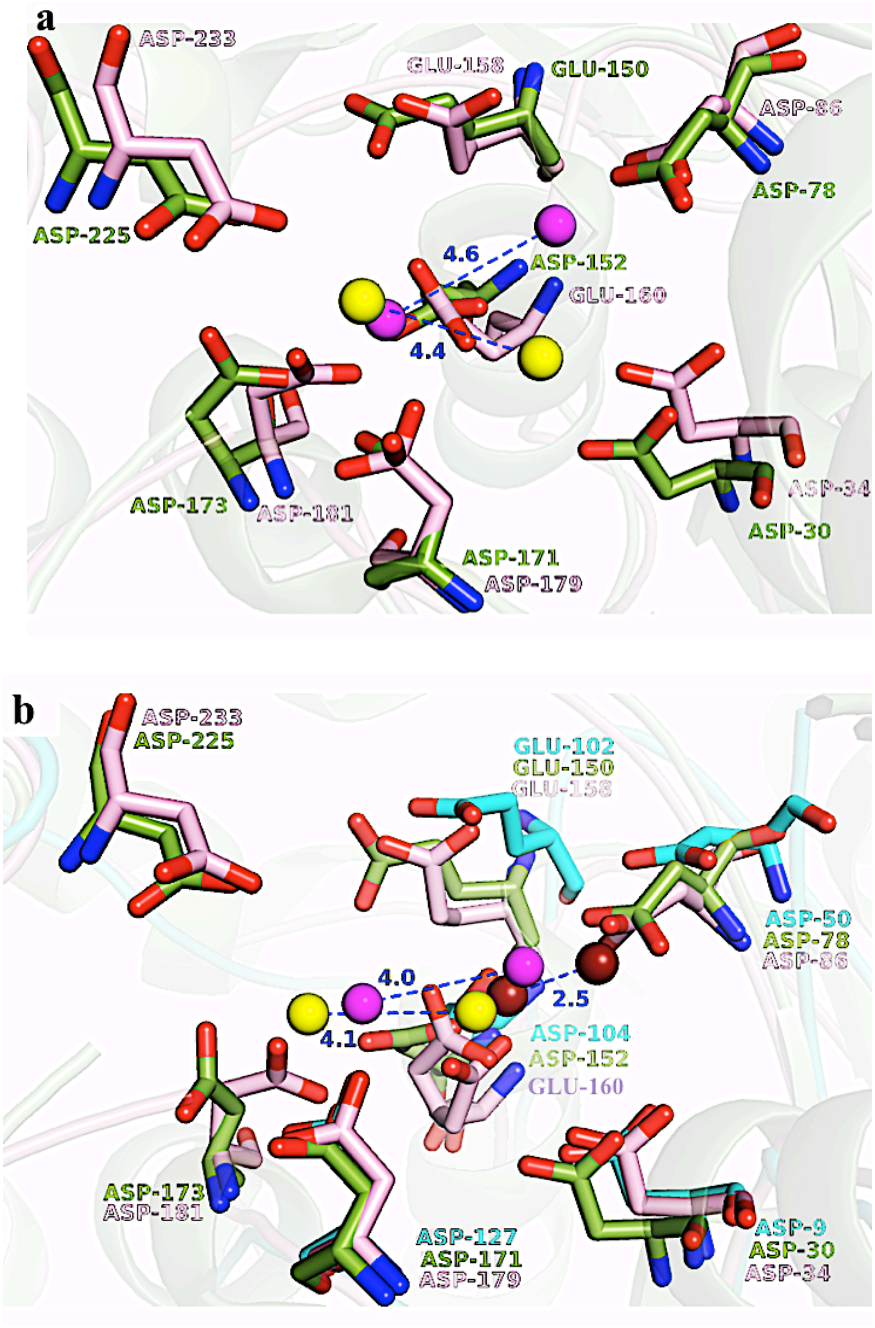



**Figure 1.19:** The active sites and metal ions in some FEN members.
**(a)** Superposition of hFEN:DNA(3Q8K) active site (pink sticks) on hEXO1:DNA (3QEA) active site (green sticks). The figure shows $2Sm^{3+}$ (magenta spheres) and $2Ba^{2+}$ (yellow spheres) for hFEN and hEXO1 respectively in the superposed active site separated with greater than 4 Å in substrate structures. **(b)** Shows the same superposition for the product structures with *E. coli* ExoIX:DNA (3ZDC) (cyan sticks) product structure. The figure illustrates the distances between the metal ions: $2Sm^{3+}$ (magenta spheres), $2Mn^{2+}$ (yellow spheres) and $2Mg^{2+}$ (light blue spheres) in the superposed active site of hFEN, hEXO1 and *E. coli* ExoIX respectively.

T5FEN is a 33 kDa molecular mass and the protein was expressed, purified, crystallised and the structure determined at 2.5 Å resolution by Ceska and coworkers (Ceska et al, 1996). Biochemical and mutagenesis studies have been done to investigate the T5 flap endonuclease characterization and mechanism. The presences of cofactor divalent metal ions in the protein active site are required for hydrolysis to occur but they are not necessary for DNA substrate binding (Garforth & Sayers, 1997). A structure-specific cleavage assay that was developed in 1999 by Pickering and colleagues for T5FEN activity has been used to examine the effect of the three lysine residues mutations (K83A, K196A and K215A) (Pickering et al, 1999). Compared to the wild type the K215A mutation does not affect the enzyme's ability to cleave the substrate while it decreased about 20-fold with the K196A mutant and a significant reduction has been reported with the K83A mutant. All of these mutations are reported to contribute to the stability of the enzyme-substrate complex and the K83A mutant has a significant role in the catalytic reaction. Additionally, K83 was demonstrated to be required for exonucleolytic activity on single-stranded DNA but is not essential for endonucleolytic activity (Garforth et al, 1999; Paul & Lehman, 1966).

The sensitivity of T5FEN activity at different pH levels and with various metal ion concentrations and types were observed. At acidic environment and low divalent metal ions concentration or presence of $Co^{2+}$, $Zn^{2+}$ or $Ni^{2+}$ the exonuclease activity is greatly reduced. It is stimulated at basic environment and seemed to require higher concentrations of metal ions or presence of $Mg^{2+}$ or $Mn^{2+}$ in reaction medium (Feng et al, 2004; Garforth et al, 1999; Garforth et al, 2001). On the other hand, if the concentration of metal ions cofactor exceeds 25 mM the enzyme loses its activity totally and the reaction is inhibited (Paul & Lehman, 1966; Tock et al, 2003) and this fact was observed for other FENs (Zheng et al, 2002). Interestingly, T5FEN cleaves double-stranded closed-circular plasmid with $Mn^{2+}$ cofactor (Garforth et al, 2001). Also the product-metal ion-protein complex is demonstrated to be more stable with $Co^{2+}$ and $Mg^{2+}$ by two-fold than with $Mn^{2+}$ (Tock et al, 2003). According to these findings, Garforth et al. (1999)

argued that the two-nuclease activities of T5FEN are distinct and different mechanisms may be used for endo- and exo- hydrolysis. In addition, the two metal-binding sites in T5FEN might bind cofactor with different affinities and as a result mutations in the amino acids that coordinate metal ions in Cat1 and Cat2 have been created (Feng et al, 2004). The results of these mutations suggested that if two residues in Cat1 are mutated (E128Q D130N) and then the enzyme is tested for substrate binding and catalysis activity the protein still has ability to bind a pseudo-Y DNA substrate but hydrolysis activity is totally lost (Feng et al, 2004). Indeed, the E128Q D130N modified protein bound DNA significantly more tightly than the WT.

In contrast, mutations in residues that coordinate a metal ion in the Cat2 site resulted in a decrease or lack of exonuclease activity but the enzyme still acts as an endonuclease. Based on isothermal titration calorimetric analysis, Feng et al. conclude that the Cat1 site has 30-fold higher affinity for cofactors and it is occupied first, which explained the selectivity of catalysis activity in different metal ions concentrations (Feng et al, 2004). For the T5FEN mechanism it has been demonstrated that the enzyme requires at least three metal ions in its reaction (Syson et al, 2008). Two metal cofactors are responsible for the hydrolysis activity while the third one appears only when the DNA substrate presents in the reaction and could bind to Asp-130.

In 2011 Tomlinson and colleagues reported that the data analysis for mutated protein D201I D204S indicated that a single divalent cofactor is required for catalysis reaction and at least one $Ca^{2+}$ ion is needed to inhibit T5FEN-WT reaction. The authors concluded that the two metal ion in T5FEN are required to place the substrate in the reactive form and the Cat1 is responsible for the chemical catalysis while Cat2 has a role in substrate-positioning only. As mentioned above Flemming (2011) has solved a T5FEN-WT structure binds to five magnesium metal ions presented in two molecules and the superposition of these molecule clarified the positions of these metal ions in the active site (Figure 1.18e) (Flemming, 2011).

### 1.3.2 The Crystal Structure of T5FEN

The 3D structure of T5FEN has the typical FEN topology (Figure 1.7a) that is composed of core $\beta$-strands surrounded by $\alpha$-helices with the active site in the centre of the enzyme. The N-terminal region of bacterial and bacteriophage FEN structures is disordered. The 19 N-terminal amino acids in the T5FEN structure can be removed without affecting the enzyme activity. This region is thought to provide a site for protein-protein interactions in other nucleases (Bhagwat et al, 1997; Ceska et al, 1996; Kim et al, 1995). An arch-like structure is present at the back of the enzyme and above the active site which is known as helical arch (Figure 1.7a). It is built from three $\alpha$-helices lined with positively charged residues (Lys-83, Arg-86, Lys-89 and Arg-93) and contains the absolutely conserved Lys-83, Arg-86 and Tyr-82 amino acids. The rear of the arch ($\alpha$-4) has the negatively charged Asp-87 and Glu-88 residues. The helical arch region forms a hole about 18 Å in diameter, large enough to accommodate a single stranded DNA flap and this hole is clearly observed in this structure because the arch region is well ordered. The architecture of this motif with these amino acids is thought to form interactions with the DNA backbone from one side and with the bases from other.

The enzyme active site has a concave shape and is located in front of the arch motif and consists of eight conserved acidic residues. In order to investigate the metal ions binding sites in T5FEN active site the crystal was soaked in 25 mM $MnCl_2$ before collecting the data (Ceska et al, 1996). Two metal sites were observed and referred to as M1 and M2 which more recently have been renamed as Cat1 and Cat2. In T5FEN a metal ions in Cat1 coordinated with five acidic residues (Asp-26, Asp-86, Glu-128, Asp-130 and Asp-153) while the ion in Cat2 binds to Asp-153, Asp155, Asp-201 and Asp-204. These metal ion-binding sites (Cat1 and Cat2) are separated from each other by 8 Å which is greater than 4 Å required for two metal ion mechanism. Furthermore, authors soaked a crystal in both 25 mM $MnCl_2$ and 2 mM $ZnCl_2$ and then data was collected and analyzed which was identified only a binding site for $Zn^{2+}$ in Cat1 (Ceska et al, 1996).

The superposition of these crystal structures active sites on T5FEN:$Mg^{2+}$ active site shows the high affinity of M1 in Cat1 site for metal ion more than M2 and M3 in Cat1 and Cat2 as seen in Figure 1.20a. The ds-DNA binding motif helix-three-turn-helix (H3TH) has been observed (Artymiuk et al, 1997) in front of the active site. It is composed of two antiparallel helices (Figure 1.20b) built from amino acids Val-189 to Ile-198 and Ala-214 to Phe-224 linked together by a flexible loop formed with 14 residues (Met-199 to Gly-213). This loop contains two aspartic acids Asp-201 and Asp-204 cross to the enzyme centre and engaged in Cat2 site formation. In addition, the H3TH motif is superimposed very well with the helix-hairpin-helix (HhH) motif from polymerase $\beta$ (PDB code IBPE, residues T93-E117). According to the superposition analysis the main difference between these two motifs is observed in the loop which is longer in H3TH motif (Artymiuk et al, 1997).

## 1.4 The Aims of the Project

Exactly how the FEN enzymes recognize and cut their substrate DNAs is unclear. A structure of the protein with DNA substrate and metal ions would provide useful information about the complex. Although eight structures of FENs have been solved and published most are of *apo* forms (see section 1.2). A few, five, structures of complexes with DNA substrates or products have been produced however, these complexes are limited to those with DNA products in which their 5' flaps were cleaved such as hFEN:DNA or for substrates lacking the 5' flap or overhang such as *Af*FEN:DNA and hEXO1:DNA or for FEN with a disordered arch as T4RNase H:DNA or the 5' single strand did not reach the arch region as *E. coli* ExoIX:DNA (Anstey-Gilbert et al, 2013). This makes it hard to investigate the whole details about DNA binding to these proteins. Additionally, the role of the arch-like structure in recognizing the ss-5' flap or overhang is also still a puzzle.

The primary aim of this project was to determine the crystal structures of FEN family members in complex with substrate DNA and/or in the presence of metal ions.

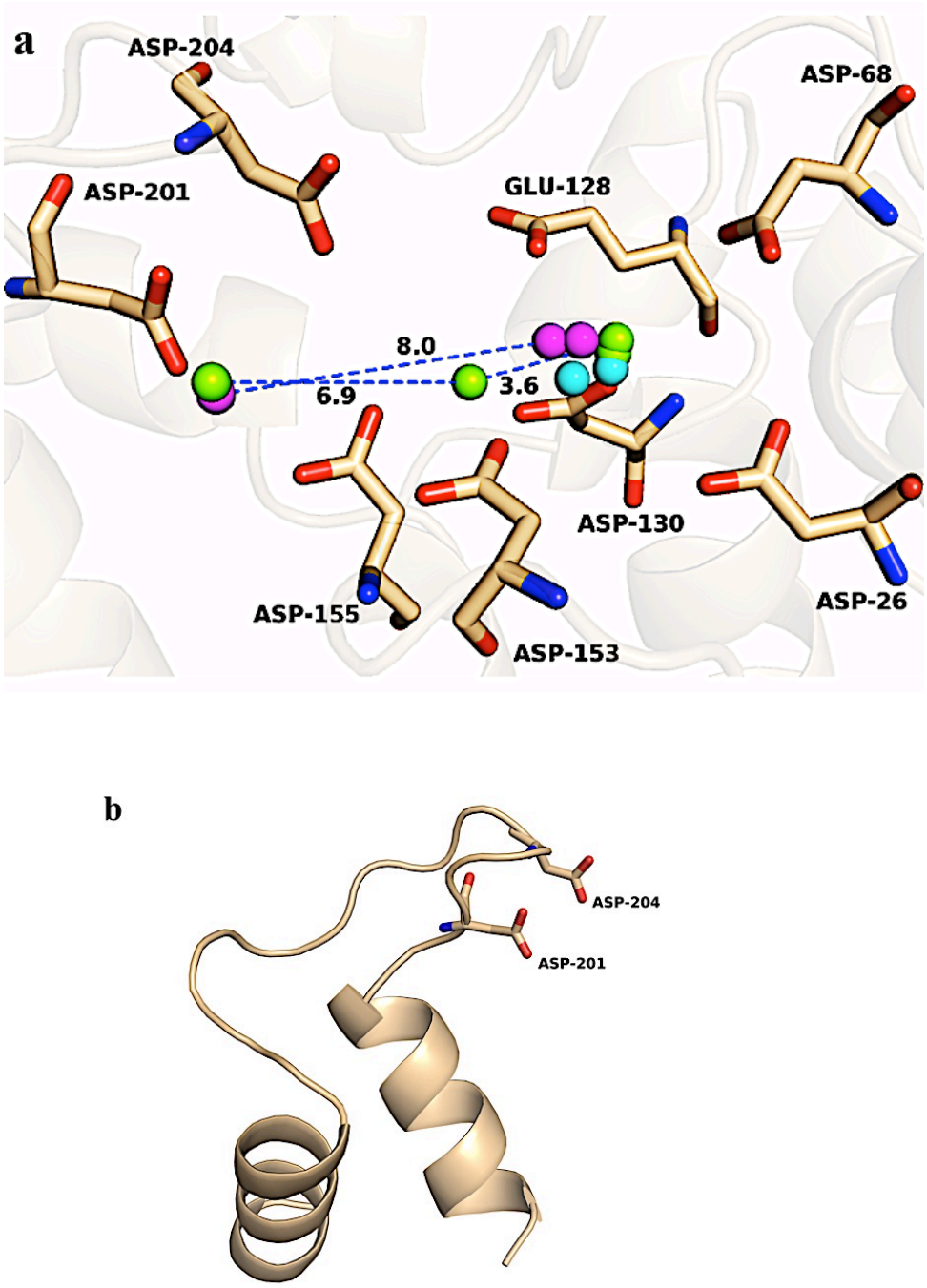



**Figure 1.20:** T5FEN active site metal ions from variety structures and the H3TH motif.

**(a)** Shows the superposition active site for some T5FEN crystal structure in complexes with metal ions: $2Zn^{2+}$ (1UT8, cyan spheres), $3Mn^{2+}$ (1UT5, magenta spheres) and $3Mg^{2+}$ (yellow spheres). **(b)** Shows the H3TH motif and the two active site residues Asp-201 and Asp-204 are extended from its loop.

The FENs to be were investigated were;

1. Bacteriophage T5FEN
2. hFEN1
3. *T.brucei* FEN

T5FEN has been chosen to complex it with DNA substrates because it has a visible ordered arch. In order to make these complexes and study the interactions with the helical arch the enzyme must be in its catalytically inactive form if metal ions are present otherwise the substrate would be cleaved. Previous studies in this laboratory using wild type protein failed to yield diffracting crystals. Therefore, two T5FEN variant proteins having mutations in their active site residues in which aspartic acid was replaced with lysine to prevent the metal ions from binding to Cat1 or Cat2 sites were used. These proteins allow the binding of the DNA substrate without any activity (Zhang, 2012).

Human flap endonuclease (hFEN1) which is a FEN superfamily member and determined as a cancer susceptibility factor has been crystallized with a double flap DNA but in this complex the 5' flap was cleaved and did not directly show whether the 5' ssDNA threaded, or was clamped or tracked by the enzyme (Tsutakawa et al, 2011). An inactive version of hFEN, D179K, was supplied by Jing Zhang (Zhang, 2012) to investigate the arch role in this enzyme as an example for FEN superfamily.

There have been no reported structures of eukaryotic FENs apart from that of hFEN. The flap endonuclease from *Trypanosoma brucei*, a parasite causing sleeping sickness was also studied. This flap endonuclease enzyme shares 46% identity with hFEN. It was hoped that solving this structure as *apo* protein and in complexes with DNA substrate would open a new window to design molecules that may be able to inhibit *T. brucei* FEN but do not affect the human one.

# Chapter 2: Material and Methods

The experimental work that was done during this project including expression and purification of T5-FEN for both the wild type and the T5FEN-D153K variant, hFEN purification and *Trypanosome brucei* flap exonuclease transformation, expression and purification are described in this chapter. Details of crystallisation and cryoprotection procedures are also explained.

## 2.1 Microbiological Methods

### 2.1.1 Growth Media

Three types of media were used. Lysogen-Broth media (Bertani, 1951) consisted of 10 g/l Bact-tryptone, 5 g/l yeast extract and 10 g/l NaCl. 2xYT media consisted of 16 g/l Bacto-tryptone, 10 g/l yeast extract and 5 g/l NaCl. Media were supplemented with 100 $\mu$g/ml carbenicillin when antibiotic selection was required. 4xYT media consisted of 32 g/l Bacto-tryptone, 20 g/l yeast extract and 10 g/l NaCl. Media were supplemented with 0.1 % glucose as a carbon source and 100 $\mu$g/ml ampicillin when antibiotic selection was required. LB agar was made by addition of agar (1.5 % w/v) to LB media.

### 2.1.2 Strains and Plasmids

An *E. coli* M72 strain carrying the pJONEX44 expression vectors (Sayers & Eckstein, 1990) with the protein coding sequence for T5 5' flap endonuclease (FEN) wild type (pJONEX44Δ19) or the T5FEN-D153K variant (pJONEX44Δ19D153K) between EcoRI and HindIII restriction sites were kindly provided by Prof. Sayers lab, Medical School, University of Sheffield (Feng, 2002; Zhang, 2012). Additionally, BL 21 (DE3) cell pellets have the expressed hFEN-D179K protein was provided by Zhang and Sarah Oates, from Prof. Sayers lab. Another strain XL1 blue transformed with PET-21a vector carrying protein-coding sequence for *Trypanosoma brucei* FEN wild type, (protein accession number XP_843679.1) was also provided

by Prof. Sayers lab, Medical School, University of Sheffield. This plasmid, pTBFEN, encoded the *T. brucei* FEN wild type full length and was codon-optimized for expression in *E. coli* by Eurofins-MWG as a custom gene synthesis. The *E. coli* strain XL1 blue was used as a primary host for transformation of the pTBFEN plasmid. The genotype for each strain is presented in Table 2.1.

Long term storage of transformed strains was done by flash freezing overnight cultures in LB and 4xYT media containing (20 % v/v) glycerol which was then stored at -80°C.

**Table 2.1:** *E. coli* strains genotype.

| ***E. coli* strain** | **Genotype** |
|---|---|
| M72 | Δ(bio-uvrB) lacZ(Am) rspL Δ(trpEA2) (λ Nam7 Nam53 c1857 ΔH1) |
| BL 21 (DE3) | $F^-$ ompT gal dcm lon $hsdS_B(r_B^- m_B^-)$ λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5]) |
| XL1-Blue | endA1 gyrA96($nal^R$) thi-1 recA1 relA1 lac glnV44 F'[ ::Tn10 $proAB^+$ $lacI^q$ Δ(lacZ)M15] hsdR17($r_K^- m_K^+$) |

## 2.1.3 Bacterial Culture

Aerobic growth was carried out in 4 liter conical flasks containing 500 ml of medium and capped with foam bungs and aluminum foil. All vessels and media were autoclaved before use. Growth was monitored by determining absorbance (turbidity) at 600 $_{nm}$ in a spectrophotometer. 2xYT media was used to express T5FEN proteins while 4xYT media was used for *Trypanosome brucei* FEN expression.

### 2.1.4 Plasmid Extraction

The pTBFEN-393 expression plasmid was extracted from XL1 blue cells using a miniprep kit (QIAprep Spin Miniprep Kit, Qiagen) and transformed into *E. coli* BL 21 (DE3) expression strain.

### 2.1.5 Transformation of Cells

The transformation procedure was carried out according to the method of Hanahan (Hanahan, 1983). 200 µl of BL 21 (DE3) competent cells were thawed on ice before the vector was adding (up to 100 ng) and gently mixed. The cells were incubated on ice for 30 mins and then subjected to a 60 seconds heat shock at 42°C and a further two mins on ice. 0.8 ml of LB medium prewarmed to 37°C was added followed by incubation for an hour at 37°C before spreading 50 µl of transformed cells on LB agar plates supplemented with 100 µg/ml ampicillin. The plate was incubated overnight at 37°C.

## 2.2 Protein Methods

### 2.2.1 Overexpression

#### *2.2.1.1 T5FEN*

The T5FEN-D153K variant and T5FEN-WT were expressed using the following protocol: *E. coli* M72 (pJONEX44D19D153K or pJONEX44D19) cultures were streaked on LB plates containing 100 $\mu$g/ml of carbenicillin and incubated at 25°C overnight. One colony was picked and grown in a 40 ml culture of 2xYT media containing 100 $\mu$g/ml of carbenicillin at 25°C overnight. The overnight culture was then used to inoculate 8 x 500 ml cultures of 2xYT media and 100 $\mu$g/ml of carbenicillin in 2 liter conical flasks with 5 ml preculture per flask. This was incubated at 30°C, 250 rpm until the $OD_{600\ nm}$ was approximately 0.7. The expression was then induced by growing the culture at 42°C, 250 rpm for an hour and at 25°C, 250 rpm for an additional four hours. Cells were harvested by centrifugation at 14,500 x g for 10 minutes at 4°C and pellets were stored at -20°C until used.

A purified protein solution for T5FEN-D155K variant was kindly supplied by Jing Zhang, Medical School, University of Sheffield. The concentration of the protein solution was 48 mg/ml in 1.2 ml storage buffer, 10 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM DTT, and 50% glycerol and this was saved at -20ºC until needed.

In order to use this variant protein for crystallisation trials it was dialyzed at 4°C against a storage buffer lacking glycerol, initially for two hours and then the buffer was changed and kept overnight at 4°C. To mix the protein with the DNA substrate the protein buffer was changed to the oligonucleotide-dissolving buffer (10 mM MES pH 6.5 and 50 mM KCl) by dialyzing. The protein was then concentrated down to the required working concentration which is 15 mg/ml using a 30 kDa molecular weight cut off Centricon centrifugal ultrafilteration device (Vivascience). Centrifugation at 4000 x g was carried out at 4°C until the desired concentration was achieved.

#### *2.2.1.2 Human 5' Flap exonuclease (hFEN-1)*

Jing Zhan and Sarah Oates, Medical school, University of Sheffield, kindly provided protein solution and BL 21 (DE3) cell pellets for truncated variant hFEN-D179K. The protein was kept at 53 mg/ml in 10 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM DTT, 0.5 mM EDTA and 50% glycerol at -20°C. The protein solution was dialyzed against the storage buffer without glycerol at 4°C for about two hours and then the buffer was changed and left overnight to remove all of the glycerol. Centrifugation at 4000 x g was carried out at 4°C in 30 kDa molecular weight cut off Centricon centrifugal ultrafilteration device (Vivascience) until the required concentration was obtained.

#### *2.2.1.3 Trypanosome brucei FEN-WT*

One colony of BL 21 (DE3) transformed with PET-21a vector was picked and grown in 40 ml culture of 4xYT media containing 100 µg/ml ampicillin and 0.1% glucose at 37ºC overnight. The overnight culture was then used to inoculate 8 x 500 ml cultures of 4xYT media, 100 $\mu$g/ml of ampicillin and

0.1% glucose in 2 liter conical flasks with 5 ml preculture per flask. This was incubated at 20°C, 250 rpm until the $OD_{600\ nm}$ was about 0.7. The expression was then induced by adding 1 mM IPTG and growing the culture was continued at 20°C, 250 rpm overnight. Cells were harvested by centrifugation at 14,500 x g for 10 minutes at 4°C and pellets were stored at -20°C until used.

## 2.2.2 Purification

### *2.2.2.1 T5FEN Proteins*

Purification of the T5FENΔ19-D153K variant and T5FEN-WT were carried out with Dr.Svetana Sedelnikova using the same protocol that was used by Flemming (Flemming, 2011). Firstly, the cell pellet was thawed and resuspended in 5 volumes of lysis buffer (50 mM Tris-HCl pH 8.0, 0.5 M NaCl) then the cells were lysed by sonication 3 x 20 sec. at 16-micron amplitude, Soniprep 150, on ice. The insoluble material was then spun down by centrifugation at 72,606 x g for 15 minutes at 4°C and the supernatant was collected and diluted four volumes with 50 mM Tris-HCl pH 8.0 to reduce the concentration of NaCl to 0.1 M. A 5 ml Heparin-HP Sepharose FastFlow (GE Health care) column was equilibrated with 50 mM Tris-HCl pH 8.0, 0.1 M NaCl and then protein was loaded onto this column. The column was washed with three volumes of equilibration buffer and the protein was eluted with a 1 M NaCl gradient in 50 mM Tris-HCl pH 8.0. The protein concentration of the peak fractions were checked using the Bradford Assay (Bradford, 1976) and samples applied to an SDS-PAGE gel to confirm the presence of the protein. Fractions containing T5 flap endonuclease were combined and diluted with $ddH_2O$ to reduce the conductivity to 15 mS/cm.

The protein sample was loaded onto an equilibrated 6 ml Resource Q (GE Health care) column at 5 ml/min. flow rate and the column was washed with 50 mM MES pH 6.5. The protein sample was then eluted with a 1 M NaCl gradient in 50 mM MES pH 6.5 and 2.5 ml fractions were collected. Protein fractions were again checked by the Bradford Assay (Bradford, 1976) and

SDS-PAGE analyses. Fractions that contained T5 flap endonuclease were collected and concentrated down to 1-2 ml by centrifugation at 4,000 x g in a 10 kDa molecular weight cut-off VivaSpin concentrator.

The sample, 2 ml, was applied to 1.6x60 cm Superdex G200 gel filtration column (GE Health care) after equilibration with 20 mM HEPES pH 7.5, 150 mM KCl, 0.1 mM EDTA, 1 mM DTT buffer, and 2 ml fractions were collected. Peak fractions containing T5 flap endonuclease were combined and concentrated to 15 mg/ml for crystallization. The sample was used for crystallisation trials and can be stored at 4°C (for up to a week) or at -80°C mixed with 50% glycerol (for long-term storage) until used.

#### *2.2.2.2 Human 5' Flap endonuclease (hFEN-1)*

The same procedure that was used for purification of the T5FEN was applied to the purification of the hFEN-D179K protein. About 90% pure protein was obtained from the first two steps of purification (heparin-Sepharose column followed by Resource Q column). Protein was concentrated to 11 mg/ml for crystallization.

#### *2.2.2.3 Trypanosoma brucei FEN-WT*

*Trypanosoma brucei* FEN was purified as described for hFEN-1 and 90% pure protein was obtained. Protein was stable in salt (NaCl) but it disappeared from the solution after salt removing. The protein was then concentrated to 18 mg/ml and 26.5 mg/ml for crystallization purposes.

## 2.3 Denaturing Gel Electrophoresis (SDS-PAGE)

The protein expression was confirmed by SDS-PAGE analysis according to the method described by (Laemmli, 1970) using the Bio-Rad Mini Protein 3 system according to the manufacturer's instructions. A 12% resolving gel was made by mixing 5 ml of 30% acrylamide (w/v acrylamide/bis solution, Bio-Rad Laboratories, Inc.), 4.69 ml 1 M Tris-HCl, pH 8.8, 2.56 ml Milli-Q water, 125 µl 10% SDS (w/v), 125 µl 10% ammonium persulphate and 12.5

µl TEMED. Once the resolving gel set, a 6% stacking gel was poured on top of it that consisted of 1.5 ml acrylamide (30% acrylamide/bis solution, Bio-Rad Laboratories, Inc.), 0.94 ml 1 M Tris-HCl pH 6.8. 2.6 ml Milli-Q water, 75 µl 10% SDS (w/v), 75 µl 10% ammonium persulphate and 7.5 µl TEMED. A perspex 1.0 mm comb (Bio-Rad) was immediately inserted.

The protein samples were supplied with 2 x loading buffer (100 mM Tris-HCl pH 6.8, 20% glycerol, 4% SDS (v/v), 0.02% (v/v) bromophenol blue) according to their volumes to reach 1x loading buffer as a final concentration. Then the samples were incubated at 95°C for 5 minutes and left to cool at room temperature. SDS-PAGE was run at 100 V initially then increased to 200 V once the protein samples had reached the resolving gel front. Running buffer was made up by mixing 14.4 g /l glycine, 3 g /l Tris-bicine, and 1 g SDS with a final pH 8.8.

The gel after that was stained with 0.1% Coomassie brilliant blue dissolved in 4:1:5 v/v ratios of methanol, acetic acid and sterile water and then shake for 15 minutes initially. The gel was then checked after 15 mins. which can give a quick idea for the purification process and was be left overnight. The same mixture lacking Coomassie blue was used as a destain solution to identify protein bands on the gel.

## 2.4 FEN:DNA Crystallisation

### 2.4.1 Preparation of Oligonucleotide

Several DNA oligonucleotides have been designed and ordered from Eurofins-MWG (Table 2.2) in order to co-crystallize with the T5FEN proteins: T5FEN-WT, T5FEN-D153K and T5FEN-D155K variants as well as with hFEN-D179K and *T. brucei*. The oligos were prepared using the same procedure and used in crystallization trials. The two types of DNA oligos that were successful in forming crystallization complexes with T5FEN and hFEN proteins are described in detail in the following sections.

Oligonucleotide DNA-substrates in the form of a 5' overhang structure were ordered from Eurofins-MWG Operon. It has the same sequence as the 5ov4 oligonucleotides that were designed and used by Anstey-Gilbert (Anstey-Gilbert et al, 2013; Hemsworth, 2009) in a complex with *E. coli* ExoIX (5'-AAAAGCGTACGC-3') (Fig. 2.1a). This oligo anneals to itself to make a DNA molecule that has an 8 base pair palindromic duplex with 4 base 5' overhangs.

Firstly, the 5ov4 oligo was used with the T5FEN-D153K variant and the concentration adjusted to 1.1 mM for the duplex molecule by dissolving it in 10 mM MES pH 6.5 and 50 mM KCl, and then it was annealed by heating to 94°C for 10 minutes and kept on the bench top at room temperature to cool before using in crystallization. Another DNA oligo was designed and ordered from Eurofins-MWG. This type of oligo was intended to have the sequence 5'-GATCTATATGCGATCGC-3' which should anneal to itself to give a DNA molecule composed of an 8 base pair palindromic duplex with a long 5' overhang of 9 bases long (Figure 2.1b). The concentration of the duplex molecule was adjusted to give 1.1 mM by dissolving the oligo in 10 mM MES pH 6.5 and 50 mM KCl oligo buffer. The oligos were self-annealed in oligo buffer by heating to 94°C for 10 minutes and left on the bench top at room temperature for 10 minutes to cool before using in crystallization.

### 2.4.2 T5FEN:DNA Co-Crystallization

For crystallization purposes appropriate amounts of T5FEN proteins (T5FEN-WT, T5FEN-D153K and T5FEN-D155K) were mixed with DNA oligos in order to get a final 1:1 molar ratio of protein:DNA. Crystallization trials were set down manually using vapour diffusion method (sitting drop and hanging drop) with different screens such as Nextal JCSG and PACT suite from QIAGEN and PEG/Ion II from Hampton research. Plates were incubated at 17°C and checked for any crystal growth after one and two weeks.

**Table 2.2:** The sequences and descriptions of DNA oligonucleotides. They were used in crystallization trials in complexes with T5FEN wild type, variant, hFEN1-D179K variant and *T. brucei* FEN.

| DNA Oligo | Sequence | Description |
|---|---|---|
| **5ov4** | 5'-AAAAGCGTACGC-3' | Anneal to itself to produce a DNA molecule consisting of an 8 base pair palindrome duplex with a 4, 7, 8 and 9 base 5' overhangs at each end. |
| **5ov7** | 5'-AAAAAAAGCGTACGC-3' | |
| **5ov8** | 5'-AAAAAAAAGCGTACGC-3' | |
| **5ov9** | 5'-AAAAAAAAAGCGTACGC-3' | |
| **5ov9** | 5'-GATCTATATGCCATCGG-3' | |
| **5ov9** | 5'-GATGTAGTTCCGTACGG-3' | |
| **5ov4** | 5'-AAAAGCGCGTACGCGC-3' | Anneal to itself to produce a DNA molecule consisting of a 12 base pair palindrome duplex with a 4 and 6 base 5' overhangs at each end. |
| **5ov6** | 5'-AAAAAAGCGCGTACGCGC-3' | |
| | 5'-AAAAGCGCGTACGCGCAAAA-3' | Anneal to itself to produce a double pseudo-Y DNA structure consisting of a 12 base pair with a 4 or 6 bases 5' and 3' pseudo-Y at each end. |
| | 5'-AAAAAAGCGCGTACGCGCAAAAAA-3' | |
| | 5'-ACTCTGCCTCAAGACGGT-3'<br>5'-ACTTTTGAGGCAGAGT-3'<br>5'-ACCGTCC-3' | Double flap DNA with 4 bases 5' flap and a 3'-single nucleotide. |

Optimizations of conditions that presented crystals growth were again set up manually in 24-well Linbro plates (Hampton Research) via hanging drop vapour diffusion technique by varying the concentrations of the precipitant and the overall pH.

### 2.4.3 hFEN-D179K:DNA Co-Crystallisation

The same procedure was applied for hFEN-D179K variant and DNA oligo but the plates were checked after a month. Optimization was done for the condition that a crystal was grown in and trials were again set up manually in 24-well Linbro plates (Hampton Recearsh) via hanging drop vapour diffusion technique by different concentrations of the precipitant.

### 2.4.4 *Trypanosoma brucei* FEN Crystallisation

#### *2.4.4.1 Wild Type Crystallisation*

Crystal trials by sitting drop method have been set up for the *T. brucei* FEN-WT full length using the robot and seven protein crystallisation screens (PACT, JSCG suite, PEG suite, PH clear, $AmSO_4$, Classic and MPD). The plates were incubated at 17°C room and checked for any crystal growth after one and two weeks but crystals start growing after a month. Optimisations of conditions that presented crystal growth were set up manually in 24-well Linbro plates (Hampton Research) via hanging drop vapour diffusion technique by varying the concentrations of the precipitant and the overall pH but no crystal growth was obtained in optimisations plates.

#### *2.4.4.2 Co-Crystallization with DNA Substrates*

*Trypanosoma brucei* FEN-WT full length was mixed with different DNA substrates in 1:1 molar ratio of protein:DNA. Crystallization trials were set using the robot and vapour diffusion method with the seven screens that were used in *apo* protein crystallization. The plates then were incubated in 17°C and checked after one week. The crystals were grown in different screen conditions.

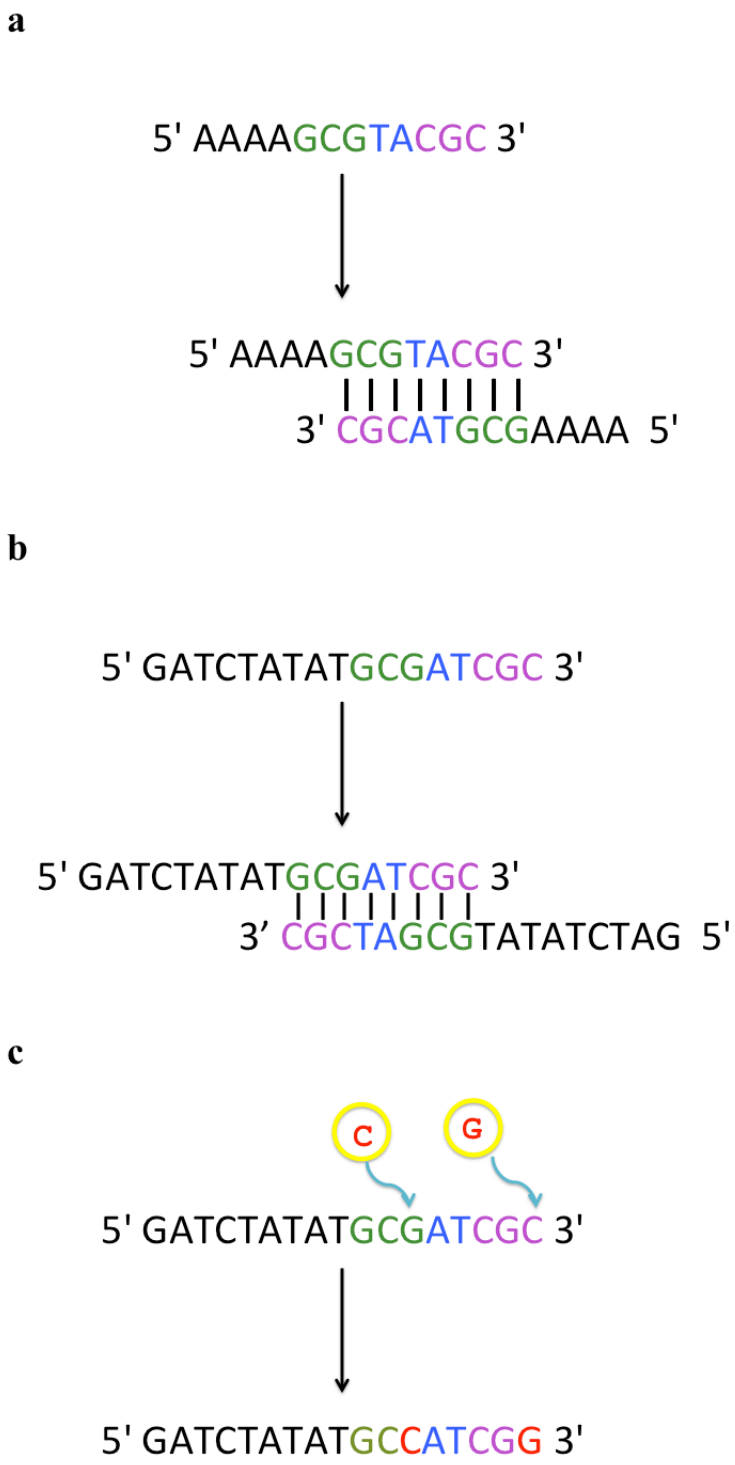



**Figure 2.21:** The DNA oligonucleotides used in this project.
**(a)** The palindromic sequence of the 5ov4 oligo. **(b)** The original palindromic sequence of the 3' overhangs oligo that was designed for crystallization as a complex with T5FEN proteins. **(c)** The modified sequence.

## 2.5 Cryoprotection and Data Collection

To start collecting data the crystals were tested in in-house using X-ray diffraction. Cryo buffers based on the screen conditions with different percentages of glycerol 5%, 10%, 15%, 20%, 25% and 30% were prepared. The frozen cryoprotectant solutions should first be tested by X-ray diffraction to avoid ice ring formation. The crystals were transferred to the best cryoprotectant solution for one second before flash cooling in a nitrogen stream at 100 K.

All crystals were initially tested for good diffraction on the in-house Rigaku Micromax 007 copper rotating anode generator that generates X-rays at a wavelength of 1.542 Å and which is equipped with mar345 detectors. Later higher resolution data were collected from good quality crystals at the Diamond Light Source (DLS) near Oxford, United Kingdom. The possible space group and data collection strategy were obtained for each crystal initially in-house using IMOSFLM software (Leslie, 1994) and later in Diamond using the Diamond autoindexing and autostrategy options of the data processing software, EDNA.

## 2.6 Data Processing

The processing of X-ray diffraction data was performed using MOSFLM in house (Leslie, 1994) and the xia2 and MOSFLM softwares at Diamond (Otwinowski & Minor, 1997). Both were carried out using three stepwise procedures: indexing, integration and scaling. During indexing the crystal lattice type, the unit cell parameters and the orientation parameters were determined. Scaling was then done using SCALA (CCP4, 1994) and free R Flag was set for later use in structure refinement.

## 2.7 Structure Determination and Refinement

### 2.7.1 The Phase Determination

The FEN proteins native and complexes with DNA substrates that have been studied during this project were solved by molecular replacement using the program PHASER (McCoy et al, 2007). This program runs a search in all alternative space groups to choose the one with the best solution which is hopefully correct.

### 2.7.2 The Search Model and Refinement

Before starting the molecular replacement process, search models should be chosen and modified if necessary and this modification include removing of solvent molecules, metal ions and any other molecules in protein complexes such as nucleic acids DNA or RNA. Chainsaw (CCP4, 1994) was used to modify the model pdb file (for T.b.FEN only) by pruning non conserved residues back to the gamma atom using the alignment between investigated protein and model sequences. For T5FEN:DNA complexes the PHASER program was used in the presence of the mtz file that was obtained from SCALA, and the PDB file for T5FEN native protein (ID-1EXN) that was solved by Ceska and his colleagues (Ceska et al, 1996) was used as a search model. Chain A from PDB file (3Q8L) and (3Q8K) that contains protein in a complex with DNA double flap substrate and product, divalent metal ion and $K^+$ ion (Tsutakawa et al, 2011) was used as a search model for hFEN-D179K:DNA structure and *T. brucei* FEN-WT respectively. These coordinates of the solution from PHASER were firstly run for 10 cycles of refinement with REFMAC5 program and then model building was performed in COOT (Emsley & Cowtan, 2004), with REFMAC5 (Murshudov et al, 1997) was used again for further structure refinement. The refinement process continues after each round of rebuilding until it reaches convergence when a best fit of the model to the map has been made and all features accounted for. At this point, the $R_{work}$, $R_{free}$ and log-likelihood values will also have stopped improving.

### 2.7.3 Model Validation

Finally, the MolProbity software (Chen et al, 2010) that can be downloaded for free from http://molprobity.biochem.duke.edu can perform a final evaluation of the structure and if any corrections are needed, they can be done in COOT before running the final refinement using REFMAC5. Additional structure validation can be obtained by uploading the pdb files into PROCHECK (Laskowski et al, 1993).

The structure determination process will be explained in more details for each structure in the following chapters.

PyMol (DeLano, 2002) is used to generate the figures in this thesis otherwise the source will be written under the figure.

# Chapter 3: Structure Determination

This chapter discusses the steps in the successful solution of the structures of T5FEN variant proteins and wild type all in complexes with DNA substrates. No previous crystallographic studies have been published for T5FEN in a complex with DNA substrates or products. Also no crystal structure has been obtained to clarify the helical arch's specific role in interactions with the 5'-ssDNA and to solve the puzzle of this region in FEN enzymes as discussed in section (1.2.1.3). During this project, T5FENΔ19, T5FENΔ19-D153K and T5FENΔ19-D155K crystal structures in 3D have been solved in complexes with DNA 5' and 3' overhang substrates. These structures were solved in a range of resolutions from 1.8 Å to 2.7 Å. The first T5FEN *apo* protein structure that was solved in 1996 by Ceska and his colleagues observed that the first 18 amino acids of the protein N-terminus were disordered (Ceska et al, 1996). These residues do not affect the enzyme activity but they formed a flexible part of the protein which does not appear in the electron density map (Ceska et al, 1996; Feng et al, 2004). Consequently, this region was deleted from the subsequent protein constructs and this truncated form, T5FENΔ19, will be referred to as T5FEN wild type in the next sections.

T5FEN-D153K and T5FEN-D155K variants were designed by the Sayers lab in order to investigate the possible new metal binding sites and to prevent the cations from binding to the enzyme active site in Cat1 and Cat2 respectively by replacing the negative charge residues (aspartic acid 153 and 155) with positive ones (lysines). These mutations reduce or abolish the enzyme activity but do not affect the protein's ability to form complexes with DNA substrates or products (Feng, 2002; Zhang, 2012). In this case, 5' flap or overhang DNA cannot be cleaved by the enzyme and can interact with the helical arch by threading or clamping methods allowing investigators to study the arch role in the DNA interactions and so they are valuable to investigate complexes with DNA.

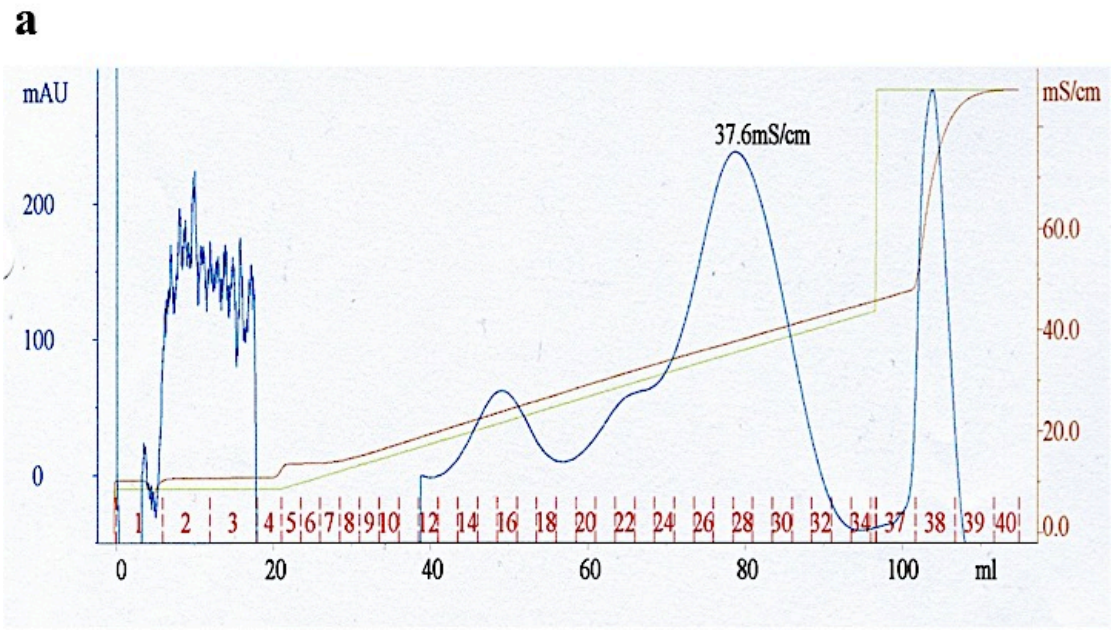
## 3.1 T5FEN Crystal Preparation

### 3.1.1 Purification

#### *3.1.1.1 T5FEN-D153K*

*E. coli* M72 strains carrying pJONEX44 with the protein coding sequence for T5 5' flap endonuclease (FEN) wild-type or variant were kindly provided by Prof. Sayers lab, Medical School, University of Sheffield. T5FEN and T5FEN-D153K proteins were expressed and purified using the protocols that were described in sections (2.2.1.1 and 2.2.2.1). The purification process (details below) of T5FEN enzymes was carried out using a heparin column followed by an anion exchange chromatography (Resource Q column) and a final size exclusion chromatography step was performed with a Superdex G200 column. The protein's purity was more than 95% which is good for crystallization trials. After that, protein was concentrated to 15 mg/ml and stored at 4°C to mix with DNA oligo and use for crystallization.

Pellets of approximately 10 g cells were lysed in lysis buffer (50 mM Tris-HCl pH 8.0 and 0.5 M NaCl) and sonicated. 50 ml of cell free extract was collected resulting in ~ 990 mg total protein. The sample was diluted four times with 50 mM Tris-HCl pH 8.0 and then applied on a 5 ml Heparin-HP cartridge. The protein then was eluted by a 75 ml gradient of NaCl concentration from 0.15 M to 0.6 M in 50 mM Tris-HCl buffer. 2.5 ml fractions were collected from the first step of purification and peak fractions 25-31 were combined (Figure 3.1) with total protein of 17.5 mg. Samples were then diluted four times with distilled water until the conductivity was 15 mS and then applied on a 6 ml Resource Q column. The protein was eluted by a 48 ml gradient of NaCl concentrations from 0.15 M to 0.3 M in buffer B (50 mM MES pH 6.5, Figure 3.2a). About 2.5 ml fractions were collected and fractions 22 and 23 were combined. Total protein was 8.2 mg which was measured by the Bradford method (Bradford, 1976).



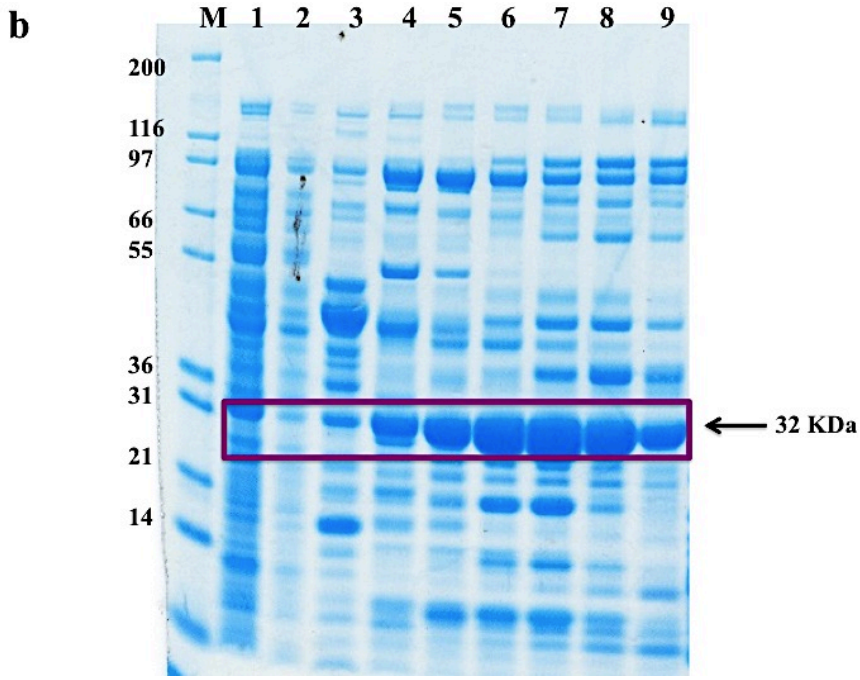




**Figure 3.22:** Initial purification step for T5FEN-D153K variant using Heparin column.
**(a)** The chromatogram of the elution profile from Heparin-HP column.
**(b)** 12% SDS-PAGE analysis of the protein samples from Heparin-HP column and the protein can be seen at the estimated size ~32 kDa. M lane: Protein molecular weight marker (Mark 12$^{TM}$), lane 1: pellet, lane 2: crude extract, lane 3: unbound protein and lanes 4-9: fractions.

The sample was concentrated down to 1.7 ml and 8.84 mg protein using a Vivaspin concentrator and then applied on a 16 x 60 superdex 200 column equilibrated with HEPES-KCl buffer. The gel filtration process was performed at a flow rate 1.5 ml / min. Two ml fractions were collected and 13-15 peak fractions were combined (Figure 3.2b) and concentrated down to 0.35 ml with 5.1 g pure protein for crystallisation (Figure 3.2c).

#### *3.1.1.2 T5FEN-D155K*

This protein was supplied by Jing Zhang, Medical School, University of Sheffield, as a pure T5FEN-D155K protein at a concentration of 48 mg/ml kept in 50% glycerol. The protein was then processed to remove the glycerol and to obtain a 15 mg/ml sample as a final concentration as described in section (2.2.1.1).

#### *3.1.1.3 T5FEN- WT*

The same protocol was used to purify T5FEN-WT through three steps: heparin column, Resource Q column and gel filtration column. Pellets were prepared from 3 liter cultures and lysed in lysis buffer (50 mM Tris-HCl and 0.5 M NaCl). Cell free extract (CFE) containing 520 mg of total protein were collected in 40 ml and diluted 3-fold with 50 mM Tris-HCl pH 8.0 before applying on a 5 ml Heparin-HP cartridge. The protein was eluted by a 75 ml gradient of NaCl concentration from 0.15 M to 0.6 M in 50 mM Tris-HCl buffer and 2.5 ml fractions were collected. A chromatogram of the T5FEN-WT elution from the heparin column can be seen in Figure 3.3a. Ten ml sample from peak fractions (20-23) containing 30 mg protein were combined and diluted to 3 times their original volume with 50 mM Tris-HCl pH 8.0 buffer. The sample was applied to a 6 ml Resource Q column and then protein was eluted by 60 ml gradient of NaCl concentration from 0.1 M to 0.3 M in 0.5 M NaCl and 2.5 ml fractions were collected. Fractions from 21 to 26 were collected and protein concentration was checked by the Bradford method (Bradford, 1976). Peak fractions 23 and 24 with 16 mg total protein were combined and then concentrated to 2 ml using a Vivaspin

concentrator. The protein sample was applied to a 16 x 60 Superdex 200 column equilibrated in HEPES-KCl buffer and gel filtration was performed at a flow rate of 1.5 ml / min (Figure 3.3b). Two ml fractions were collected and fractions 13-15 were combined to 5.7 ml with 8.5 mg total protein (Figure 3.4a). A Vivaspin device was used to concentrate the protein to ~15 mg/ml for crystallization purposes. The three steps of purification to obtain the pure protein were investigated using SDS-PAGE analysis and can be seen in Figure 3.4b.

## 3.1.2 Crystallization

### *3.1.2.1 DNA 5ov4 and 3over6 Substrates Preparation*

To make a complex between T5FEN-D153K, D155K and WT enzymes and a DNA substrate 5ov4: (5′-AAAAGCGTACGC-3′) and 3ov6: (5'-GATCTATATGCCATCGG-3') the oligo must be prepared before use in crystallization to obtain 1.1 mM concentration of duplex in buffer 10 mM MES pH 6.5 and 50 mM KCl. More details about the DNA 5ov4 and 3ov6 substrates can be found in section (2.4.1).

### *3.1.2.2 Crystallization Process*

Wild type and variant versions of the T5FEN enzyme at 15 mg/ml were mixed with 1.1 mM 5ov4 and 3ov6 oligos at 1:1 molar ratio. Consequently, 25 µl of oligo were mixed with 50 µl of protein. Crystallisation trials were set down manually using the sitting drop vapor diffusion method with different screens such as Nextal JCSG, PACT, and PEG/Ion II. After two weeks crystal trials were checked under the microscope and some crystals were found. The T5FEN-D153K variant did not crystallize with 3ov6 oligo but good quality crystals were grown with 5ov4 DNA in different conditions of Nextal JCSG screen as shown in (Figure 3.5). On the other hand, the T5FEN-D155K variant and T5FEN-WT did not produce any crystals with the 5ov4 oligo but they formed crystals in the presence of the 3ov6 oligo in some conditions of Nextal JCSG and PACT screens (Figure 3.6).

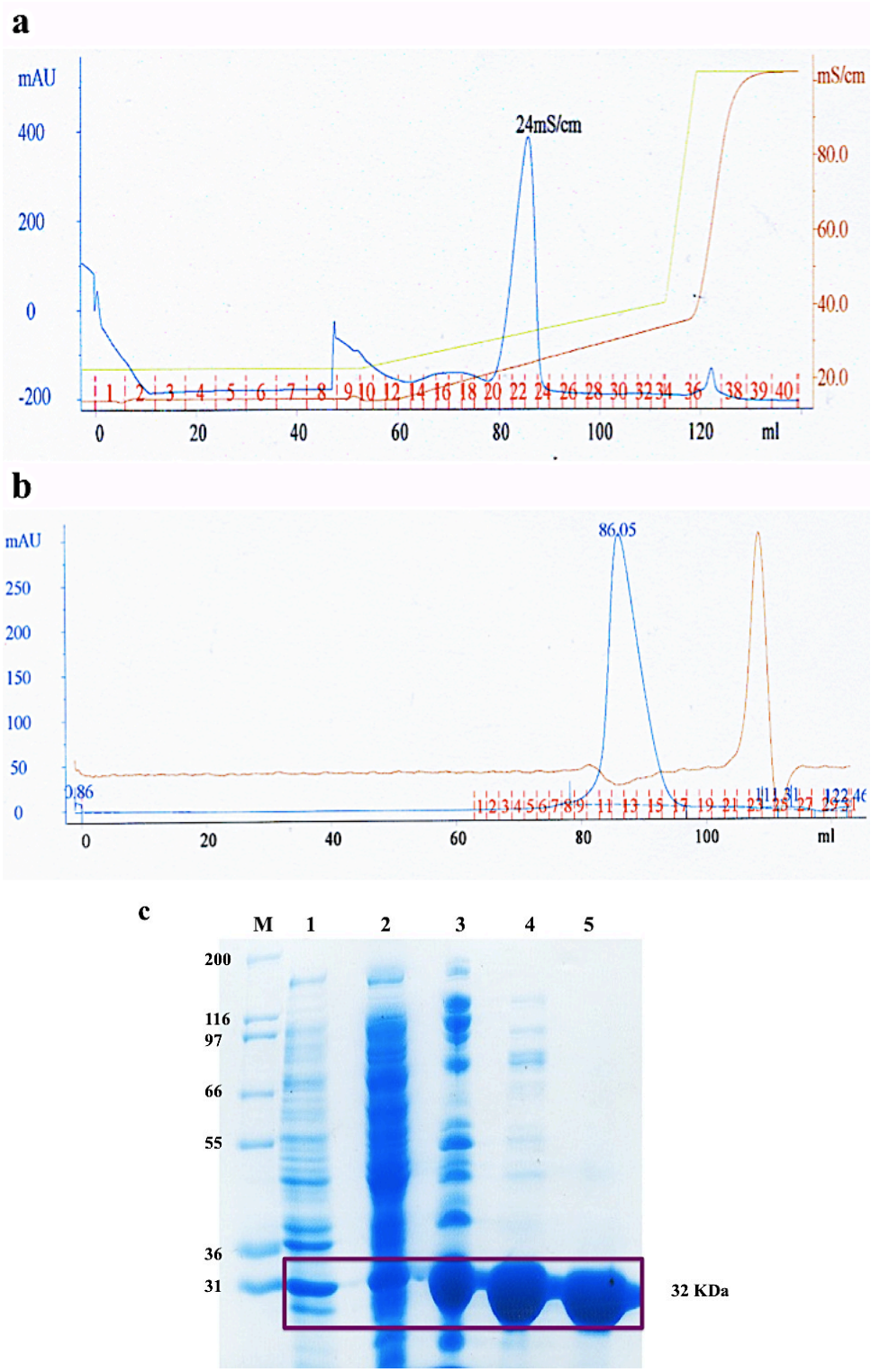



**Figure 3.23:** Chromatogram of the elution profile for T5FEN-D153K from Resource Q column and 16 x 60 Superdex 200 column.
The peak fractions can be seen in 22 and 23 in **(a)** and the chromatogram of the elution profile from a 16 x 60 Superdex 200 column in **(b)**. **(c)** Shows 9% SDS-PAGE analysis of T5FEN-D153K protein from gel filtration column and the protein can be seen in the estimated size ~32 kDa. M lane: Protein molecular weight marker (Mark 12$^{TM}$), lane 1: pellet, lane 2: crude extract, lane 3: unbound protein and lanes 4 and 5: fractions.

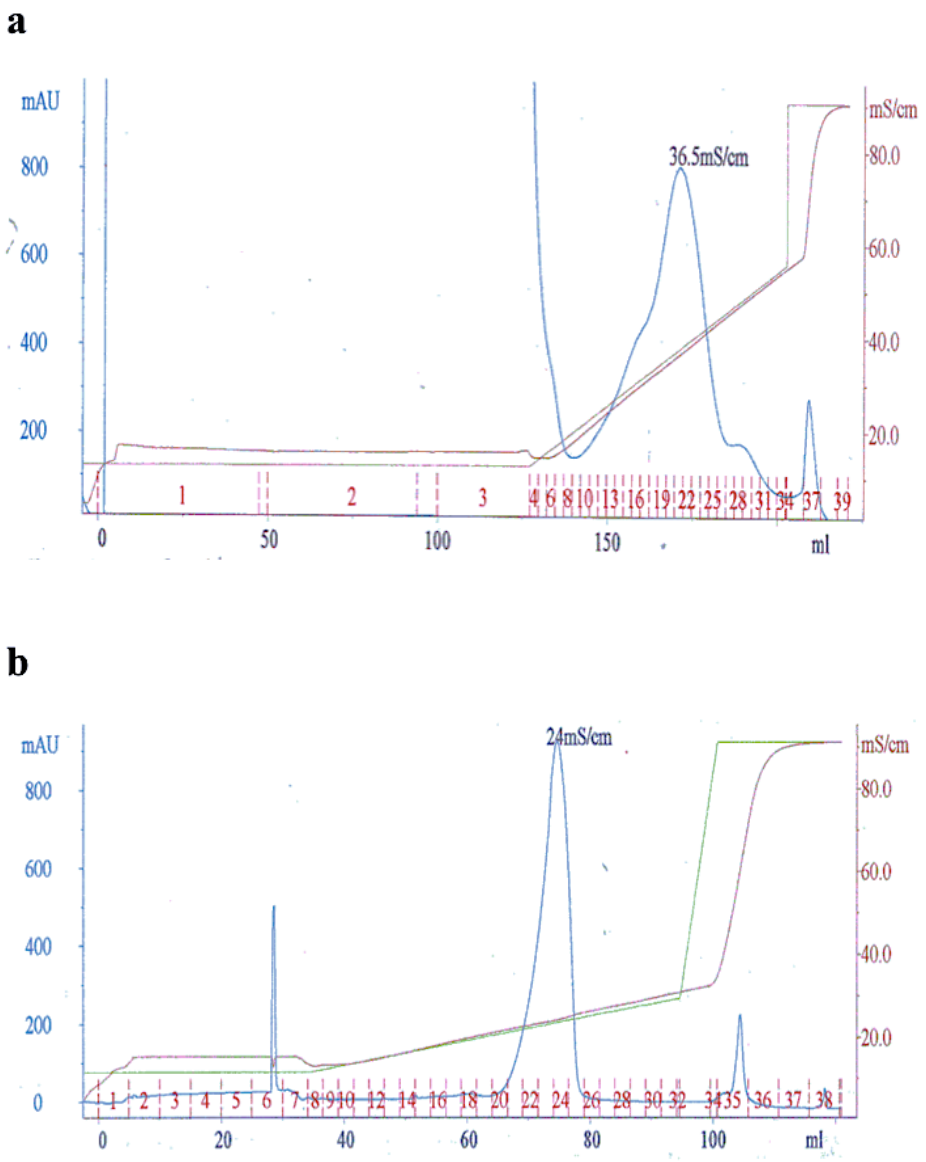



**Figure 3.24:** Chromatogram of the elution profile for T5FEN-WT.
**(a)** From Heparin-HP column and the peak fractions can be observed in 20-23. **(b)** From Resource Q column and the peak fractions can be seen in 21-26.

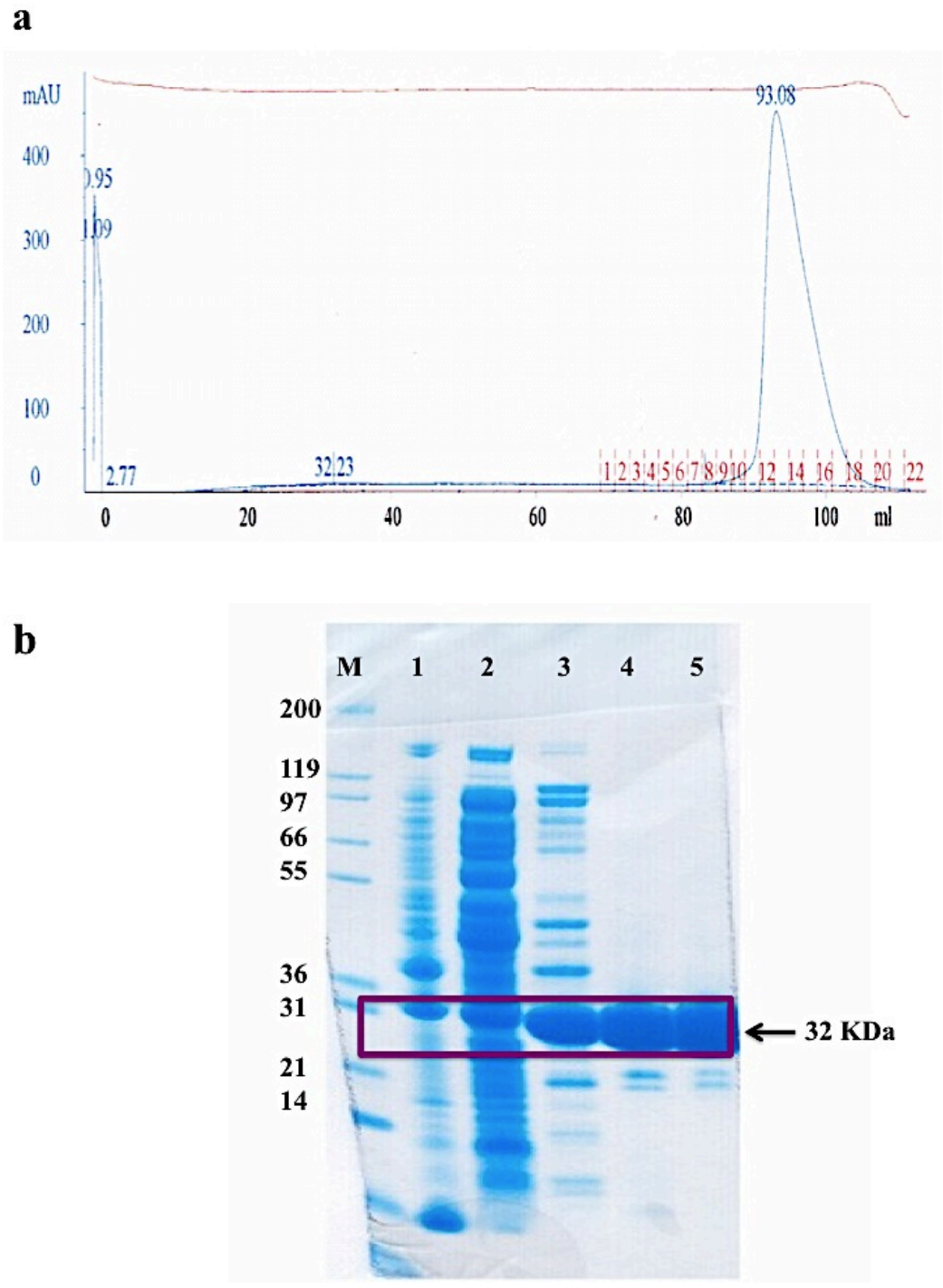
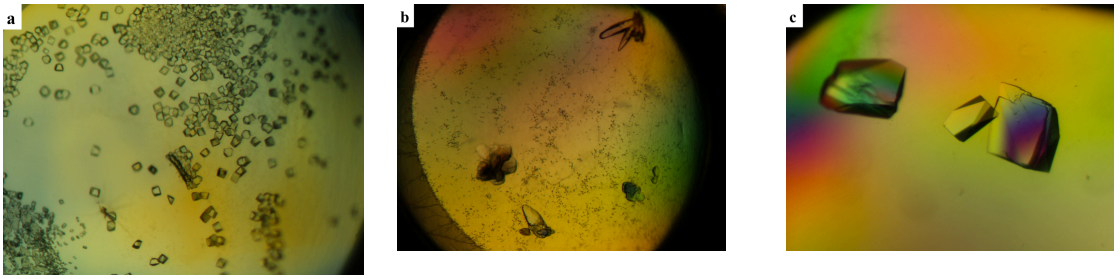
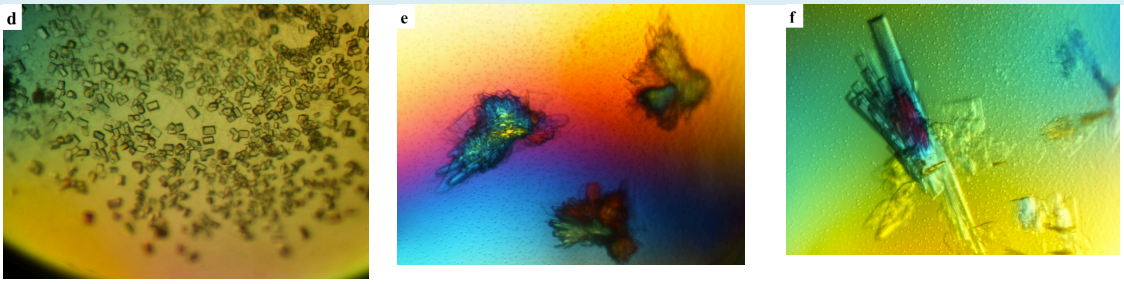


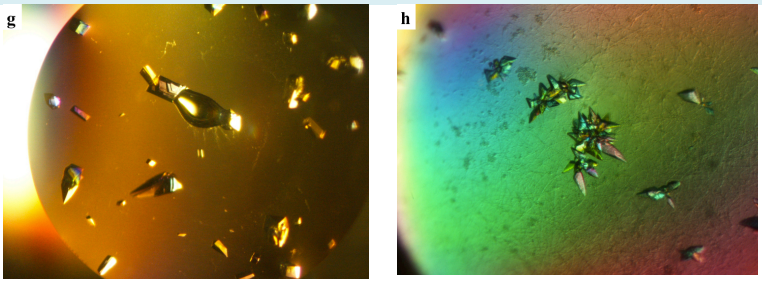

**Figure 3.25:** Shows the last step of purification process using 16 x 60 superdex-200 gel filtration column and SDS-PAGE analysis for T5FEN-WT.
**(a)** Chromatogram of the elution profile from a 16 x 60 superdex-200 gel filtration column. The peak fractions can be seen from 13-15. **(b)** SDS-PAGE using 4-12% Bis-Tris NOVAX gel. Protein purity improvement can be observed through three purification steps used to purify T5FEN-WT. M: protein molecular weight (Mark 12$^{TM}$), lane 1: pellets, lane 2: cell free extract, lane 3: after Heparin-HP column, lane 4: after Resource Q column and lane 5: after gel filtration column.




| 0.8 M $(NH4)_2SO_4$, 0.1 M $C_6H_8O_7$ pH 4 | 0.2 M $(NH4)_2SO_4$, 0.1 M bis-Tris pH 5.5, 25% PEG 3350 | 0.2 M $Li_2SO_4$, 0.1 M bis-Tris pH 5.5, 25% PEG 3350 |
|---|---|---|

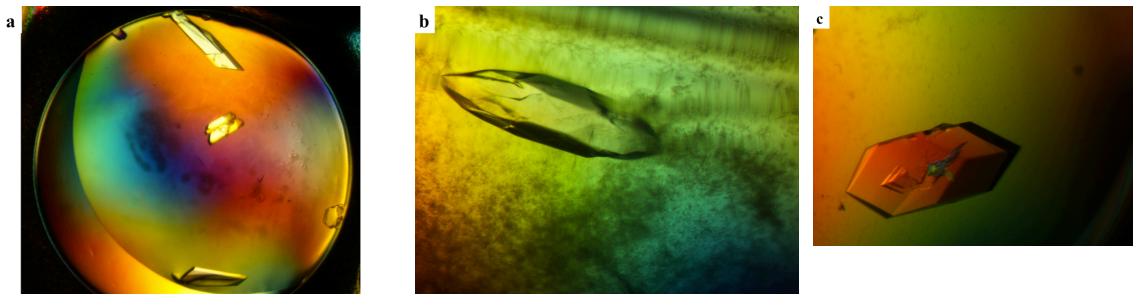



| 0.2 M $MgCl_2$, 0.1 M bis-Tris pH 5.5, 25% PEG 3350 | 0.2 M NaCl, 0.1 M $(CH3)_2AsO_2Na$ pH 6.5, 2 M $(NH4)_2SO_4$ |
|---|---|

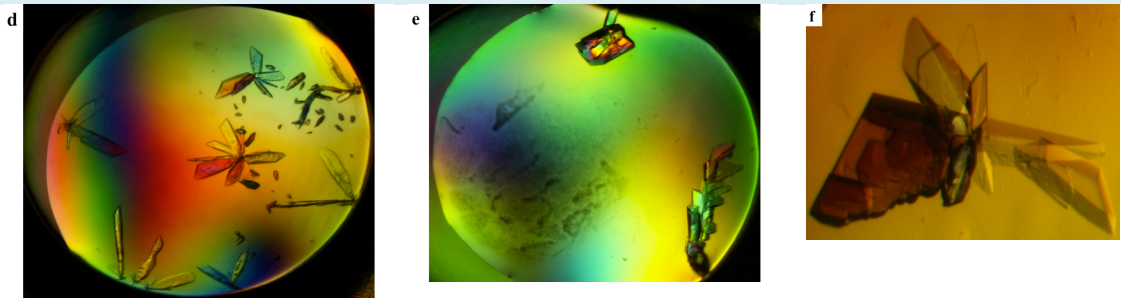



| 0.2 M $MgCl_2$, 0.1 M bis-Tris pH 5.5, 22% PEG 3350 | 0.2 M Ca(OAc)2, 0.1 M $(CH3)_2AsO_2Na$ pH 6.5, 46% PEG 300 |
|---|---|

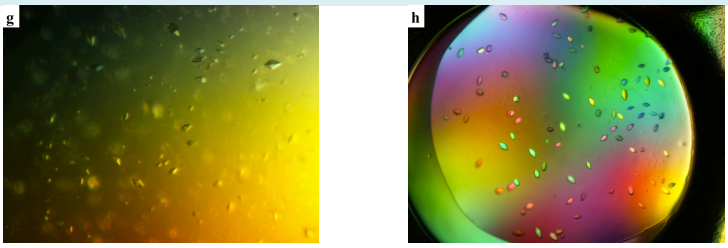
**Figure 3.26:** Different crystals for T5FEN-D153K variant complexed with 5ov4 DNA substrate were grown in different conditions of Nextal JCSG screen.




| 0.1 M Na2HPO4 pH 4.2, 40% $C_2H_6O$, 5% PEG 1000 | 0.8 M Na3PO4, 0.8 M K3PO4, 0.1 M HEPS pH 7.5 | 0.2 M LiSO4, 0.1 M Tris pH 8.5, 1.26 M $(NH4)_2SO4$ |
|---|---|---|




| 0. 2 M $KNaC_4H_4O_6.4H_2O$, 20% PEG 3350 | 0.2 M HCOONa, 0.1 M Bis-Tris propane pH 6.5, 20% PEG 3350 | 0.2 M CaCl2, 0.1 M C2H3NaO2 pH 5, 20% PEG 6000 |
|---|---|---|




| 0.2 M $NH_4C_2H_3O_2$, 0.1 M Bis-Tris pH 5.5, 24% PEG 3350 | 0.1 M MIB buffer pH 9, 25% PEG 1500 (WT) |
|---|---|

**Figure 3.27:** Different crystals for T5FEN-D155K variant and T5FEN-WT.
The crystals were grown in the presence of 3ov6 substrate in different conditions of Nextal JCSG and PACT screens.

In Figure 3.5 G and H crystals were grown in optimization buffers. The original condition for the G crystals was 0.2 M $MgCl_2$, 0.1 M bis-Tris pH 5.5, 25% PEG 3350. The precipitant (PEG 3350) concentration was increased from 20% to 30% and the new crystals grew in 22% PEG 3350. The screen condition for the H crystals was 0.2 M $Ca(OAc)_2$, 0.1 M $(CH3)_2AsO_2Na$ pH 6.5, 40% PEG 300. The concentration of the precipitant solution (PEG 300) was varied from 36% to 46% and the new crystals grew in 46% PEG 300. Crystals for wild type and variant protein took about two weeks to grow at 17°C. These crystals were tested on the in-house apparatus to choose the best crystals that gave a good diffraction pattern and these were sent to the Diamond Light Source (DLS) for collecting data.

Two crystals for each T5FEN variants, D153K and D155K, and one for the wild type have been chosen to present in this project and referred to them as C1 and C2 for T5FEN-D153K:5ov4 complexes and C3 and C4 for T5FEN-D155K:3ov6 complexes while CW will be used for the wild type complex.

## 3.2. T5FEN-D153K Complex with 5ov4

### 3.2.1 Data Collection

To collect data, crystals were mounted in loops and dipped in cryoprotectant buffer which is composed of the mother liquor solution and one of the anti-freezing buffers and frozen in liquid nitrogen. Cryoprotectant was prepared in different percentages of glycerol starting from 5-30%, frozen in liquid nitrogen and then tested in the in-house system to choose the best one (see appendix 2.4.3). The first test for cryoprotectant solution confirmed that 5% glycerol could be used for collecting data from crystals but in general from 5%-30% glycerol was used.

For C1 (Figure 3.7a) data were collected on Beamline I03 at the Diamond synchrotron to 2.16 Å resolution (Figure 3.7b). The set of 450 images was collected at 0.2° rotation increments with 2 seconds exposure per image at wavelength 0.976253 Å and 20% transmission.

For C2 (Figure 3.7c) data were collected on Beamline I-24 at the Diamond Light Source and the crystal diffracted to 2.2 Å (Figure 3.7d). The 0.977023 Å wavelength X-rays were used by exposing the crystal for 0.20 seconds at 26% transmission of the beam with 0.20˚ rotation of the crystal per image. About 900 images were collected for this crystal.

### 3.2.2 Data Processing

X-ray diffraction data for T5FEN:DNA crystals were processed in the same way using Xia2 (Winter, 2010) at Diamond. It was carried out using three stepwise procedures: indexing, integration and scaling. During indexing the crystal lattice type, the unit cell parameters and the orientation parameters were determined. Then the diffraction spot intensities were integrated in the appropriate lattice type. Scaling and merging data then were done using SCALA (CCP4, 1994) and .mtz file was generated for each crystal. The free R set was selected for later use in structure refinement.

The spacegroup for C1 was determined to be $P2_12_12_1$ with an orthorhombic lattice type. It has a parameters of a = 63.99, b = 105.98, c = 106.85 Å and $\alpha = \beta = \gamma = 90°$. Orthorhombic unit cell was also determined for C2 with $P2_12_12_1$ spacegroup again. The lattice dimensions for the second crystal are a = 44.72, b = 109.94, c = 127.34 Å and $\alpha = \beta = \gamma = 90°$. Data collections and processing statistics for these protein crystals are summarized in Table 3.1.

### 3.2.3 Model Building

#### *3.2.3.1 Matthews Probability Calculation*

To start molecular replacement it was important to estimate the number of molecules in the asymmetric unit (AU) of the crystal. For this purpose the Mathews number (Vm) was calculated for each crystal (Matthews, 1968) using the web page http://www.ruppweb.org.Mattprob/ (Kantardjieff & Rupp, 2003).

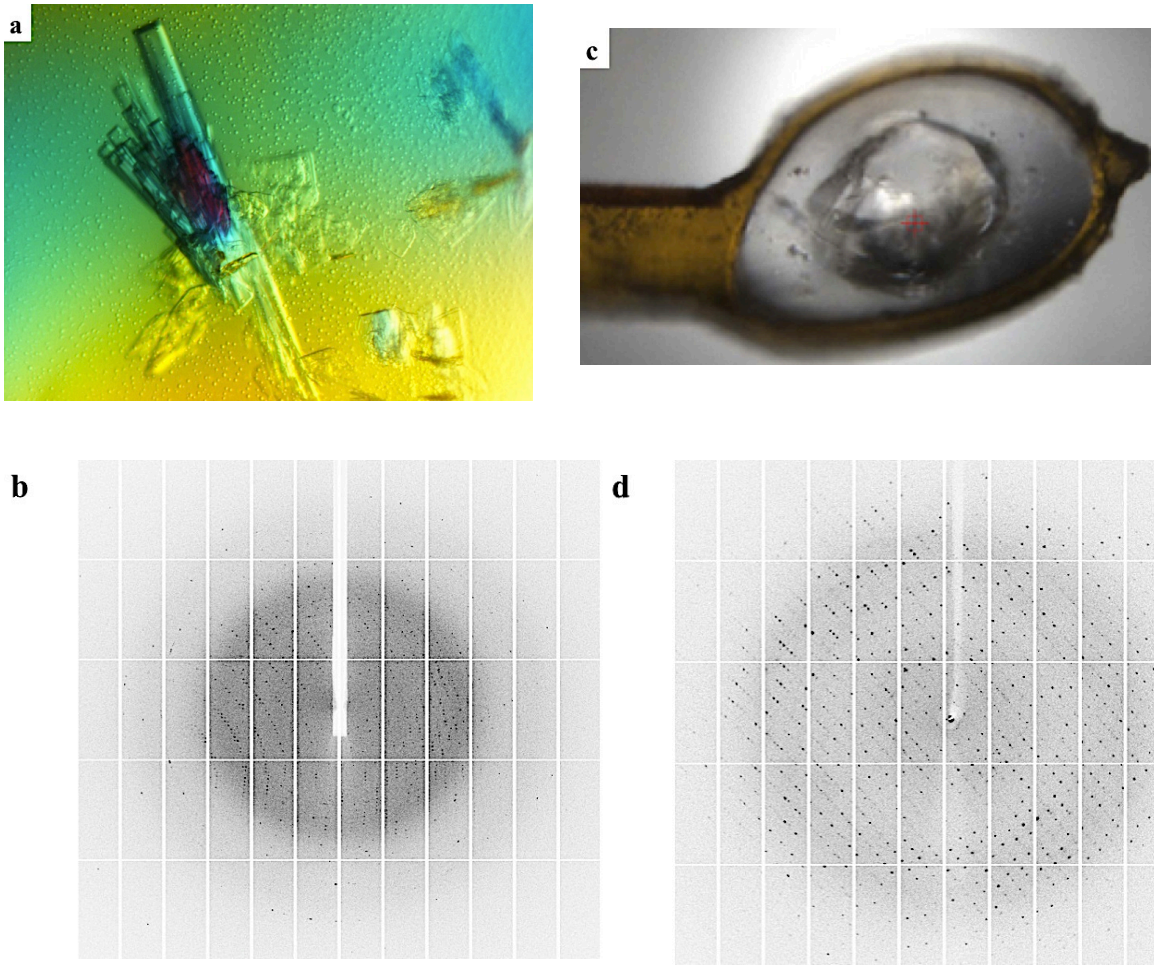



**Figure 3.28:** C1 and C2 of T5FEN-D153K in complex with 5ov4 DNA
**(a)** and **(c)** Crystals of C1 and C2 respectively.
**(b)** and **(d)** The diffraction pattern of the crystals (2.16 Å max. resolution) for C1 and (2.22 Å max. resolution) for C2 respectively.

**Table 3.3:** Data collection and processing statistics for C1 and C2 of T5FEN-D153K:DNA.

| **Data collection and processing statistics** | | **T5FEN-D153K** | | | |
|---|---|---|---|---|---|
| | | **C1** | | **C2** | |
| X-ray source | | 103 | | I-24 | |
| Wavelength | | 0.97630 Å | | 0.977023 Å | |
| Space group | | $P2_12_12_1$ | | $P2_12_12_1$ | |
| Crystal system | | Orthorhombic | | Orthorhombic | |
| Cell dimension | | | | | |
| a (Å) | $\alpha$ (°) | 63.99 | 90° | 44.72 | 90° |
| b (Å) | $\beta$ (°) | 105.98 | 90° | 109.94 | 90° |
| c (Å) | $\gamma$ (°) | 106.85 | 90° | 127.34 | 90° |
| Resolution range overall (Å) | | 47.71-2.16 | | 44.72-2.22 | |
| Resolution range highest shell (Å) | | 2.21-2.16 | | 2.28-2.22 | |
| Number of observations | | 126849 (9585) | | 182387 (6726) | |
| Number of unique reflections | | 39459 (2934) | | 31727 (2177) | |
| $R_{merge}$ | | 0.140 (1.093) | | 0.094 (0.792) | |
| $R_{pim}$ | | 0.076 (0.594) | | 0.038 (0.405) | |
| I/$\sigma$I | | 10.4 (1.9) | | 13.4 (2.2) | |
| Completeness | | 99.2 (99.7) | | 99.4 (94.8) | |
| Multiplicity | | 3.2 (3.3) | | 5.7 (3.1) | |
| Matthews coefficient (Å$^3$/Da) | | 2.83 | | 2.44 | |
| Solvent content% | | 56.49% | | 49.67% | |
| Number of molecules in the AU. | | 2 | | 2 | |

*Data in brackets indicates highest resolution shell.

$$R_{merge} = \sum_{hkl}\sum_{i}\left|Ii(hkl) - \langle I(hkl)\rangle\right| / \sum_{hkl}\sum_{i} Ii(hkl)$$

$$R_{pim} = \sum_{hkl}\left[\frac{1}{N-1}\right]^{1/2}\sum_{i}\left|Ii - \langle I\rangle\right| / \sum_{hkl}\sum_{i} Ii$$

Where <*I(hkl)*> is the mean intensity of the reflection.

These probabilities based on the unit cell volume for each crystal. There were four possible solutions in the normal Vm range of C1 crystal. There were 1, 2, 3 or 4 molecules in the asymmetric unit. For the resolution of the crystal's data, the most likely was two molecules in the asymmetric unit with 56.49% solvent content and a Vm of 2.83 $Å^3$/Da. The C2 of this variant had three possible solutions in the normal Vm range. 1, 2 or 3 molecules were in the asymmetric unit and the strongest possible solution was two molecules in the asymmetric unit with a solvent content of 49,67% and a Vm of 2.44 $Å^3$/Da.

### *3.2.3.2 Structure Solution and Refinement*

#### 3.2.3.2.1 The Search Model

After determination of the number of molecules in the asymmetric unit of each crystal, the structures of T5FEN-D153K variant in complexes with 5ov4 DNA substrates were solved using the molecular replacement method. First of all, an appropriate search model was chosen which belongs to T5FEN native crystal structure (PDB accession code 1EXN) (Ceska et al, 1996). The model was prepared and modified by removing of solvent molecules and chain B of the 1EXN coordinate was used.

#### 3.2.3.2.2 The Phase Determination

The PHASER software (McCoy et al, 2007) was used to solve the structure using the chain B of the 1EXN coordinate and the SCALA output free.mtz file for each structure of T5FEN-D153K:DNA complex. The sequence identity between the model and the observed structure was set to 100% and the search looked for two copies (about 600 amino acids) in the asymmetric unit. PHASER found a single solution for C1 and C2 structures with two protein chains in a spacegroup of $P2_12_12_1$. Phasing rotation function Z score (RFZ), translation function Z score (TFZ), log-likelihood gain (LLG) and the total log-likelihood gain (LLG) for the first and the second chains in C1 and C2 structures are summarized in Table 3.2.

**Table 3.4:** The PHASER statistics for C1 and C2 of T5FEN-D153K:DNA

| T5FEN-D153K:DNA | First chain | | | | Second chain | | | Total LLG |
|---|---|---|---|---|---|---|---|---|
| | RFZ | TFZ | LLG | | RFZ | TFZ | LLG | |
| **C1** | 14.8 | 31.7 | 711 | | 13.1 | 56.7 | 2824 | 3176 |
| **C2** | 9.6 | 18.2 | 285 | | 10.7 | 31.8 | 1144 | 1344 |

### 3.2.3.2.3 Model Building and Refinement

The coordinates generated by PHASER were refined using REFMAC5 (Murshudov et al, 1997) and ten cycles to refinement were run initially. The coordinates output from the refinement process were used to rebuild the structure using the COOT program (Emsley et al, 2010). The protein chains were built by adding amino acids main or side chains with clear density and removing the others with no or poor density. Each rebuilding cycle in COOT was followed by refinement in REFMAC5 until the protein chains were completed.

### 3.2.3.2.4 Identification of the DNA

In both C1 and C2 structures there was evidence of density for DNA between the two protein chains and a model of the DNA was built using COOT by adding adenosine nucleotides that were subsequently changed during the refinement. The building of the DNA in both structures was started in the duplex region of the molecule close to the H3TH motif in the protein structure and continued in the 5' overhang direction. The nucleotide sequence and numbering were corrected to agree with our own sequence, which agreed well with the density, and the structure refined further using REFMAC5. In the beginning of the model building process for C1 structure, the 5' overhang has two adenosines at the 5' end and the electron density for the other two adenosines did not reappear again (Figure 3.8a).

Consequently, the observed 5' overhang was composed of only two dAs while the other two nucleotides may be disordered or possibly have been removed by a contaminating nuclease activity. In contrast, the 5' overhang in C2 contained only two nucleotides but as the density was improved after rebuilding and refinement cycles it allowed the addition of the other two adenosines to give a 5' overhang with four nucleotides (Figure 3.8b). The DNA 5' overhang can be seen totally threaded through the arch in C2 structure while partial threading was observed in C1 crystal structure according to the electron density maps (Figure 3.8).

#### 3.2.3.2.5 Identification of the Metal Ions

There was no any evidences for metal ion presence in C1 crystal structure active sites. A very weak peak was noticed in the chain B potassium binding-site and was occupied with the $K^+$ ion and refined the structure by REFMAC5. The output files were opened in COOT and the bond lengths agreed with this metal ion while the B-factor was high which could refer to the weakness of the peak. In C2/A there was no evidence for the presence of the metal ions in the active site but there was a peak in the $K^+$ binding site. This peak was identified as a $K^+$ ion, which was added, and the structure was refined again in REFMAC5. The output files were opened in COOT and the bond lengths and B-factor were confirmed the identification of this metal ion. C2/B had two peaks of density in the active site, one of them close to Asp-26 and Asp-68 residues while the other one close to Asp-201 and Asp-204 and both of these peaks interact with the DNA backbone. Upon checking the crystallization buffer used, it has 200 mM $MgCl_2$ in its component which can occupy these two peaks. The $Mg^{2+}$ ions were added and the REFMAC5 was run. Additionally, in the potassium-binding site within the H3TH motif there were also two peaks for the metal ions. One of them was occupied by the $K^+$ ion and the other one was occupied firstly by $Mg^{2+}$, $Ca^{2+}$ ions and water molecule before run REFMAC5. The pdb and mtz* files were opened in COOT and the bond lengths and B-factors agreed with the $K^+$ ion occupied the first peak but disagreement with any one of the other ions or water molecules in the second peak were observed.

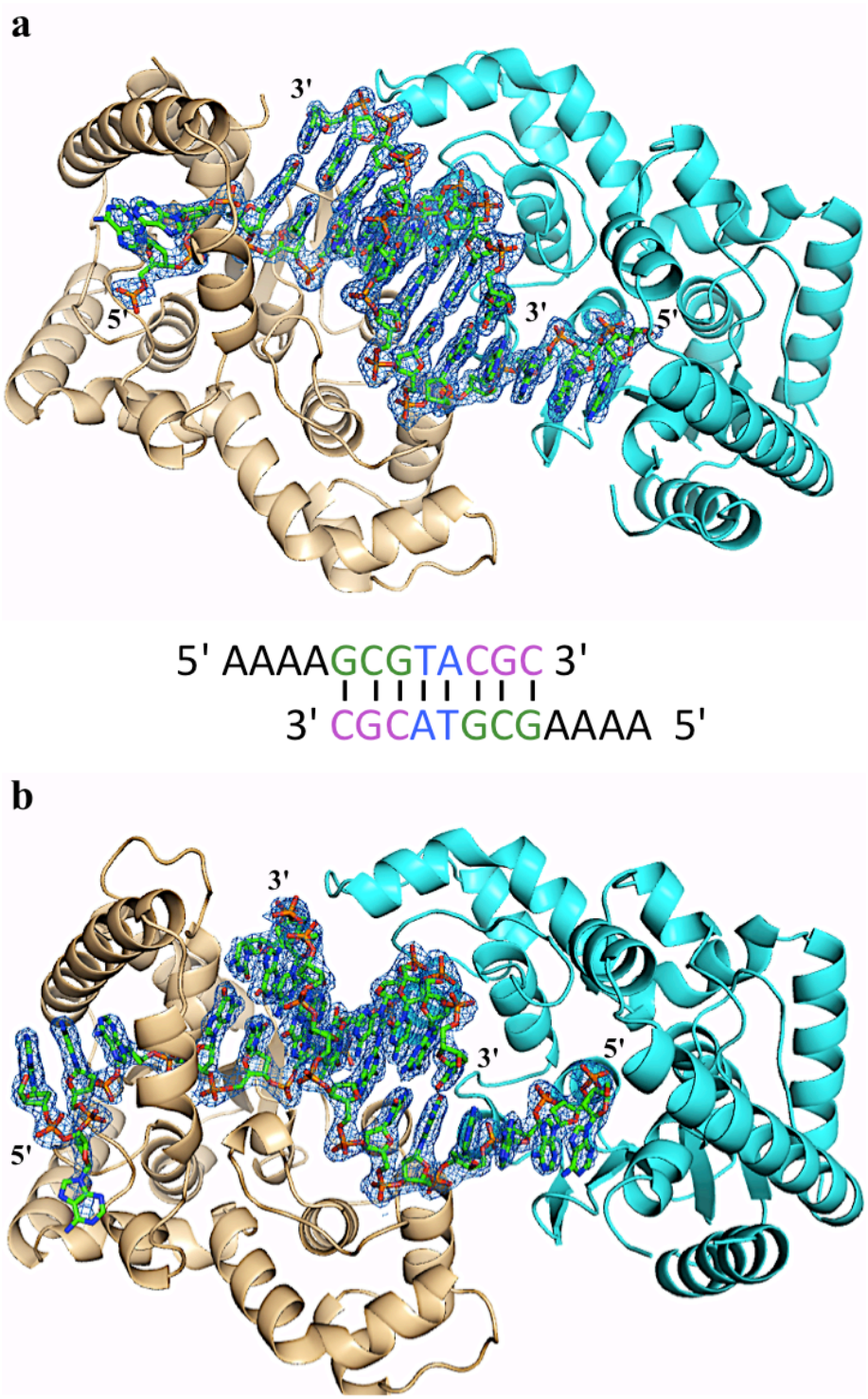



**Figure 3.29:** 2Fo-Fc electron density map for the 5ov4 DNA in C1 and C2 of T5FEN-D153K:DNA. Each structure, C1 and C2, has two protein chains, A (cyan) and B (wheat) and the DNA substrate can be seen between these chains.
**(a)** Shows the partially threading by two nucleotides in C1 while **(b)** shows fully threading of four nucleotides in C2. The 5ov4 palindromic DNA oligo sequence can be seen between C1 and C2 structures.

This peak site then was occupied with another $K^+$ ion with 60% occupancy and 40% for the previous one in the first peak and the REFMAC5 was run again. The COOT was used to check the result from REFMAC5 and it was more acceptable for the second peak to be occupied with $K^+$ ion according to the B-factors which means the $K^+$ ion in this structure has two positions within the H3TH motif.

**3.2.3.2.6 Addition of Solvent**

The other solvents that were present in the crystallization and cryoprotectant buffers such as the glycerol and the polyethylene glycol were added when they agreed with the electron density map. Water molecules were added automatically using the "Find waters" option in COOT which were then checked manually. The water molecules that were located greater than 3.5 Å from surrounding atoms or have no density were removed. The structures were checked for any unmodelled density that might be PEG or glycerol molecules, or other solvents.

In C1 structure there were electron density present in similar places around the two molecules, A and B. These electron density took triangular shapes and by checking the crystallization buffer [**0**.2 M NaCl, 0.1 M (CH3)2AsO2Na pH 6.5, 2 M (NH4)2SO4] it has 2 M of ammonium sulphate solution. This component in solution give two ions: NH4+ and $(SO4)^{2-}$ and the electron density were agree with the sulphate ions more than ammonium ions or any other ions present in crystallization buffer. The $(SO4)^{2-}$ ions were added and the structure refined. By checking the structure in COOT the B-factors were in agreement with sulphate ions that appeared in thirteen copies around the complex.

After rebuilding everything and adding the DNA and the solvent molecules the refinement converged with $R_{work}$ and $R_{free}$ achieved reasonable values of (21% / and 28%) for C1 and (19% / and 27%) for C2 structures. The two complexes were uploaded into MolProbity and Procheck to estimate the

structure quality. Residues with clashes or poor rotamers were corrected in COOT before running the final refinement with REFMAC5.

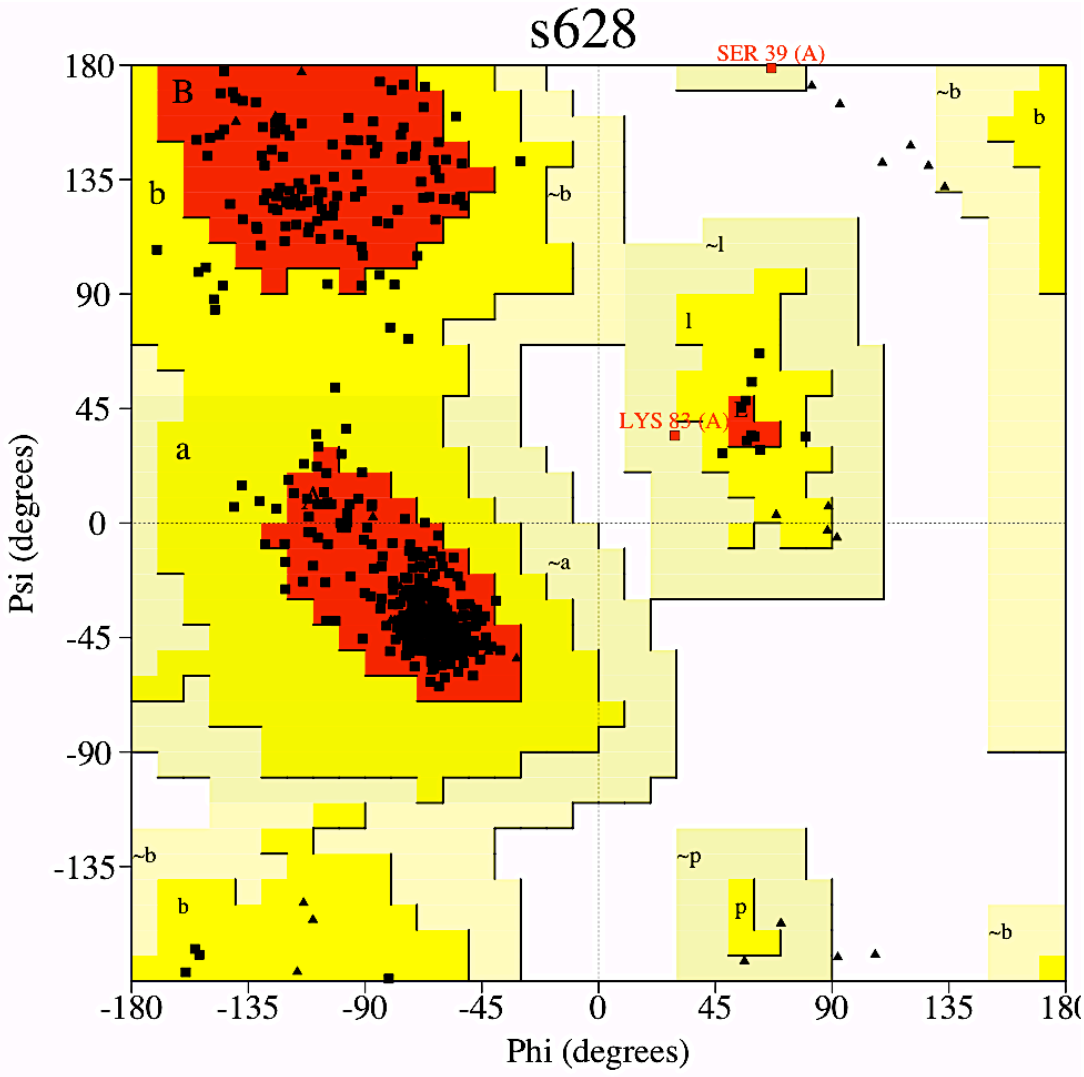
#### *3.2.3.3 Structures Stereochemistry*

Both models of C1 and C2 were input into PROCHECK (Laskowski et al, 1993) and the Ramachandran plots are shown in Figure 3.9 for C1 and Figure 3.10 for C2 structures. According to Ramachandran plots it can be seen that no residues in the disallowed regions for C1 while one was present for C2, and only two residues were found in the generously allowed regions for C1 and one for C2. Also these structures input into the MolProbity software (Davis et al, 2007) to provide additional structure validation through clash score, rotamer and Ramachandran plots. MolProbity detected the possible clashes between all structure atoms and adjustment the side-chain amide rotamers to satisfy potential hydrogen bonding interactions. It also defines the allowed regions on the Ramachandran plot using the updated datapase of structures.

For C1 MolProbity gave 3.62 clashscore corresponding to 99$^{th}$ percentile and 0.38% Ramachandran outliers for two residues out of 571 across the two molecules (Figure 3.9). 2.50% of side-chain conformations marked as rotamer outliers and 98$^{th}$ percentile MolProbity score. The C2 structure reported higher clashscore of 4.82 corresponding to 98$^{th}$ percentile with two Ramachandran outliers was recorded and 2.07% of side-chains exhibited poor rotamers (Figure 3.10) while the overall MolProbity score came for 96$^{th}$ percentile. These results for the two structures gave strong evidences that the final models of T5FEN-D153K complexed with 5' overhang DNA are of a good quality and the final refinement statistics are written in Table 3.3.

**Table 3.5:** Data refinement statistics for C1 and C2 crystals.

| Data Refinement Statistics | T5FEN-D153K:DNA | |
|---|---|---|
| | **C1** | **C2** |
| Number of reflections | 37381 | 30073 |
| $R_{work}/R_{free}$ | 0.21/0.28 | 0.1969/0.2730 |
| No. Atoms (protein/other) | 4257/754 | 4255/764 |
| No. Waters/ Average B-factors | 197/32.415 | 258/42.936 |
| B-factors of bond atoma ($Å^2$) | | |
| Overall | 32.606 | 38.673 |
| Main chain | 30.116 | 37.098 |
| Side chain | 34.481 | 39.870 |
| DNA (X & Y) | 39.133/44.554 | 42.507/37.801 |
| R.m.s deviation from ideality | | |
| Bond lengths (Å) | 0.0190 | 0.0191 |
| Bond angles (°) | 1.8266 | 1.9018 |
| Ramachandran plot. Proportion in residues in: | | |
| Allowed regions% | 92.0 | 9.0 |
| Additional allowed regions% | 7.6 | 9.2 |
| Generously allowed regions% | 0.4 | 0.2 |
| Disallowed regions% | 0.0 | 0.2 |
| Average RMS B-factors (main/side chains) | 0.688/1.998 | 0.790/2.424 |
| Average RMS B-factors of DNA (X & Y) | 4.646/4.004 | 4.646/4.504 |




Plot statistics

| | | |
|---|---|---|
| Residues in most favoured regions [A,B,L] | 446 | 92.0% |
| Residues in additional allowed regions [a,b,l,p] | 37 | 7.6% |
| Residues in generously allowed regions [~a,~b,~l,~p] | 2 | 0.4% |
| Residues in disallowed regions | 0 | 0.0% |
| Number of non-glycine and non-proline residues | 485 | 100.0% |
| Number of end-residues (excl. Gly and Pro) | 51 | |
| Number of glycine residues (shown as triangles) | 34 | |
| Number of proline residues | 12 | |
| Total number of residues | 582 | |

**Figure 3.30:** Diagram for the final model of C1-T5FEN-D153K:5ov4 shows Ramachandran plot.
The figure was produced using PROCHECK http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/

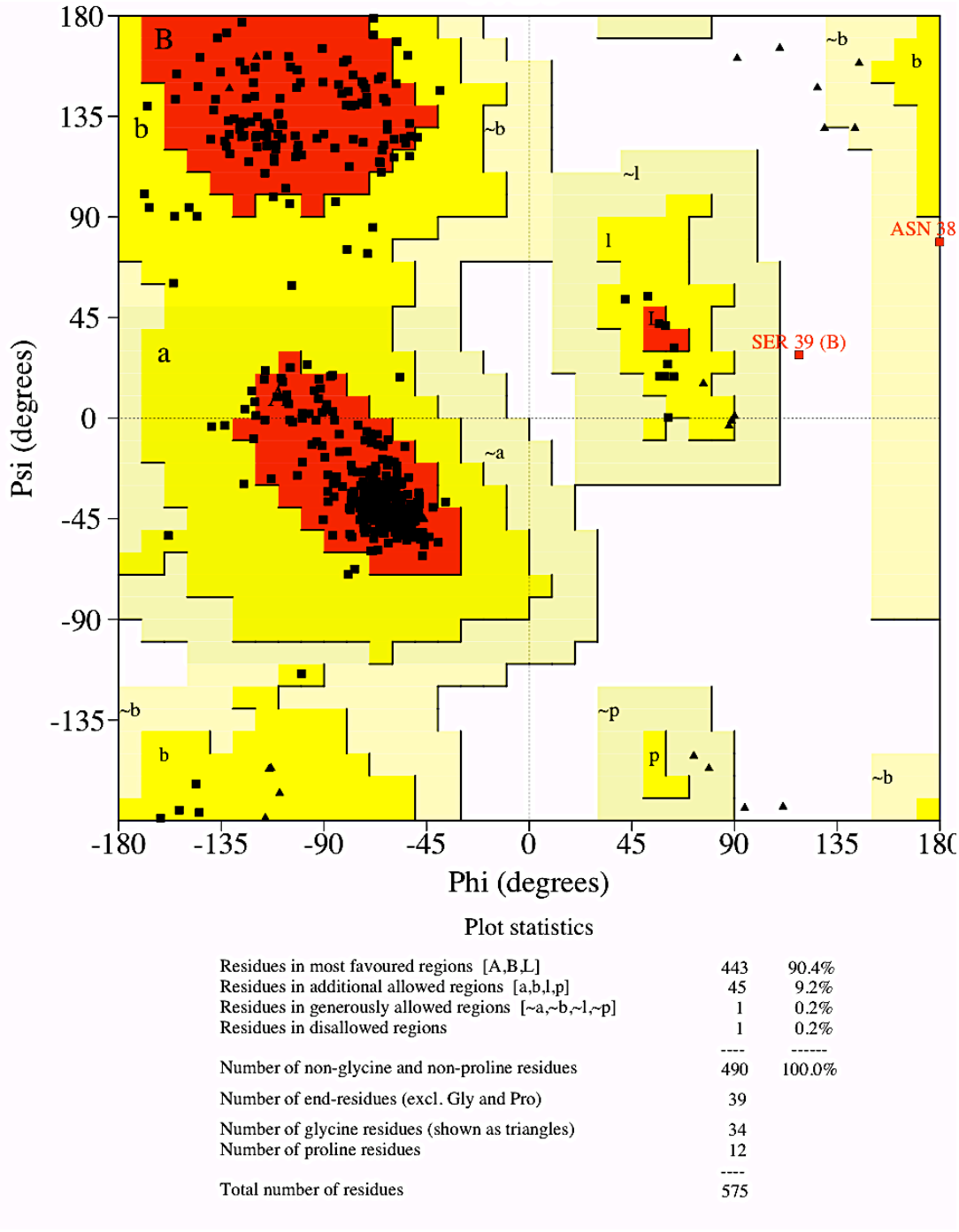



**Figure 3.31:** Ramachandran plot for C2 final model.
Figures produced by PROCHECK. http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/

## 3.3 T5FEN-D155K and WT Complexes with 3ov6

### 3.3.1 Data Collection

The T5FEN-D155K variant and the WT were crystalized in complexes with the same DNA substrate, 3ov6, and they solved with similar process. T5FEN-D155K:DNA gave good quality crystals and most of them diffracted to a high resolution. Unfortunately, two crystals only have the protein complexed with 3' overhang DNA substrate and they are diffracted to 1.8 Å and 2.2 Å while T5FEN-WT:DNA crystal is diffracted to 2.9 Å. For collecting data the crystals were cryoprotected by adding 30% (v/v) glycerol to the crystallization buffer that crystals were grew in.

Data were collected for C3 structure (Figure 3.11a) in the presence of 30% of glycerol mixed with 0.2 M $CaCl_2$, 0.1 M $NaO_2C_2H_3$ pH 5 and 20% (w/v) PEG 6000 precipitant solution. A data set of 360 images was collected on Beamline I02 at Diamond using 0.9795 Å wavelength and the crystal was rotated in 0.1° with 0.5 second exposure per image and 100% transmission. This crystal resulted in the highest resolution, 1.8 Å, (Figure 3.11d) in all the T5FEN structures solved in complexes with DNA oligos.

The C4 crystal (Figure 3.11b) was also cryoprotected with 30% glycerol mixed with 0.2 M $NH_4C_2H_3O_2$, 0.1 M Bis-Tris pH 5.5, 24% PEG 3350 mother liquor solution. Data were collected on Beamline I03 at the Diamond. 0.9200 Å wavelength X-rays were used, exposing the crystal for 0.4 second at 100% transmission of the beam with 0.2° rotation of the protein crystal per image. About 500 images were collected at 2.2 Å resolution (Figure 3.11e). To collect data for T5FEN-WT:DNA complex the crystal (Figure 3.11c) was also cryoprotected by adding 30% glycerol to crystallization buffer containing 0.1 M MIB pH 9 and 25% (w/v) PEG 1500. The crystal was sent to the Diamond Light Source, Oxfordshire, and diffracted to 2.9 Å (Figure 3.11d). A data set was collected consisting of 360 images in 0.1° rotations with 0.5 second exposure per image using 0.9795 Å wavelength and 100% transmission.

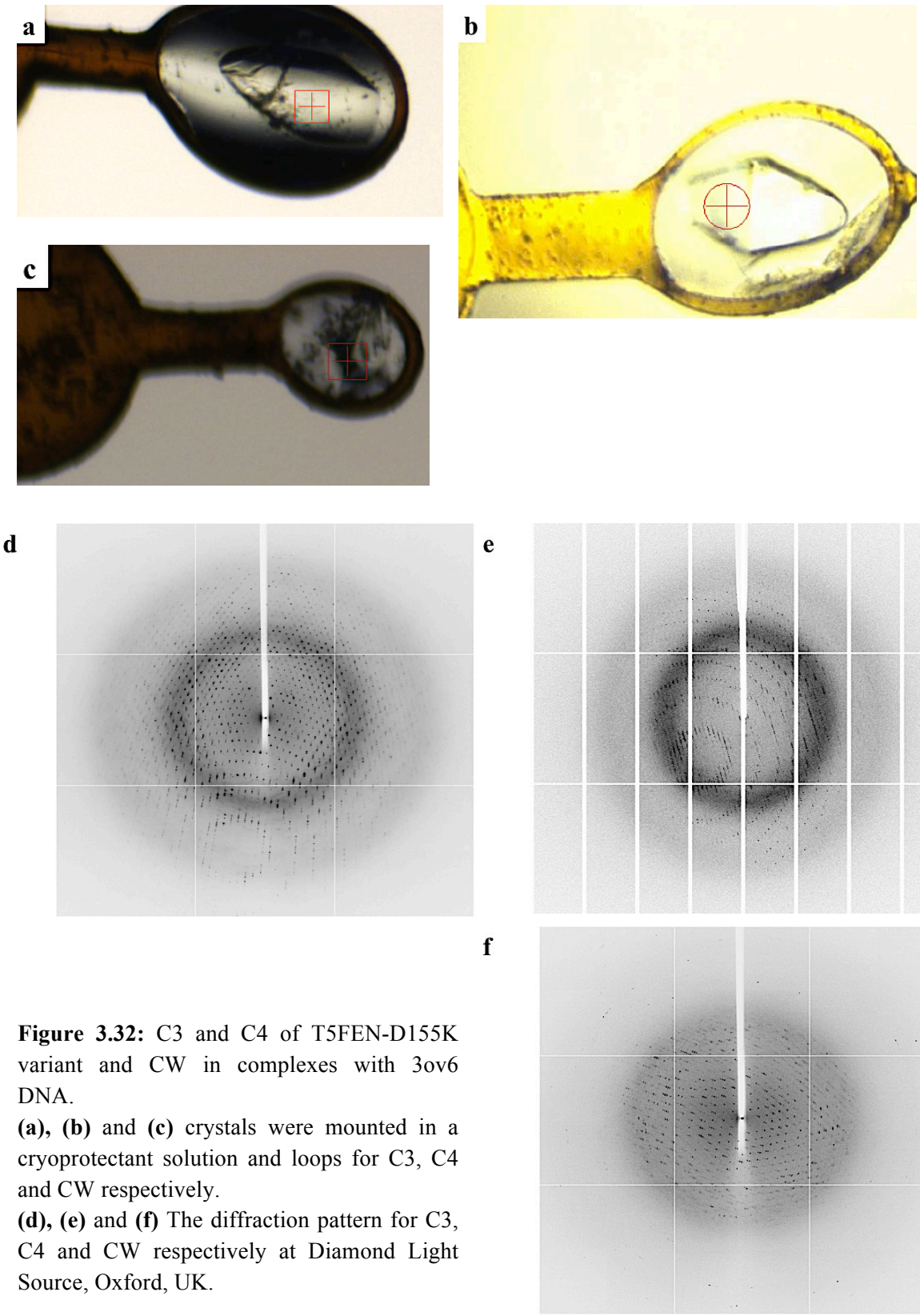



**Figure 3.32:** C3 and C4 of T5FEN-D155K variant and CW in complexes with 3ov6 DNA.
**(a), (b)** and **(c)** crystals were mounted in a cryoprotectant solution and loops for C3, C4 and CW respectively.
**(d), (e)** and **(f)** The diffraction pattern for C3, C4 and CW respectively at Diamond Light Source, Oxford, UK.

### 3.3.2 Data Processing

The crystals of T5FEN-D155K variant in complexes with 3' overhang DNA were processed in the same way as C1 and C2 crystals using xia2 (Winter, 2010). They are determined in tetragonal $P4_12_12$ space group and C3 has the dimension of a = b = 67.51, c = 187.81 Å and $\alpha = \beta = \gamma = 90°$. C4 recorded the parameter of a = b = 67.19, c = 186.96 and $\alpha = \beta = \gamma = 90°$. The processing of T5FEN wild type crystal was indicated that it is also in $P4_12_12$ space group with cell parameters of a = b = 67.44, c = 188.22 Å and $\alpha = \beta = \gamma = 90°$. This space group is new for T5FEN structures and the data statistics for these crystals are summarized in Table 3.4.

### 3.3.3 Model Building

#### *3.3.3.1 Matthews Probability Calculation*

C3, C4 and CW crystals shared similar unit cell parameters. The Matthews number calculation for all of them indicated that there were two prediction solutions for the molecules number in the asymmetric unit. These solutions were one or two molecules in the asymmetric unit. The most probable solution was the first prediction with one molecule in the asymmetric unit, 63% solvent content and 3.34 $Å^3$/Da Vm for C3, C4 and CW.

#### *3.3.3.2 Structure Solution and Refinement*

##### 3.3.3.2.1 The Search Model

A T5FEN-D153K variant 3D crystal structure in a complex with magnesium ions was solved by Flemming at 1.5 Å resolution (Flemming, 2011). Chain A of this structure's coordinates was use as a search model to obtain the structures of C3, C4 and CW using molecular replacement method. The search model was prepared before used in the same way that was described above (see section 3.2.3.2.1).

**Table 3.6:** Data collection and processing statistics for C3, C4 and CW of T5FEN-D155K:DNA.

| **Data collection and processing statistics** | | **T5FEN-D155K, WT** | | | | | |
|---|---|---|---|---|---|---|---|
| | | C3 | | C4 | | CW | |
| X-ray source | | 102 | | 103 | | 102 | |
| Wavelength | | 0.97960Å | | 0.9200 Å | | 0.97960 Å | |
| Space group | | $P4_12_12$ | | $P4_12_12$ | | $P4_12_12$ | |
| Crystal system | | Tetragonal | | Tetragonal | | Tetragonal | |
| Cell dimension | | | | | | | |
| a (Å) | $\alpha$ (°) | 67.51 | 90° | 67.19 | 90° | 67.44 | 90° |
| b (Å) | $\beta$ (°) | 67.51 | 90° | 67.19 | 90° | 67.44 | 90° |
| c (Å) | $\gamma$ (°) | 187.81 | 90° | 186.96 | 90° | 188.22 | 90° |
| Resolution range overall (Å) | | 67.51-1.86 | | 26.06-2.20 | | 62.74-2.90 | |
| Resolution range highest shell (Å) | | 1.91-1.86 | | 2.26-2.20 | | 2.98-2.90 | |
| Number of observations | | 1047650 (78323) | | 137590 (10592) | | 177628 (8335) | |
| Number of unique reflections | | 37488 (2720) | | 22372 (1642) | | 10012 (649) | |
| $R_{merge}$ | | 0.083 (1.073) | | 0.098 (0.695) | | 0.145 (0.925) | |
| $R_{pim}$ | | 0.016 (0.198) | | 0.039 (0.269) | | 0.033 (0.235) | |
| I/$\sigma$I | | 30.8 (6.4) | | 10.9 (2.6) | | 16.9 (2.5) | |
| Completeness | | 99.8 (99.7) | | 99.3 (100.0) | | 97.3 (89.4) | |
| Multiplicity | | 27.9 (28.8) | | 6.2 (6.5) | | 17.7 (12.8) | |
| Matthews coefficient ($Å^3$/Da) | | 3.34 | | 3.30 | | 3.34 | |
| Solvent content% | | 63.21% | | 62.68% | | 63.21% | |
| Number of molecules in the AU. | | 1 | | 1 | | 1 | |

*Data in brackets indicates highest resolution shell.

$$R_{merge} = \sum_{hkl}\sum_{i}\left|Ii(hkl) - \langle I(hkl)\rangle\right| / \sum_{hkl}\sum_{i} Ii(hkl)$$

$$R_{pim} = \sum_{hkl}\left[\frac{1}{N-1}\right]^{1/2}\sum_{i}\left|Ii - \langle I\rangle\right| / \sum_{hkl}\sum_{i} Ii$$

Where <*I(hkl)*> is the mean intensity of the reflection.

#### 3.3.3.2.2 The Phase Determination

The PHASER software (McCoy et al, 2007) was set again for each structure to search for one subunit in the asymmetric unit (300 residues) using the SCALA output free-mtz files and a resolution range from 54.8 Å to 2.5 Å for C3 and C4 and from 46.2 Å to 2.9 Å for the CW structure. The PHASER was run with all choices of alternative spacegroup for the mtz file predicted spacegroup ($P4_12_12$) and 100% similarity with the search model. A single solution was obtained from PHASER with spacegroup of $P4_32_12$ for all complexes. PHASER statistics for these three structures are summarized in Table 3.5. The mtz spacegroup was set for all complexes to the one that was obtained by PHASER as a single solution using sortmtz option in CCP4 (CCP4, 1994).

**Table 3.7:** PHASER statistics for C3, C4 and CW

| Complex | RFZ | TFZ | LLG | Total LLG |
|---|---|---|---|---|
| **C3** | 13.2 | 39.0 | 1251 | 1648 |
| **C4** | 12.7 | 36.0 | 1393 | 1570 |
| **CW** | 8.1 | 29.6 | 892 | 892 |

#### 3.3.3.2.3 The Model Building and Refinement

The output files from PHASER were refined using REFMAC5 (Murshudov et al, 1997) and ten cycles refinement were run initially. The coordinates that were generated by REFMAC5 were opened in COOT (Emsley et al, 2010) and manually building for the protein residues was started followed by refinement process for all T5FEN:3ov6 structures.

#### 3.3.3.2.4 Identification of the DNA

In the electron density map that was opened in COOT there was an extra strong density not related to the protein in each structure of C3, C4 and CW (Figure 3.12). The DNA backbone and bases can be easily recognized in these density maps which gave evidence for the presence of the DNA molecules. The building of the DNA molecule was started in C3 from the duplex region next to the H3TH motif of the protein in the 5' prime

direction. In some parts of the map there were evidences that this DNA is 3' prime and not 5' prime but because it was designed as a 5' overhang (see section 2.4.1) the building of the DNA was continued in the same way (5'). When the DNA building process was finished it was clear from the electron density this molecule is crystallized as 3' overhang and not 5' overhang. The DNA molecule was deleted and rebuilt again but this time with 3' direction and refined in REFMAC5. By the end of this process the shape of the DNA molecule was clear and composed of ten base pairs and a six base 3' overhang symmetrically by two-fold axis. What was happened with this DNA substrate can be explained as this: the DNA was designed to have the sequence 5'GATCTATATGCGATCGC3' (Figure 3.13a) and was ordered as 5'GATCTATATGCCATCGG3' (Figure 3.13b). Additionally, the sequence 5'ATCTATATC3' can be easily annealed to itself and as a result 3' overhang was generated and prevent producing of the 5' overhang which have not noticed in this DNA substrate (see section 2.4.1). In addition, the nucleotide number ten, guadinosine-10, was clamped by the arch residues as seen in Figure 3.13c. The DNA substrate was then built in other structures, C4 and CW in 3' direction as done for C3 structure.

#### 3.3.3.2.5 Identification of the Metal Ions

The C3 has one peak in the active site close to M1 in Cat1 site and another one found far away from the active site and interacted with the dC-15. These two peaks can be occupied by calcium or sodium ions that were present in the crystallization buffer. The B-factors, the geometry and the bond lengths indicate that a $Ca^{2+}$ ion is more likely than a $Na^{+}$ ion and so these two peaks were modeled with $Ca^{2+}$ ions and the structure was refined again. A strong peak was also presented within the H3TH motif in the potassium-binding site which was identified as a $K^{+}$ ion and another cycle of refinement was run. In C4 structure there were two peaks determined as metal ions positions. One of them was in the same place of the $K^{+}$ ion that was noticed in C3 and occupied with the potassium ion. The other peak was in the duplex region of the DNA molecule between dT-5 and the symmetric dT-7 and was surrounded with two waters.

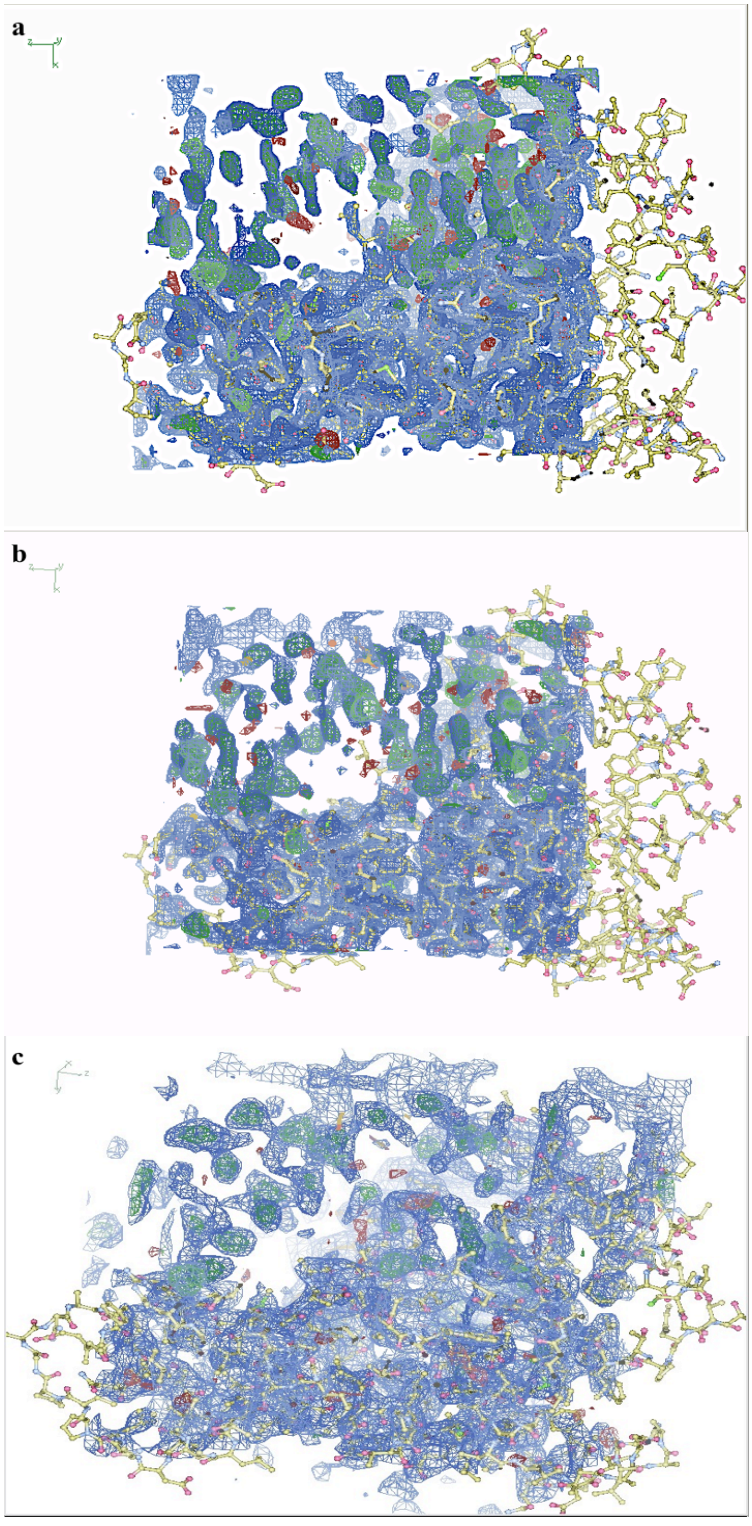



**Figure 3.33:** The electron density map for C3 and C4 variant and wild type in complexes with 3ov6 DNA.
**(a)**, **(b)** and **(c)** Show an evidence for DNA presence (green density) in C3, C4 and CW respectively.
Figures were generated using COOT (Emsley et al, 2010).

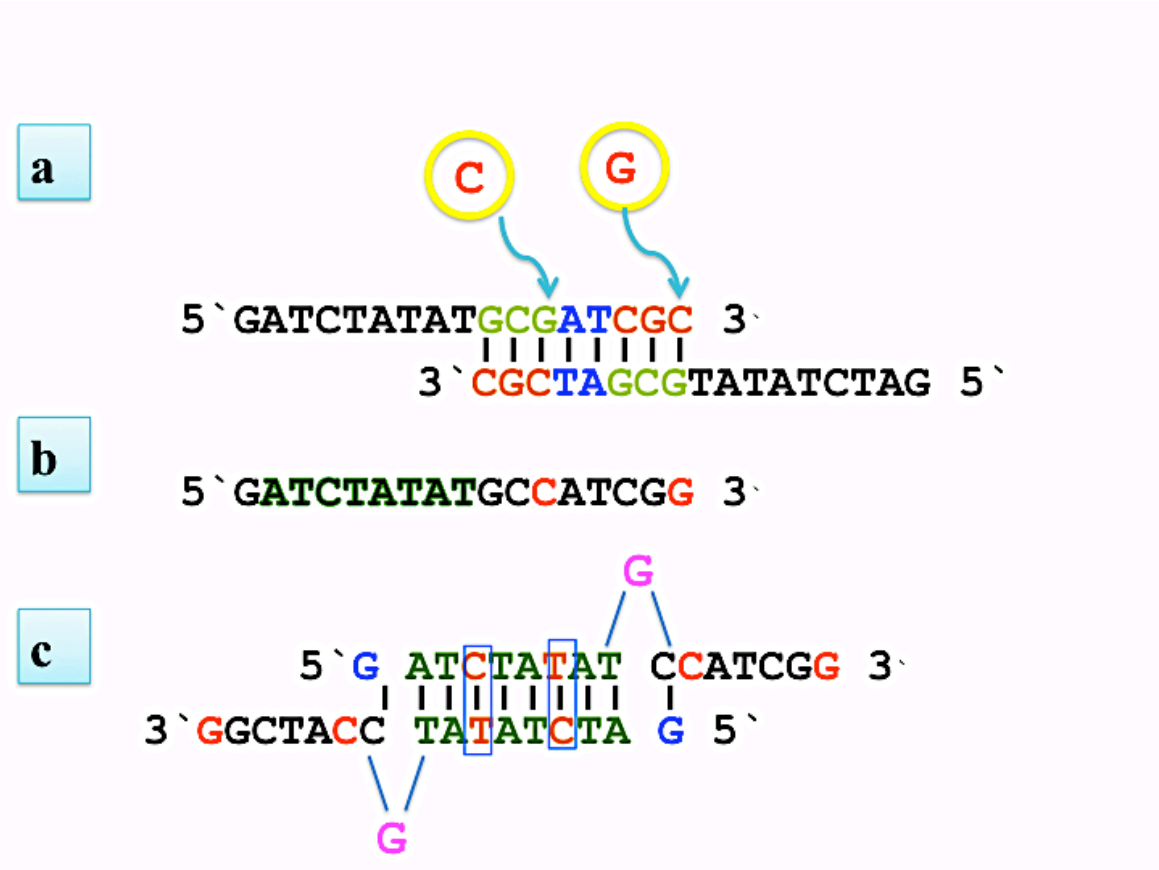



**Figure 3.34:** Diagram for DNA 3ov6 substrate sequence.
**(a)** Shows the original designed sequence of the DNA substrate.
**(b)** Shows the sequence for DNA oligo that was ordered from Eurofins.
**(c)** Shows the final model of the 3ov6 DNA that is complexed with T5FEN-D155K variant and wild type enzyme. The dG-10 is colored in magenta while the two mismatches (dT-C) in blue boxes.

This peak was occupied with sodium ion according to the metal geometry and the B-factor and the structure was then refined. In contrast, no any peaks for the metal ions were noticed in the wild type complex neither in the active site nor in the potassium-binding region which can be due to the low resolution of this complex.

#### 3.3.3.2.6 Addition of Solvent

The maps then were checked for the presence of any solvent molecules that were found in the crystallization and cryoprotectant buffers such as the glycerol and the polyethylene glycol which were added. The water molecules were added automatically and then were checked manually as described above (see section 3.2.3.2.5) and another cycle of refinement was run.

### *3.3.3.3 Structures Stereochemistry*

The final models for the two structures of T5FEN-D155K variant and WT in complexes with 3ov6 DNA are composed of 271 residues for one molecule in asymmetric unit complexed with one molecule of DNA 3' overhang in each structure. All of these models were input into PROCHECK (Laskowski et al, 1993) and MolProbity (Davis et al, 2007) softwares one by one for structure validation. C3 structure has the highest resolution (1.86 Å) has been obtained for T5FEN enzymes in complexes with 5' or 3' overhang DNA during this work. Figures 3.14, 3.15 and 3.16 are illustrated the Ramachandran plots for C3, C4 and wild type structures respectively.

According to the MolProbity analysis for C3, the clashscore was 2.47 corresponding to 99$^{th}$% percentile and no Ramachandran outliers were detected (Figure 3.14). 1.31% of conformations marked as rotamer outliers while the overall MolProbity score came for 100$^{th}$ percentile. Clashscore for all atoms in the C4 structure was 3.83 corresponding to 99$^{th}$% percentilewhile Ramachandran outliers reported 0.37% for one amino acid out of 271 across the molecule (Figure 3.15). 3.46% of sidechains poor rotamers and 95$^{th}$ percentile for MolProbity score. The CW had the highest

clashscore among these three structures with a score of 9.03 corresponding to 97$^{th}$ percentile which is likely due to the data resolution (~3 Å). In addition, 1.85% was recorded for Ramachandran outliers for five residues out of 271 across the molecule and 7.46% of side-chains exhibited poor rotamers (Figure 3.16) with a 83$^{rd}$ percentile MolProbity score.

Theses results for these crystal structures provide strong evidence that the final models of T5FEN-D155K complexed with 3' overhang DNA are good quality. The final model statistics are present in Table 3.6.

**Table 3.8:** Data refinement statistics for T5FEN-D155K:DNA and wild type crystals.

| **Data Refinement statistics** | **T5FEN-D155K:DNA and WT:DNA** | | |
|---|---|---|---|
| | C3 | C4 | CW |
| Number of reflections | 35532 | 21164 | 9474 |
| $R_{work}/R_{free}$ | 0.1824/0.2165 | 0.1935/0.263 | 0.2395/0.3036 |
| No. Atoms (protein/other) | 2163/614 | 2180/512 | 2077/370 |
| No. Waters/ Average B-factors | 256/41.067 | 134/41.994 | 28/40.325 |
| B-factors of bond atoms (Å$^2$) | | | |
| Overall | 33.372 | 37.416 | 49.349 |
| Main chain | 28.945 | 34.131 | 47.877 |
| Side chain | 36.208 | 39.644 | 50.528 |
| DNA (X) | 40.22 | 44.829 | 56.746 |
| R.m.s deviation from ideality | | | |
| Bond lengths (Å) | 0.0273 | 0.0215 | 0.0153 |
| Bond angles (°) | 2.645 | 2.029 | 1.946 |
| Ramachandran plot. Proportion in residues in: | | | |
| Allowed regions% | 93.0 | 91.9 | 86.6 |
| Additional allowed regions% | 6.9 | 7.7 | 12.1 |
| Generously allowed regions% | 0.0 | 0.4 | 0.8 |
| Disallowed regions% | 0.0 | 0.0 | 0.4 |
| Average RMS B-factors (main chain/side chain) | 1.018/3.329 | 0.818/2.652 | 0.476/1.519 |
| Average RMS B-factors of DNA (X) | 5.411 | 4.819 | 4.372 |

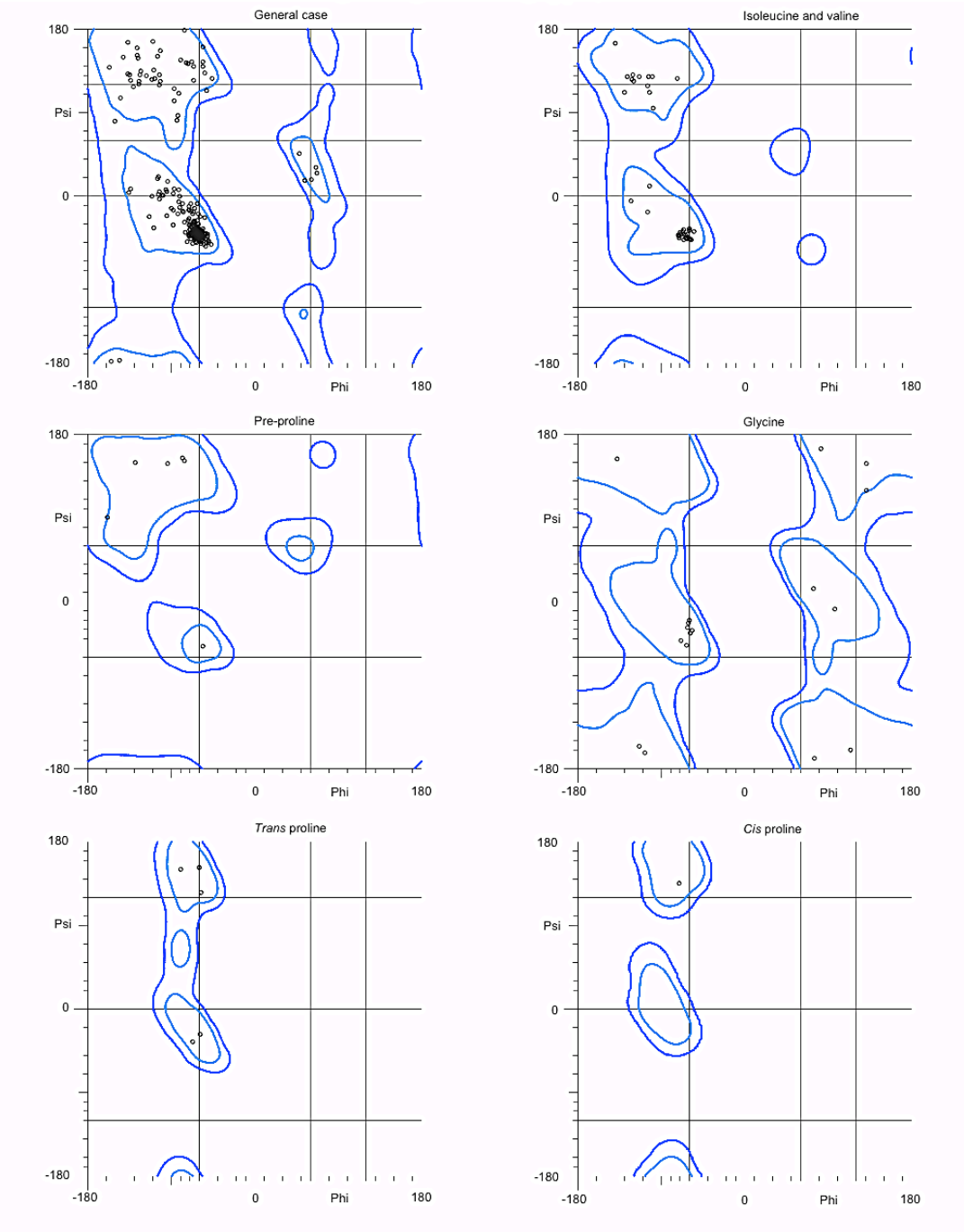



**Figure 3.35:** Shows Ramachandran plot for the final model of C3 structure. Figure was produced by MolProbity http://molprobity.biochem.duke.edu/

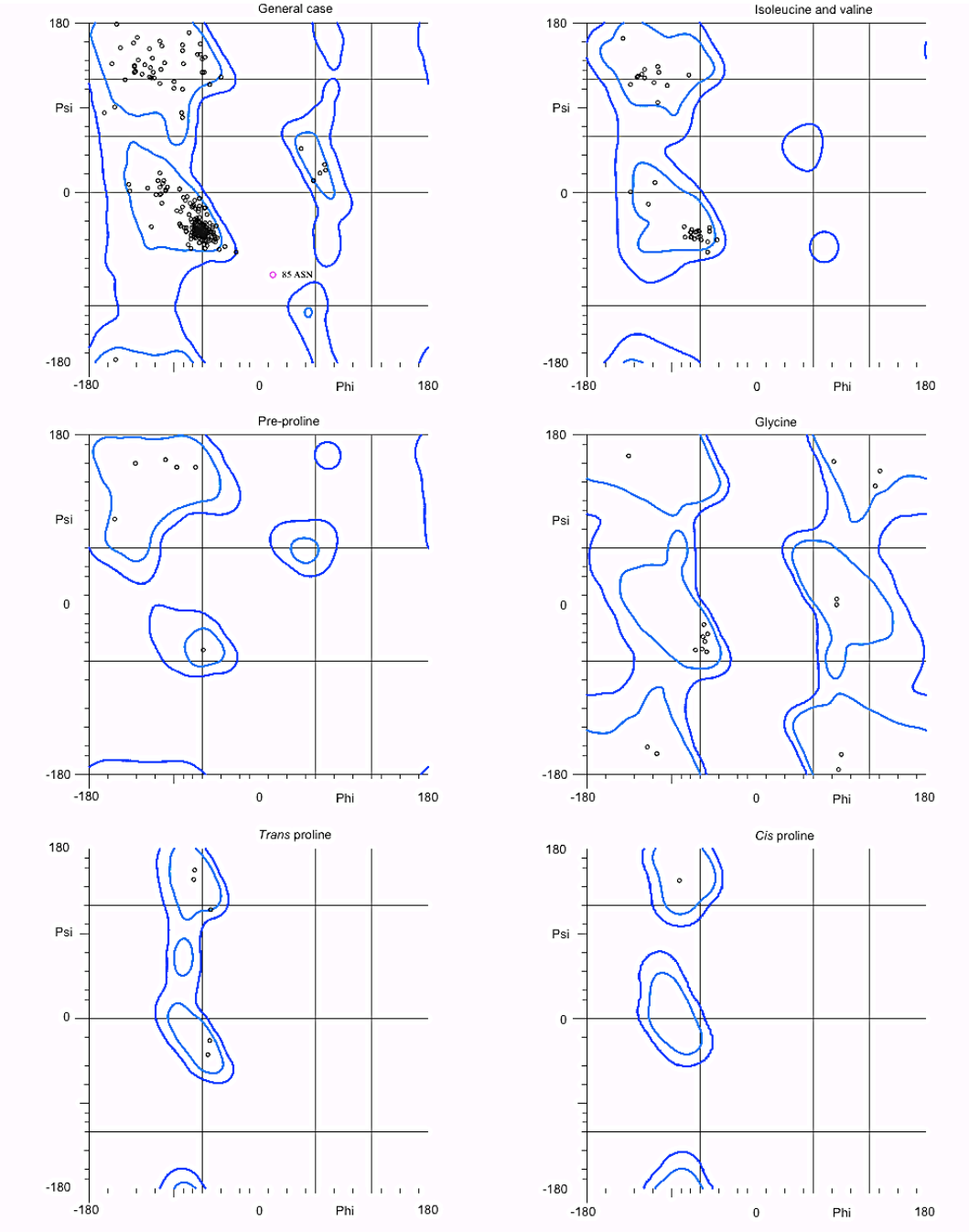



**Figure 3.36:** Shows Ramachandran plot for C4 structure.
Figure was produced by MolProbity http://molprobity.biochem.duke.edu/

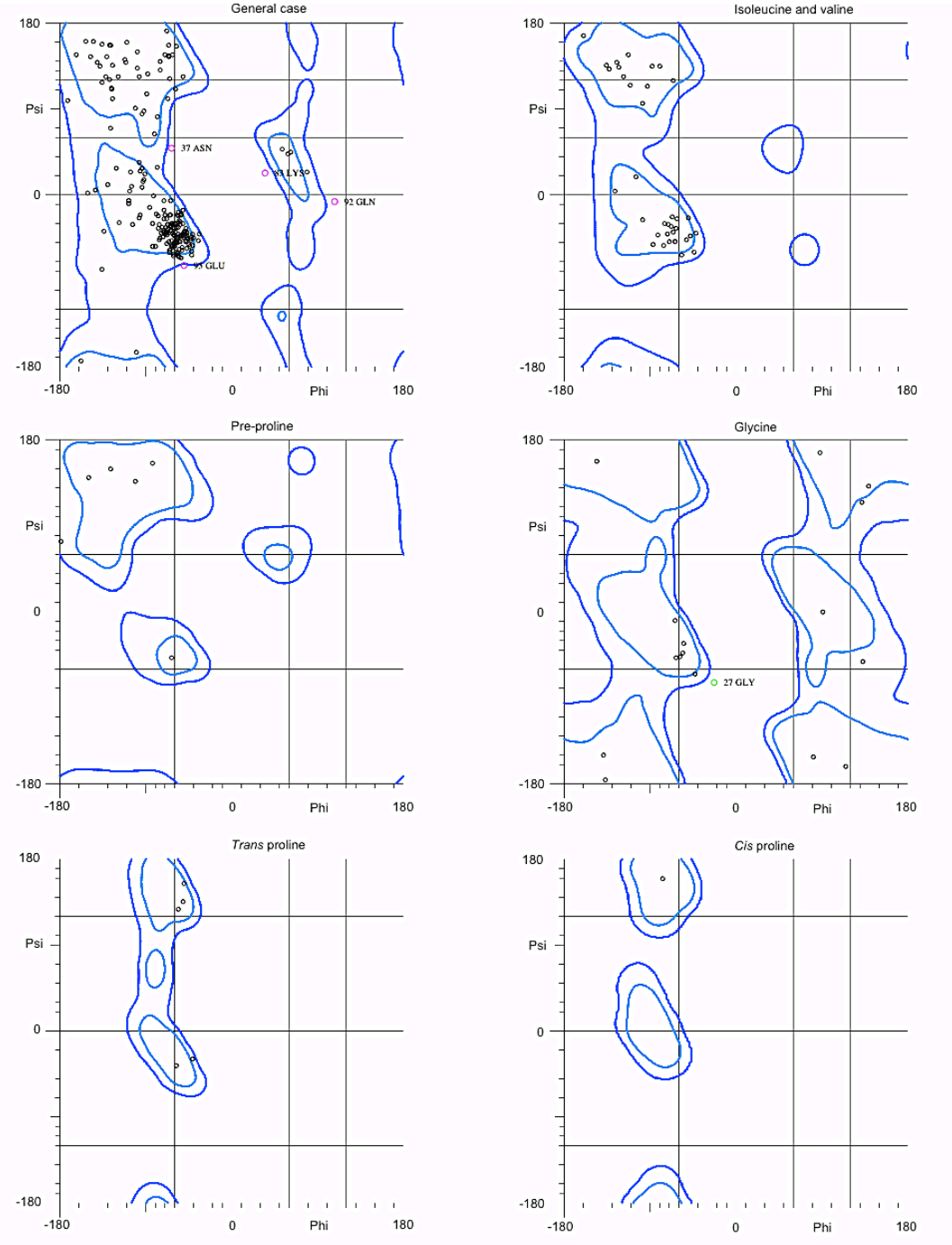



**Figure 3.37:** Shows Ramachandran plot for CW structure.
Figure was produced by MolProbity http://molprobity.biochem.duke.edu/

# Chapter 4: T5FEN-D153K in a Complex with 5ov4 DNA

This chapter describes and discusses the two structures of the Asp153Lys variant of bacteriophage T5 D15 exonuclease (T5FEN-D153K) in complexes with a palindromic DNA substrate (5ov4). The structural features of these complexes will be compared with each other and with T5FEN native protein complexed with magnesium ions. The T5FEN-D153K complexed with 5ov4 DNA crystallized in many conditions. Two crystals were solved at high resolution using molecular replacement method as described in chapter three.

## 4.1 Results

### 4.1.1 Overall Structure of C1

This structure was solved in space group $P2_12_12_1$ at 2.16 Å resolution and two molecules are present in the asymmetric unit complexed with one palindromic DNA that is composed of eight base pairs with four base 5' overhangs. The protein molecules are called chain A and B (Figure 4.1). Chain A is built with 271 residues from 20-291 while chain B lacks the last amino acid (Gln-291) and so is composed of residues 20-290. The two monomers display identical fold to that described in section 1.3.2 for T5FEN with an ordered arch. Residue His-36 is disordered in chain A and some residue side-chains are totally truncated such as Arg-20, Asn-38, Lys-40, Lys-41, Arg-207, Asn-226 and Pro-235 while Lys-280 side chain is partially disordered due to the poor electron density in these locations. Additionally, four residues from 38-41 are disordered completely in chain B with some other fully truncated side-chains including Arg-20, Asn-37, Lys-56, Lys-71, Lys-89, Tyr-90, Leu-100, Glu-102, Gln-103, Glu-106, Lys-109, Arg-207, Lys-239 and Lys-241 or partially deleted such as Glu-95, Glu-96, Lys-116 and Lys-163. The absent residues in chain A and B are in the loop connecting $\alpha 1$ and $\alpha 2$.

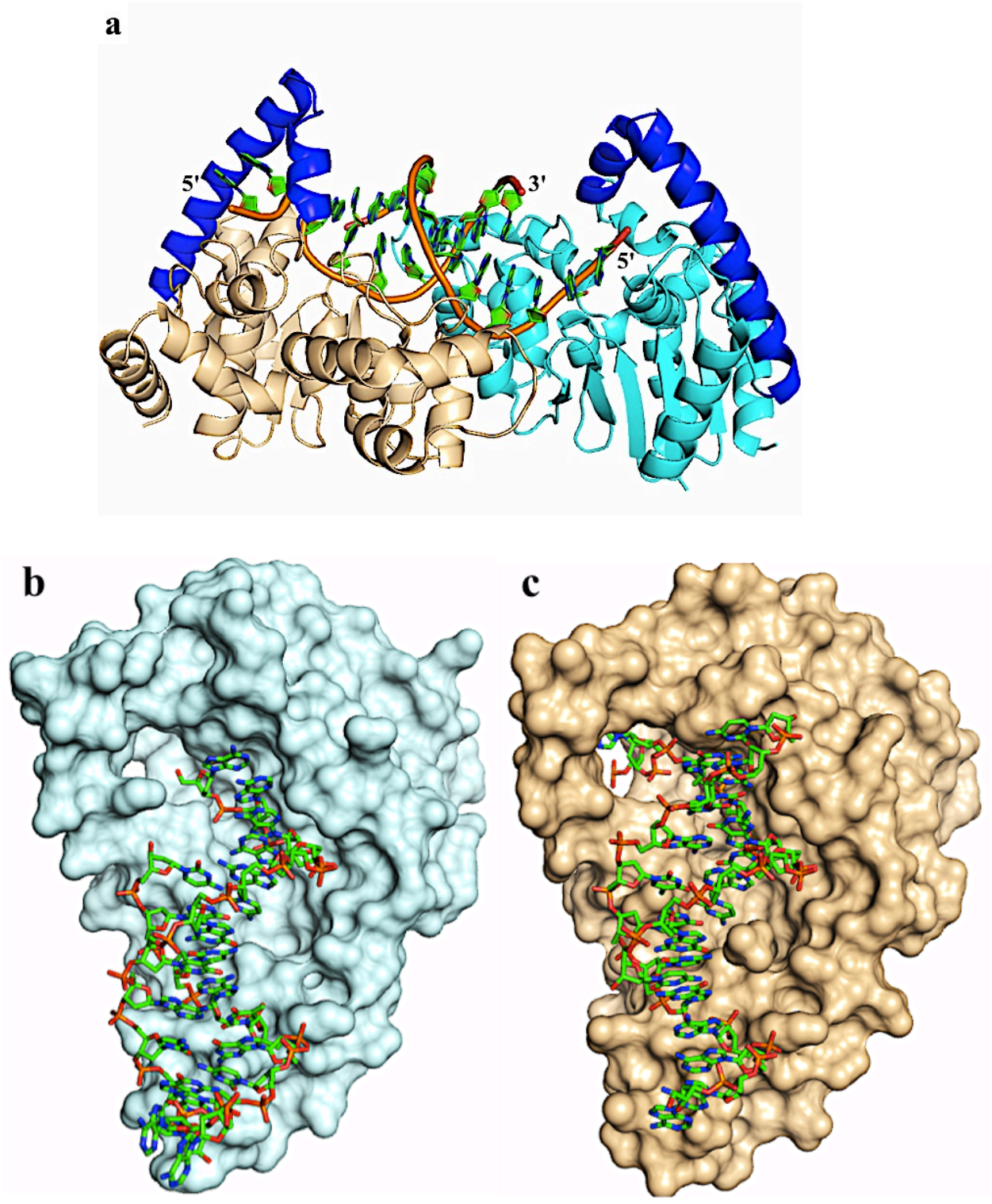



**Figure 4.38:** Diagrams for C1-T5FEN-D153K:DNA structure.
**(a)** Shows two protein molecules, C1/A (cyan) and C1/B (wheat) complexed to one palindromic DNA molecule (orange backbone and green bases). The arch-like structure is colored in blue. **(b)** and **(c)** illustrate views in Y-axis for C1/A and C1/B with the DNA substrate respectively.

#### *4.1.1.1 Monomers Interface*

The two chains A and B interact with the neighboring molecule in the crystal through hydrogen bonds and salt bridges. According to the protein-protein interface analysis between the crystal protein chains carried out using the pdbsum service (www.ebi.ac.uk/pdbsum), chain A interacts with chain B' from the neighboring molecule (chain B', Figure 4.2) *via* three hydrogen bonds formed from its residues Lys-116 and Thr-117 main-chain carbonyl groups to chain B' residues Asn-21 side chain ND2 and Trp-148 side chain NE1 (Figure 4.3a). However, no salt bridges were found for this chain A. Additionally, chain B of this complex also interacts with chain A" from the neighboring molecule (chain A", Figure 4.2) *via* a salt bridge and nine hydrogen bonds, five of them are considered as salt bridges in the same time (Figure 4.3b). These hydrogen bonds form by chain B Lys-138, His-142, Gln-275, Thr-282 and Glu-290 side chains to chain A" Glu-81, Glu-88, Glu-77 side chains and Lys-71 main chain amide nitrogen and Arg-125 side chain respectively (Figure 4.3b). The other hydrogen bonds form by Leu-139 and Glu-290 main chain carbonyl groups to Lys-71 side chain NZ and Arg-125 side chain two amine groups (Figure 4.3b). The hydrogen bonds that are formed by chain B Lys-138, His-142 and Glu-290 were also considered to be salt bridges. However, another salt bridge form between Glu-287 side chain from chain B and Arg-125 side chain of chain A" (Figure 4.3b). These protein-protein interactions between T5FEN-D153K:5ov4 monomers are summarized in Tables 4.1, 2 and 3.

#### *4.1.1.2 The Presence of Ammonium Sulphate*

Ammonium sulphate (Figure 4.4) was present in the crystallization buffer in concentration of 2 M. Thirteen sulphate ions $(SO_4)^{2-}$ appeared in the structure (Figure 4.5a) and coordinate with the protein residues. One $(SO_4)^{2-}$ ion was found between the two DNA strands, X and Y and interacts with their bases by forming hydrogen bonds to strand-X dC-6 and strand-Y dC-10 nitrogens-4 and a water molecule (Figure 4.5b). Chain A was surrounded with five $(SO_4)^{2-}$ ions while the other seven ions interact with chain B (Figure 4.5a).

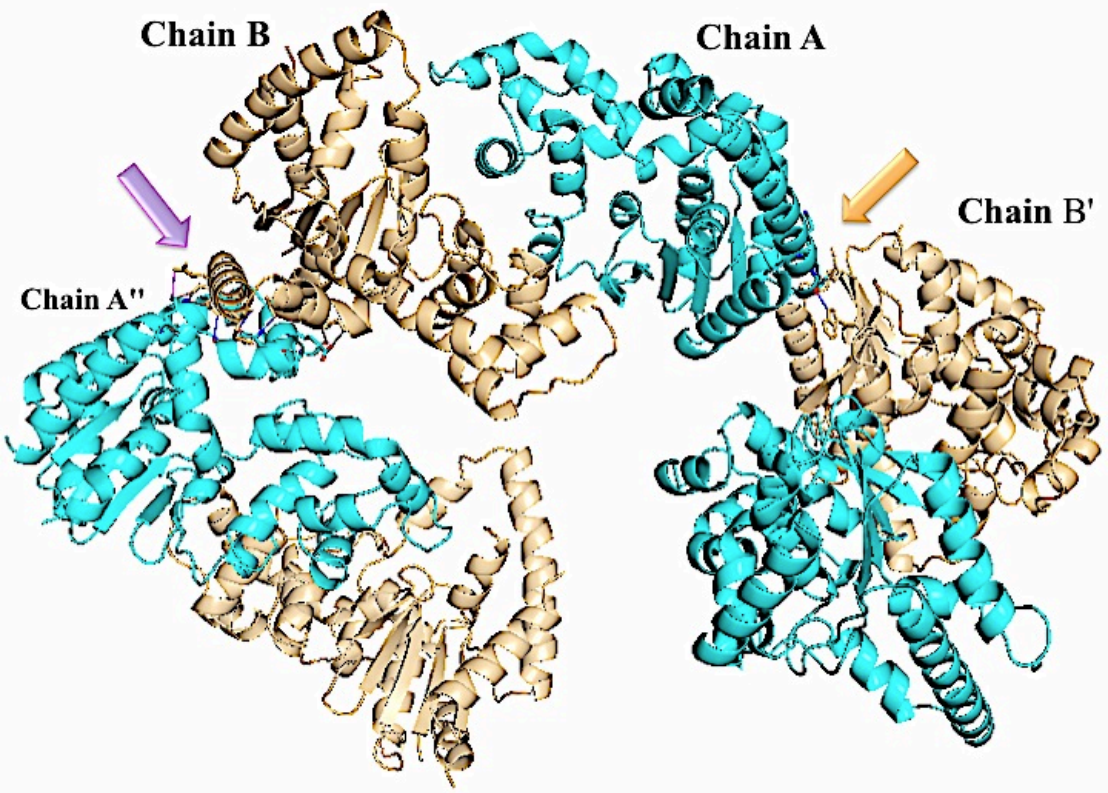



**Figure 4.39:** Diagram for monomers interface of C1 in the crystal.
The figure represents the protein-protein interface between the complex monomers with neighboring molecules. Chain A (cyan) interacts symmetrically with the neighboring molecule chain B' (wheat) and the interaction area is shown by orange arrow. Chain B (wheat) of the complex also interacts symmetrically with the neighbor molecule chain A'' (cyan) and the interaction area is shown by magenta arrow.

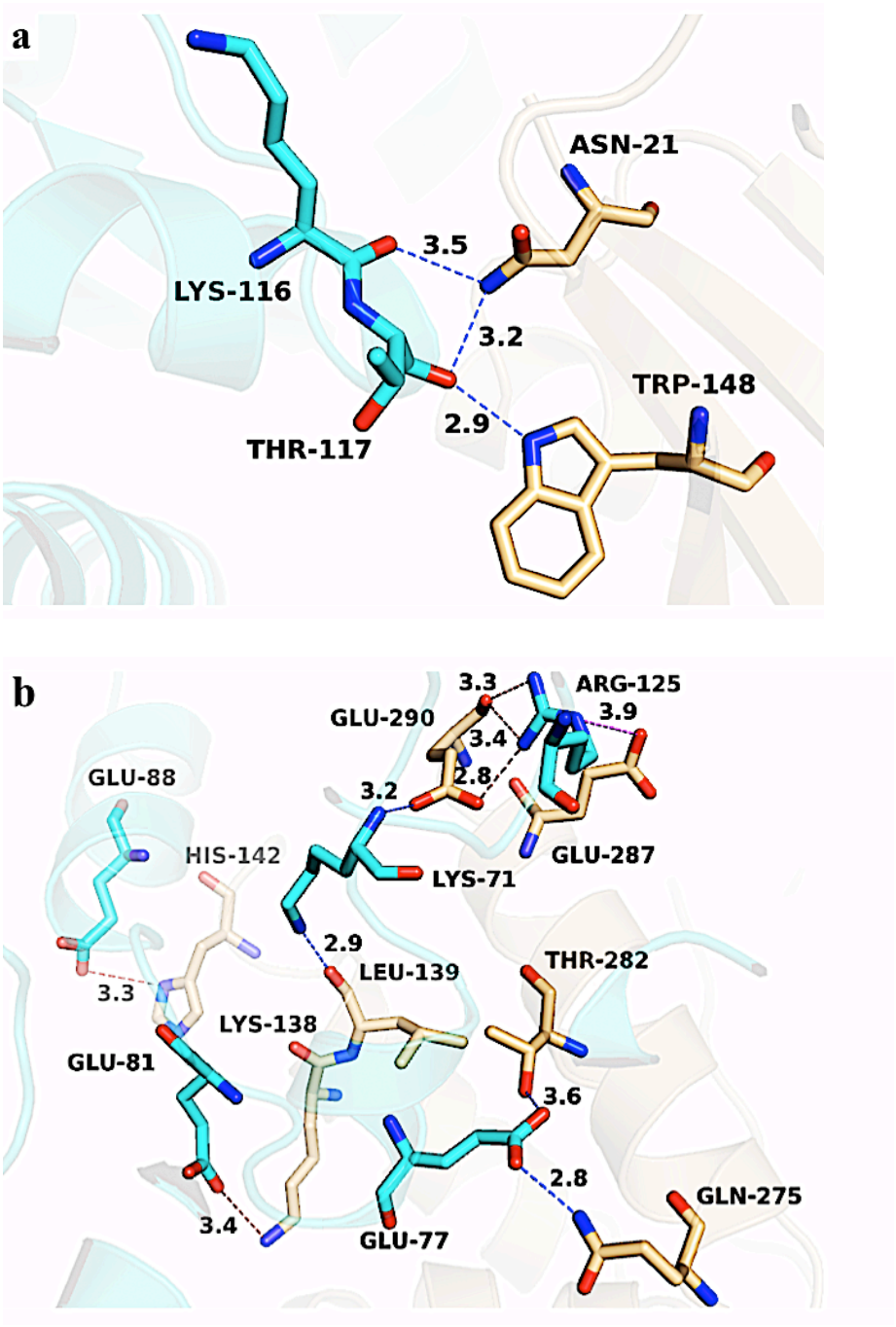



**Figure 4.40:** Protein-protein interactions between C1 monomers.
**(a)** Shows the interactions between chain A (cyan sticks) and the neighboring molecule chain B' (wheat sticks) *via* three hydrogen bonds formed by two residues. **(b)** Illustrates the interactions between chain B (wheat sticks) and the neighboring molecule chain A" (cyan sticks) through nine hydrogen bonds (blue), five of them considered as salt bridges (ruby) and a salt bridge (magenta).

**Table 4.9:** Hydrogen bonds between C1/A and the neighboring molecule chain B'.

| Chain | Resi. ID | Distance Å | Chain | Resi. ID |
|---|---|---|---|---|
| **A** | Lys-116 [O] | 3.46 | **B'** | Asn-21 [ND2] |
| **A** | Thr-117 [O] | 2.86 | **B'** | Trp-148 [NE1] |
| **A** | Thr-117 [O] | 3.17 | **B'** | Asn-21 [ND2] |

**Table 4.10:** Hydrogen bonds between C1/B and the neighboring molecule chain A".

| Chain | Resi. ID | Distance Å | Chain | Resi. ID |
|---|---|---|---|---|
| **B** | Lys-138 [NZ] | 3.37 | A" | Glu-81 [OE2] |
| **B** | Glu-139 [O] | 2.89 | A" | Lys-71 [NZ] |
| **B** | His-142 [ND1] | 3.41 | A" | Glu-88 [OE2] |
| **B** | Gln-275 [NE2] | 2.81 | A" | Glu-77 [OE2] |
| **B** | Thr-282 [OG1] | 3.61 | A" | Glu-77 [OE1] |
| **B** | Glu-290 [OE1] | 3.19 | A" | Lys-71 [N] |
| **B** | Glu-290 [OE2] | 2.76 | A" | Arg-125 [NH1] |
| **B** | Glu-290 [O] | 3.35 | A" | Arg-125 [NH1] |
| **B** | Glu-290 [O] | 3.33 | A" | Arg-125 [NH2] |

**Table 4.11:** Salt bridges between C1/B and the neighboring molecule A".

| Chain | Resi. ID | Distance Å | Chain | Resi. ID |
|---|---|---|---|---|
| **B** | Lys-138 [NZ] | 3.37 | A" | Glu-81 [OE2] |
| **B** | His-142 [ND1] | 3.41 | A" | Glu-88 [OE2] |
| **B** | Glu-287 [OE1] | 3.92 | A" | Arg-125 [NE] |
| **B** | Glu-290 [OE2] | 2.76 | A" | Arg-125 [NH1] |
| **B** | Glu-290 [O] | 3.35 | A" | Arg-125 [NH1] |
| **B** | Glu-290 [O] | 3.33 | A" | Arg-125 [NH2] |

In C1/A the first $(SO_4)^{2-}$ ion surrounds by Arg-33, Phe-34 and Phe-169 residues but it does not make any interactions with them through hydrogen bonds (Figure 4.6a) while the second ion interacts with Lys-69 NZ and Arg-125 guanidinium group *via* hydrogen bonds (Figure 4.6b). The other group of $(SO_4)^{2-}$ ions is found around C1/B and they occupy the same places that have been seen in chain A (Figure 4.5a). The first ion coordinates with Arg-33 through hydrogen bonds formed by its guanidinium group and also surrounds with Phe-34 and Phe-169 without interactions (Figure 4.6c). Another ion forms hydrogen bonds to Lys-69 NZ and Arg-125 guanidinium group (Figure 4.6d). The interactions of C1-T5FEN:DNA chain A and B with other $(SO_4)^{2-}$ ions are written in Table 4.4.

#### *4.1.1.3 The 5ov4 DNA Interactions*

One molecule of 5ov4 overhang DNA is complexed with two protein molecules A and B in the asymmetric unit. The interactions between the DNA substrate and the protein monomers are very similar in both T5FEN-D153K structures and can be divided into several sites.

##### 4.1.1.3.1 Interactions with H3TH:$K^+$ Motif

The H3TH motif is well known as a DNA binding motif and found in many DNA binding proteins as described in section (1.2.1.1.1). In C1/A and C1/B, H3TH motif binds to the 5ov4 DNA in the 3' part (strand Y & X respectively) of the duplex through four residues; Gly-211, Gly-213, Lys-215 and Arg-216. Hydrogen bonds are formed between the main chain amide nitrogen of these amino acids and deoxycytidine-10/Y and adenosine-9/Y phosphate oxygens of the DNA backbone for chain A. Additional interaction can be observed by chain A H3TH motif through Ala-214 main chain amide nitrogen to dA-9/Y phosphate oxygen (Figure 4.7a). Chain B interacted with deoxycytidine-6/X and guanosine-5/X phosphate oxygens (Figure 4.7B). Indirectly, C1/B also interacted with the DNA phosphate group by three residues Met-199, Val-209 and Ile-212 *via* potassium ion (Figure 4.7b).

**Table 4.12:** The $(SO_4)^{2-}$ ions present in C1 structure and their interactions with the protein residues.

| Chain | $(SO_4)^{2-}$ Ions | | Coordination Residues |
|---|---|---|---|
| **A** | 1 | - | Arg-33, Phe-34 and Phe-169 |
| | 2 | O2 | Arg-178 (NH2) |
| | | O3 | His-176 (ND1) |
| | | O4 | H2O |
| | 3 | O3 | Glu-174 (N) |
| | | O4 | Glu-174 (O, N), Arg-173 (NH2) |
| | 4 | O1 | Lys-69 (NZ) |
| | | O2 | Arg-125 (NH2) |
| | | O4 | Arg-125 (NE) |
| | 5 | O1 | Ala-44 (N), Ser-45 (N) |
| | | O2 | Pro-42 (O) |
| | | O3 | Ser-45 (OG) |
| | | O4 | Ala-44 (N) |
| **X & Y** | 6 | O2 | dC-6-X (N4), H2O |
| | | O4 | dC-10-Y (N4) |
| **B** | 7 | O1 | Arg-178 (NH1) |
| | | O3 | Arg-178 (N), Lue-177 (N) |
| | | O4 | H2O |
| | | O4 | Asp-276 (OD1) |
| | 8 | O1 | Thr-28 (OG1) |
| | 9 | O1 | Lys-69 (NZ) |
| | | O3 & O4 | Arg-125 (NE) & (NH2) |
| | | O3 | Asp-187 (OD1) |
| | 10 | O3 | Arg-33 (NE) & (NH2) |
| | 11 | O1, O4 | Gln-275 (N) |
| | | O4 | Asp-276 (OD1) |
| | 12 | O1, O4 | Thr-264 (O) |
| | | O2 | H2O |
| | | O4 | EDO-2 (O1) |
| | 13 | O1, O2 | 2H2O |
| | | O3 | Asp-187 (OD1) |

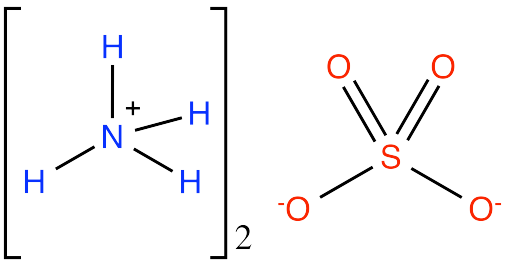



**Figure 4.41:** The chemical formula of $(NH_4)_2SO_4$.

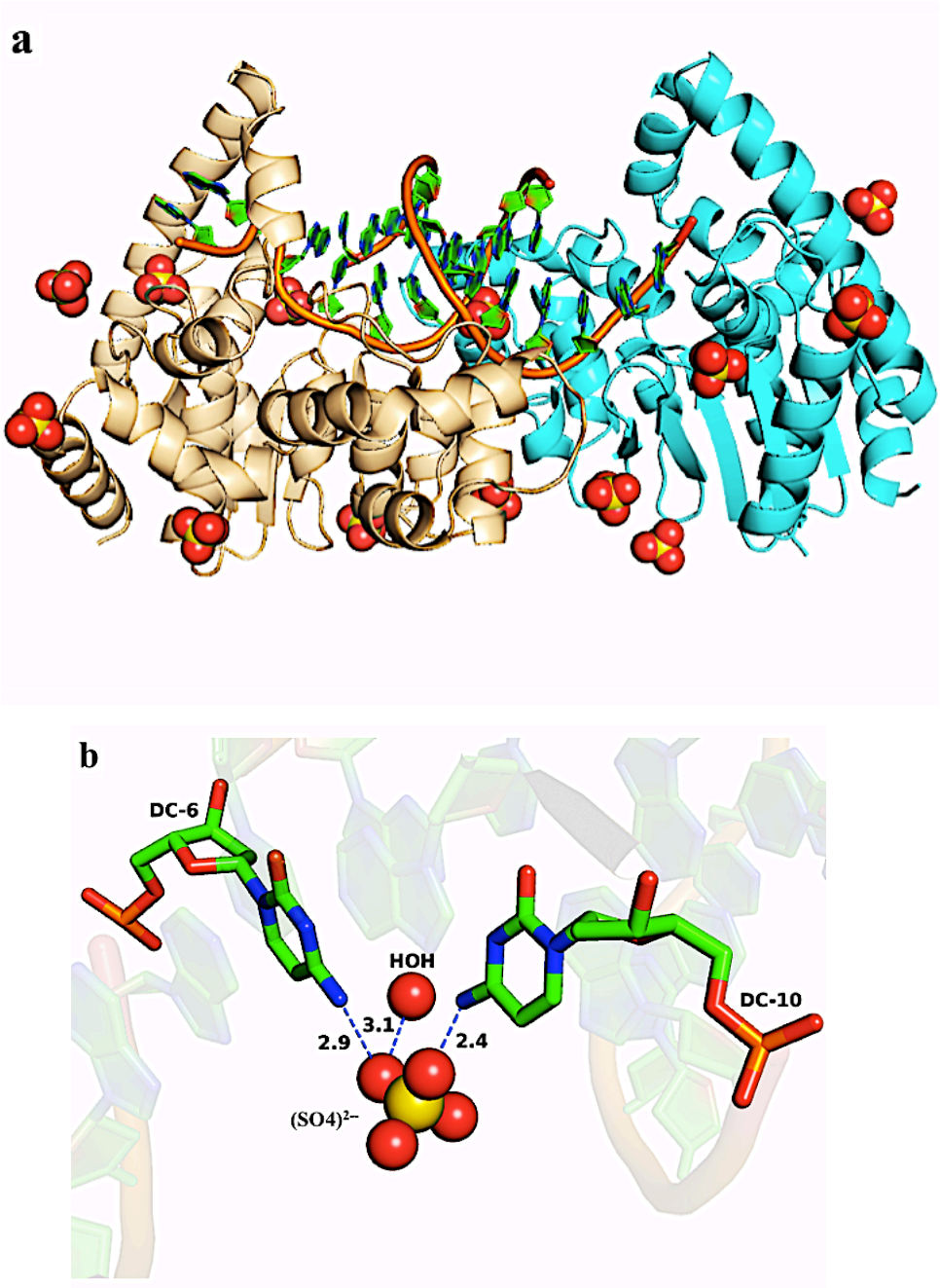



**Figure 4.42:** The presence of sulphate ions in C1 structure.
**(a)** Shows the two chains of the complex, chain A (cyan) and chain B (wheat) surrounded with sulphate $(SO_4)^{2-}$ions (red and wheat spheres). **(b)** Illustrates one of the $(SO_4)^{2-}$ions (red and wheat spheres) between the DNA strands, X and Y (orange backbone and green bases) which coordinates with the DNA bases and a water molecule.

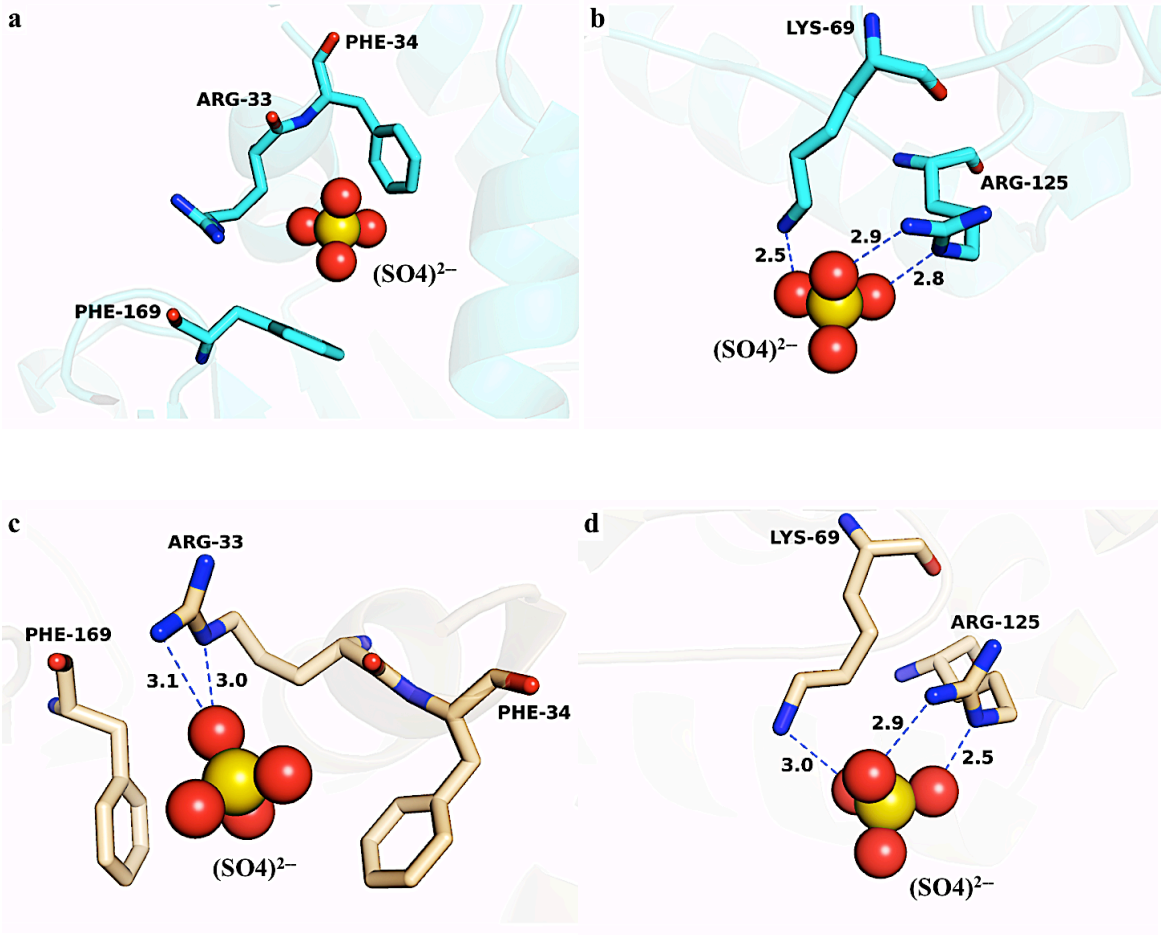



**Figure 4.43:** The interactions between chain A residues (cyan sticks) and B (wheat sticks) with $(SO_4)^{2-}$ions (red and wheat spheres) through hydrogen bonds.
**(a)** and **(b)** show the presence of two $(SO_4)^{2-}$ions close to Arg-33, Lys-69 and Arg-125 respectively in chain A. **(c)** and **(d)** show the same residues in chain C1/B.

Hydrogen bonds form by the main chains carbonyl of these residues to coordinate the $K^+$ ion which interacts with dC-6/X phosphate oxygen. Interestingly, in this complex structure the potassium ion is found in one chain only, C1/B, with low peak of density which does not obtain in its binding site in chains A. Gly-211, Ile-212 and Gly-213 residues are formed the GhG sequence which is conserved in FEN family members except for T4RNase H enzyme.

#### 4.1.1.3.2 Interactions with Hydrophobic Wedge

In C1/A the DNA passes above the active site and moves to the arch but it does not reach the arch hole and does not thread through it. Instead, the 5' overhang here stacks against the hydrophobic wedge, Phe-32, while the duplex region (3' end of strand Y) from this side of the structure far from the hydrophobic residues (Figure 4.8a). The duplex DNA in C1/B of the complex extended and passed through the enzyme active site and continued to interact with the hydrophobic wedge by the 3' end (strand X) of the duplex (Figure 4.8b). Hydrogen bonds form by Lys-35 and His-36 side-chains to dG-5/Y phosphate group and the later residue, His-36, which is completely disordered in chain A stacks against dG-5/Y base (Figure 4.8c). Additionally, His-36 main chain carbonyl interacts with dC-12/X base of the DNA duplex *via* hydrogen bond (Figure 4.8c).

#### 4.1.1.3.3 The 5' overhang Interactions

Chain A interacts with the single strand-X 5' overhang backbone by forms bifurcated hydrogen bonds from Arg-86 amino groups to dA-2/X phosphate oxygens. Additional interaction can be observed between Gly-154 and Asp-155 main chain amide nitrogen and the 5' overhang dA-3/X phosphate oxygen directly for the first residue and indirectly through water for the later one (Figure 4.9a). Furthermore, Asp-155 side chain and tyrosine-175 side-chain hydroxyl group also sent hydrogen bonds to interact with the dA-3/X and dA-4/X phosphate oxygen respectively *via* water molecules whilst His-183 main chain carbonyl oxygen interacts with the same phosphate oxygen, dA-4/X, *via* another water molecule (Figure 4.9a).

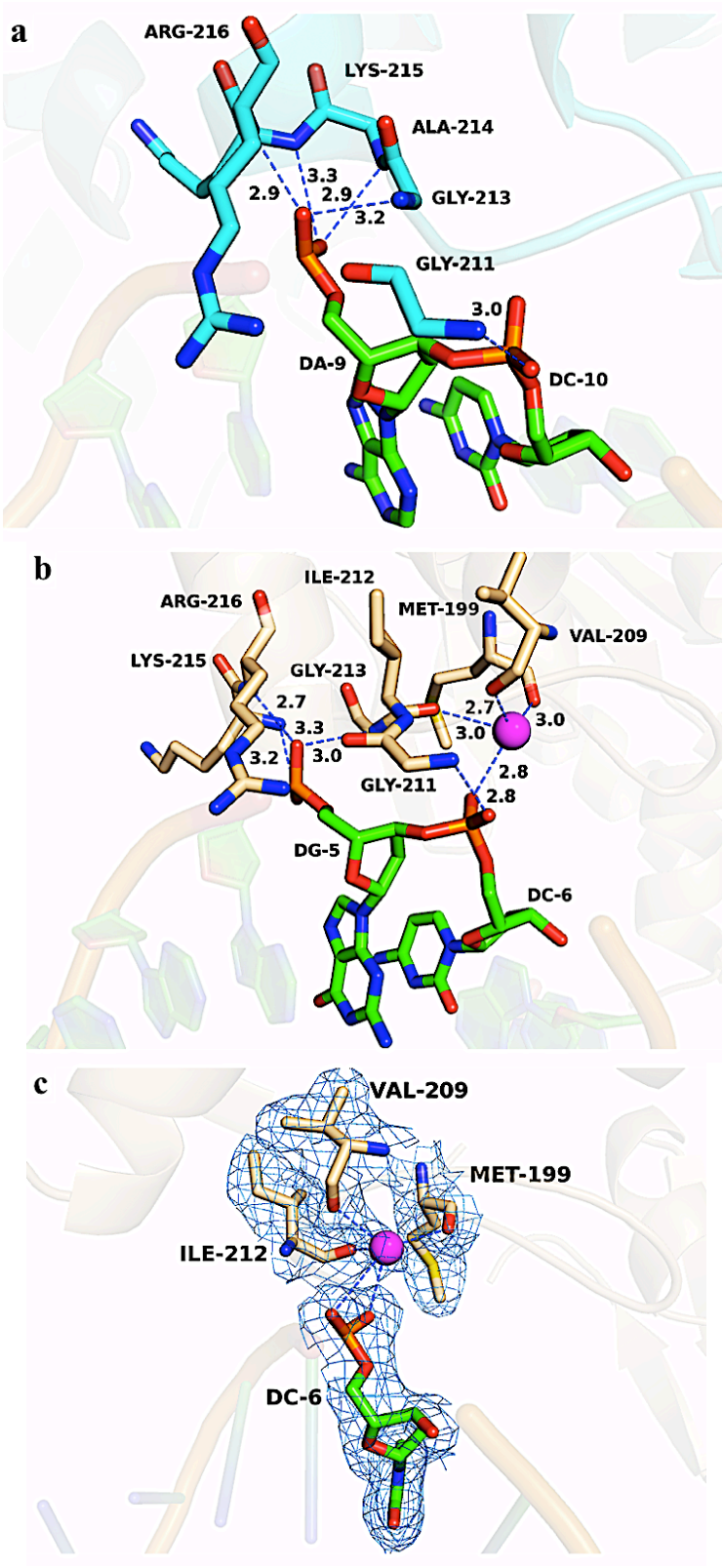



**Figure 4.44:** The H3TH motif interactions with 5ov4 DNA (orange backbone and green bases) in C1-T5FEN-D153K structure.
**(a)** Shows the interactions between C1/A H3TH motif residues (cyan sticks) with the DNA. **(b)** Shows the same interactions with C1/B H3TH motif residues (wheat sticks). The figure also illustrates the presence of the $K^+$ (magenta sphere) in its binding site in C1/B. **(c)** 2Fo-Fc electron density map for C1/B $K^+$ ion.

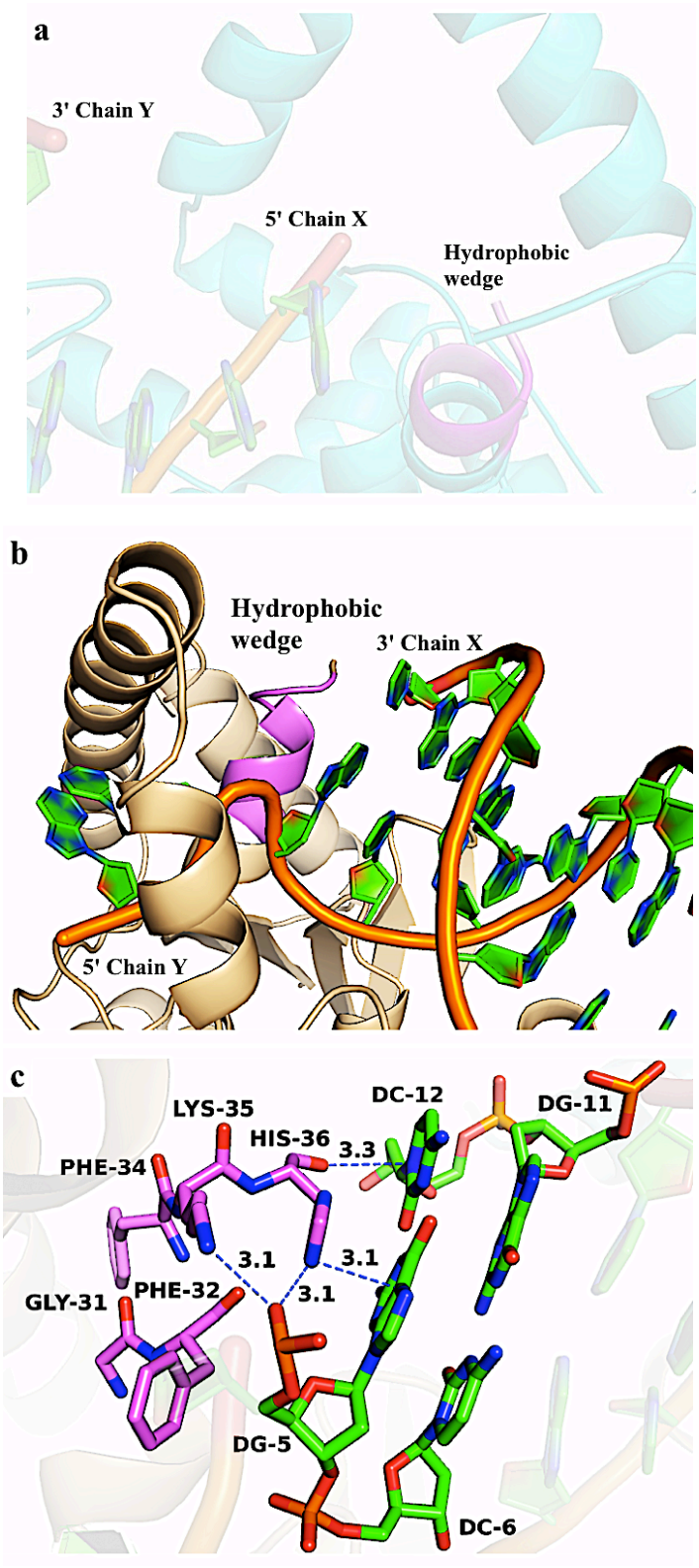



**Figure 4.45:** Diagrams for the hydrophobic wedge in C1-T5FEN-D153K:DNA. **(a)** Illustrates the hydrophobic wedge (magenta) in chain A stacks against the 5' overhang. **(b)** Shows the hydrophobic wedge (magenta) in chain B stack against the 3' duplex end of the 5ov4 DNA and a partially thread for the 5' overhang through the arch (wheat). **(c)** The interactions between the hydrophobic wedge (magenta sticks) in C1/B and the DNA through hydrogen bonds.

Interestingly, there is a positive density around dA-2/X phosphate group which joined the first and the second bases together and a negative density presents in the middle of this phosphate group as seen in Figure 4.9b. In chain B of this complex, the 5' overhang moves directly toward the arch and makes a curve-like structure before threaded partially through the arch hole (Figure 4.9c). Only two bases of the DNA substrate (dA-3 and dA-4) can be observed in this 5' end of the DNA while the other two (dA-1 and dA-2) are missing in this complex. Deoxyadenosine-3 and deoxyadenosine-4 that are inserted through the arch in C1/B make some interactions with the arch residues. Arginine-93 side-chain amino group forms hydrogen bonds to dA-4/Y base while Arg-86 interacts through its amino groups nitrogen with dC-6/Y phosphate oxygen by forming hydrogen bond indirectly *via* two water molecules (Figure 4.9c).

#### *4.1.1.4 The Active Site and Metal Ions*

This crystal structure is for the T5FEN-D153K mutant which means the metal ions should not bind to M1 of the active site. In this complex no metal ions were observed either in chain A or in chain B active sites (Figure 4.10a & b) respectively. The mutated residue, D153K, was clearly visible in both chain's active sites.

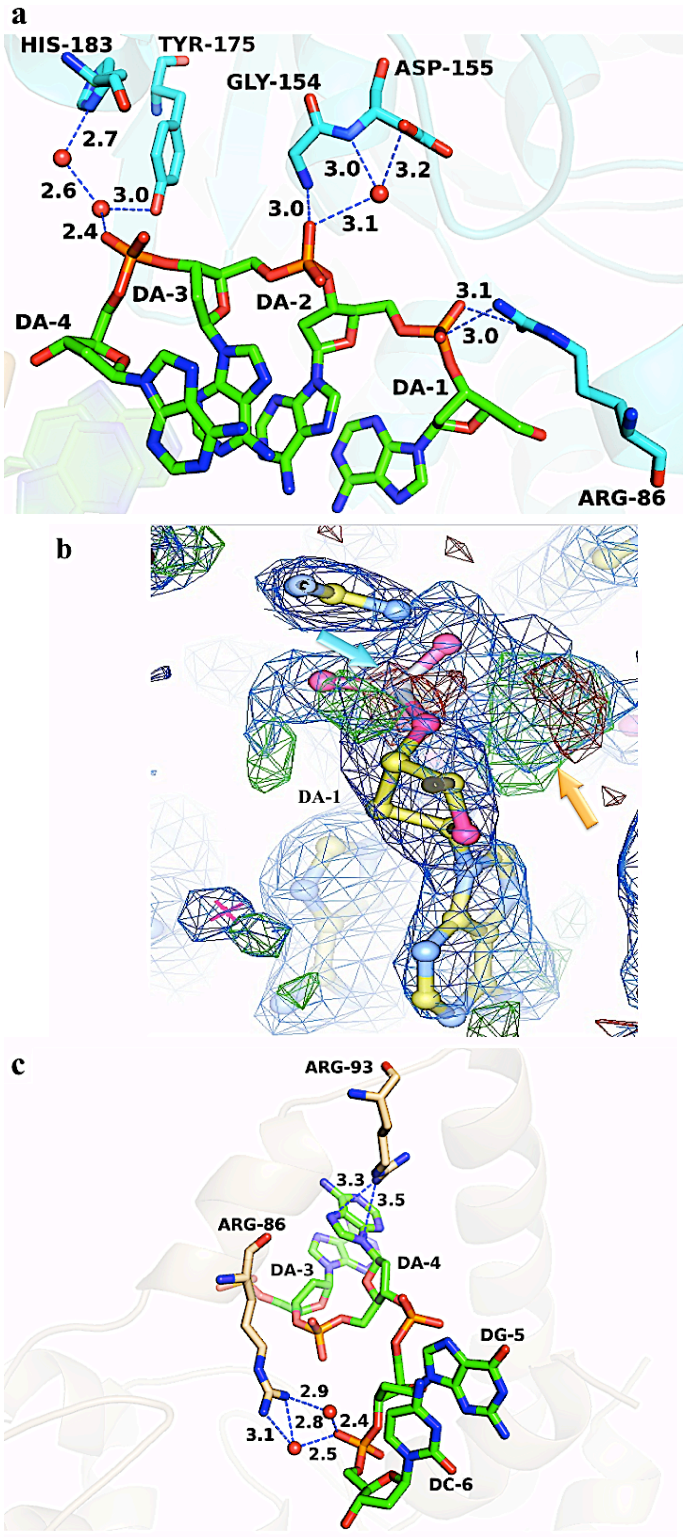



**Figure 4.46:** The DNA 5' overhang interactions with C1 structure.
**(a)** Shows the interactions between C1/A (cyan sticks) with strand-X 5' overhang (green bases). **(b)** Illustrates the first base in strand X with the positive density (green) pointed by orange arrow around dA-2 phosphate group that joined the first two bases together and the negative density (red) in the middle shown by cyan arrow. **(c)** Shows the interactions between C1/B (wheat sticks) with strand-Y 5' overhang (green bases).

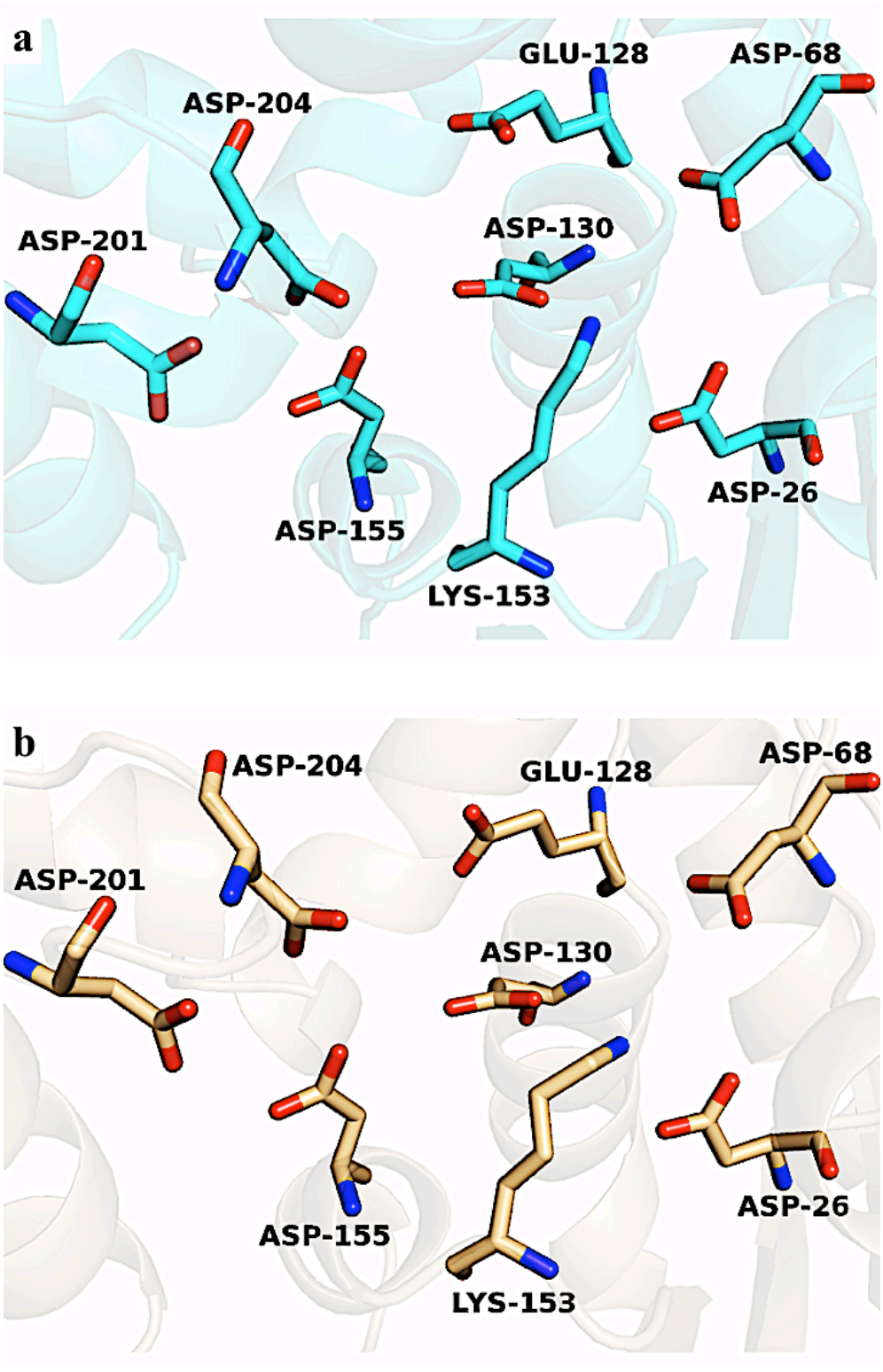



**Figure 4.47:** The C1 conserved active site residues.
**(a)** Shows C1/A (cyan sticks) active site. **(b)** Shows C1/B (wheat sticks) active site residues. Both chains active sites are metal ion-free.

### 4.1.2 Overall Structure of C2

Two protein chains, A and B were present in the asymmetric unit complexed with one molecule of 5' overhang DNA (Figure 4.11) and each one consists of 271, equivalent to residues 20-291 of the native enzyme. The DNA is not symmetrically bound in this cocrystal as more extensive interactions with protein chain B were observed than with the polypeptide in chain A. The two monomers display the same architecture of T5FEN with an ordered arch. In molecule A the electron density for the side chains of residues Arg-20, Lys-56, Glu-95, Lys-98, Glu-102, Glu-106, Lys-109, Thr-118, Glu-190, Arg-222, Lys-239, Lys-241, Glu-250, Glu-251, and Gln-291 are too poor to allows building so they truncated totally while Leu-54, Lys-69, Glu-113, Lys-163, Leu-202 and Arg-207 side chains partially deleted and residues Ser-39 and Lys-40 are disordered and cannot be seen at all.

In contrast, the electron density for the amino acids Arg-20, Asn-38, Ser-39, Lys-41, Glu-88, Lys-89, Glu-95, Lys-109, Lys-163, Glu-190 and Lys-241 side chains are totally deleted while Lys-40, Lys-69, Glu-96, Leu-98 and Lys-239 side chains are partially truncated. These disordered residues: Ser-39 and Lys-40 in chain A form part of the loop between the two helices ($\alpha$1 and $\alpha$2).

#### *4.1.2.1 Monomers Interface*

The interface between two monomers A and B was investigated through the pdbsum service (www.ebi.ac.uk/pdbsum). According to this analysis there are ten hydrogen bonds at the interface between chains A and B in the same unit cell and three salt bridges (Figure 4.12a). Further analysis of the protein-protein interface between these monomers shows symmetrically interactions between chain A of the complex and chain B' of the neighboring molecule through three hydrogen bonds and three salt bridges (Figure 4.12b). These interface interactions are summarized in Table 4.5, 6 and 7.

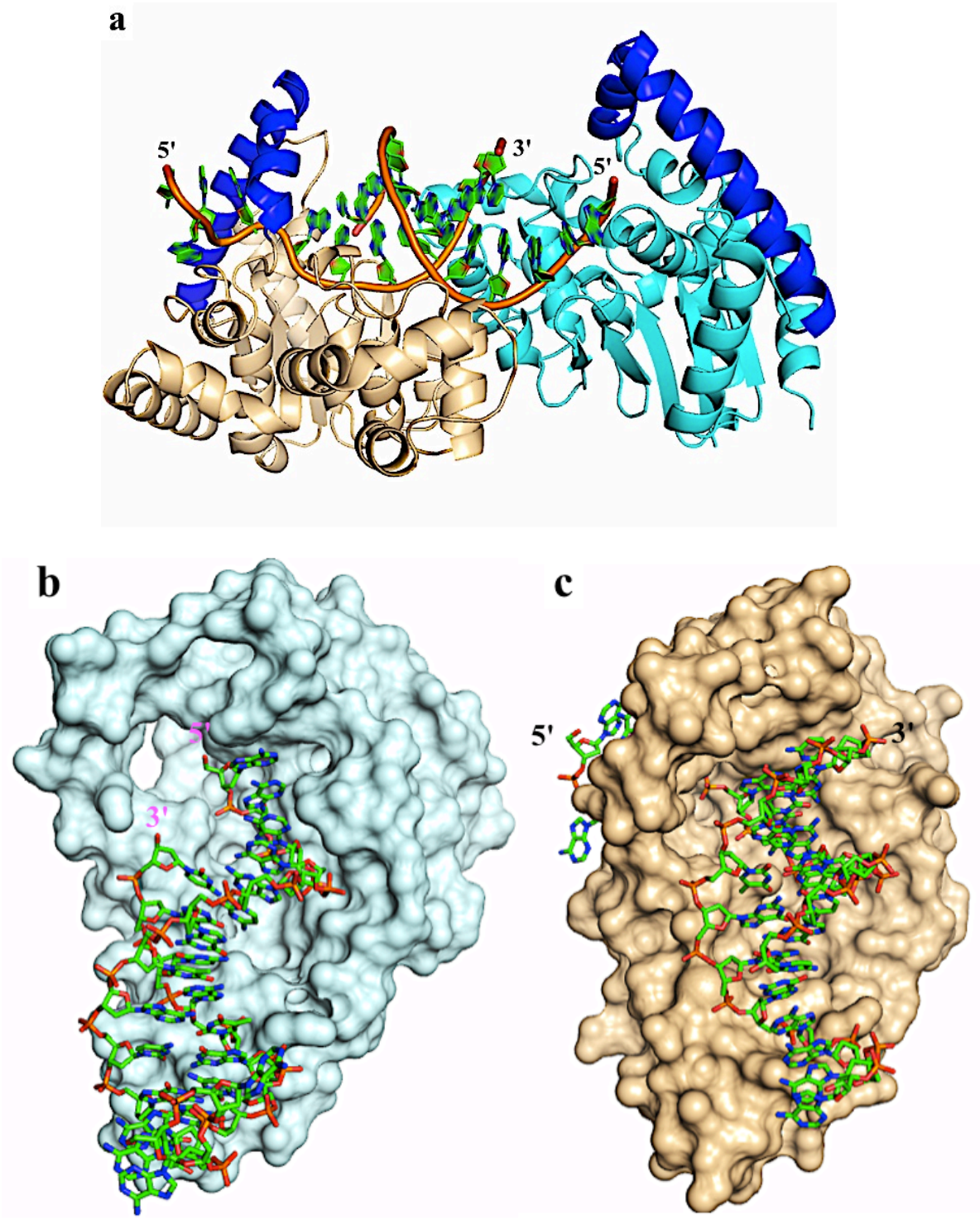



**Figure 4.48:** Diagrams for C2-T5FEN-D153K:DNA structure.
**(a)** Shows two protein molecules, C1/A (cyan) and C1/B (wheat) complexed to one palindromic DNA molecule (orange backbone and green bases). The arch-like structure is colored in blue. **(b)** and **(c)** illustrate views in Y-axis for C1/A and C1/B with the DNA substrate respectively.

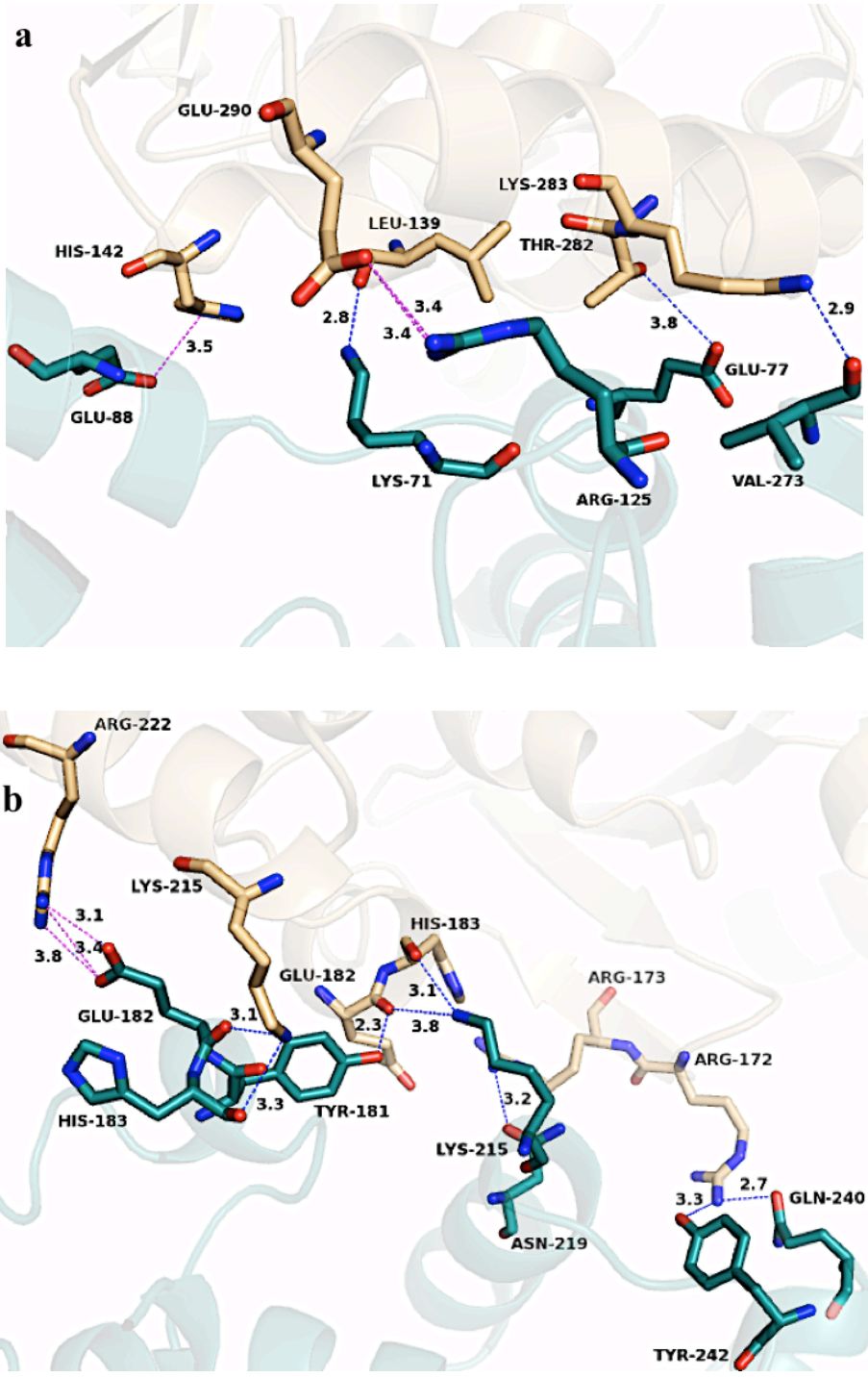



**Figure 4.49:** Protein-protein interactions in C2 structure.
**(a)** Shows C2/A (deep teal) interacts with chain B' (wheat) from the neighboring molecule. **(b)** Illustrates C2/B interactions with chain A from the same complex. These interactions made through hydrogen bonds (blue dashes) and salt bridges (magenta dashes).

**Table 4.13:** Hydrogen bonds between C2/A and the neighboring molecule chain B'

| Chain | Resi. ID | Distance Å | Chain | Resi. ID |
|---|---|---|---|---|
| **A** | Lys-71[NZ] | 2.81 | **B'** | Leu-139 [O] |
| **A** | Glu-77 [OE1] | 3.76 | **B'** | Thr-282 [OG1] |
| **A** | Val-237 [O] | 2.93 | **B'** | Lys-283 [NZ] |

**Table 4.14:** Hydrogen bonds between C2/B and C2/A in the same unit cell

| Chain | Resi. ID | Distance Å | Chain | Resi. ID |
|---|---|---|---|---|
| **B** | Arg-172 [NH2] | 3.35 | **A** | Tyr-242 [OH] |
| **B** | Arg-172 [NH2] | 2.71 | **A** | Gln-240 [OE1] |
| **B** | Arg-173 [NH1] | 3.18 | **A** | Asn-219 [OD1] |
| **B** | Lys-215 [NZ] | 3.07 | **A** | Glu-182 [O] |
| **B** | Lys-215 [NZ] | 3.32 | **A** | His-183 [O] |
| **B** | Arg-222 [NH1] | 3.43 | **A** | Glu-182 [OE2] |
| **B** | Arg-222 [NH2] | 3.09 | **A** | Glu-182 [OE1] |
| **B** | Glu-182 [O] | 3.76 | **A** | Lys-215 [NZ] |
| **B** | Glu-182 [O] | 2.26 | **A** | Tyr-181 [OH] |
| **B** | His-183 [O] | 3.11 | **A** | Lys-215 [NZ] |

**Table 4.15:** Salt bridges between C2/A and the neighboring molecule chain B' and C2/A and C2/B in the crystal.

| Chain | Resi. ID | Distance Å | Chain | Resi. ID |
|---|---|---|---|---|
| **A** | Arg-125 [NH1] | 3.35 | **B'** | Glu-290 [OE1] |
| **A** | Arg-125 [NH2] | 3.41 | **B'** | Glu-290 [OE1] |
| **A** | Glu-88 [OE2] | 3.47 | **B'** | His-142 [NE2] |
| **B** | Arg-222 [NH1] | 3.43 | **A** | Glu-182 [OE2] |
| **B** | Arg-222 [NH2] | 3.09 | **A** | Glu-182 [OE1] |
| **B** | Arg-222 [NH2] | 3.78 | **A** | Glu-182 [OE2] |

### *4.1.2.2 Protein DNA Interactions*

The DNA duplex molecule is found between the two chains A and B which interacts strongly with chain B (Figure 4.11). The double strands DNA (X and Y) are interacting with the T5FEN variant enzyme in several sites.

#### 4.1.2.2.1 Interactions with H3TH:$K^+$ Motif

C2/A interacts with internal nucleotides of the duplex from DNA strand (Y) directly through Gly-211, Gly-213, Lys-215 and Arg-216 residues and indirectly by Met-199, Val-209 and Ile-212 *via* potassium ion (Figure 4.13a). The main chain amide nitrogens of Gly-211, Gly-213, Lys-215 and Arg-216 form hydrogen bonds to the dC-9 and dA-10 phosphate oxygens. The GhG sequence is also presents here by Gly-211, Ile-212 and Gly-213 residues (Figure 4.13a). In contrast, the H3TH motif in C2/B of this structure binds to the 5ov4 DNA in the more remote part of the duplex from the active site through five residues also: Gly-211, Gly-213, Lys-215 and Arg-216. These amino acids form hydrogen bonds between their main chains amide nitrogens and the phosphate oxygens of dC-6/X and dG-5/X (Figure 4.13b). Additional interactions are formed with Met-199, Val-209 and Ile-212 and the DNA backbone indirectly *via* a potassium ion presents in two positions in chain B of this complex (Figure 4.13b).

#### 4.1.2.2.2 Potassium Binding Site ($K^+$)

A potential binding site for potassium ion was identified in the T5FEN:DNA complex in both protein chains A and B (Figure 4.13a & b and Figure 4.14a & b). The site was located within the H3TH motif in a similar position to those seen in three other FEN members whose structures have been published (Orans et al, 2011, Tsutakawa, 2011 #361, Anstey-Gilbert, 2013 #501). Potassium ions were present in the buffers used to prepare the complex at approximately 200 mM (see section 2.2.2.1 and 2.4.1) in the form of KCl. In chain A a $K^+$ ion presents and it was coordinated with dC-10/Y phosphate oxygen, two water molecules and main chain carbonyl oxygens of Met-199, Val-209 and Ile-212 (Figure 4.13a).

Similarly a potassium ion was also observed in C2/B as that seen in C2/A. In this chain, B, the $K^+$ ion occupies two positions with sharing occupancy. It occupies 40% of the first position and coordinates with dC-6/X phosphate oxygen, two waters and main chain carbonyl oxygens of Met-199, Val-209 and Ile-212 (Figure 4.13b) while it occupies 60% of the second position and coordinates by four water molecules and dC-6/X phosphate oxygens (Figure 4.13b).

#### 4.1.2.2.3 Interactions with the Hydrophobic Wedge

The 5ov4 DNA is bound along chain A and B of the T5FEN variant enzyme and extends to the active site. In C2/A it passes above the active site and stacked against the hydrophobic wedge, which is built by residues from 30-36, by the 5' overhang of strand-X (Figure 4.15a & b). The 3' end of the duplex DNA (strand-Y) in C2/A did not reach the hydrophobic wedge and as a result it did not interact with this motif in this chain. In C2/B, DNA passes through the active site and then interacts with the helix wedge by the 3' end (strand-X) of the duplex (Figure 4.16a & b). The DNA then moves toward the arch hole and the 5' end overhang with its four dAs threads through the helical arch in C2/B (Figure 4.16a). Interestingly, His-36 in C2/B is rotated totally by its main and side chains with ~ 180° from its place in chain A (Figure 4.16b).

#### 4.1.2.2.4 Intermolecular Interactions Underneath the Helical Arch

The 5' end of the DNA strand-X passes above the active site in C2/A and is coordinated by Lys-153 *via* hydrogen bonds formed to the dA-2 phosphate oxygen from one side and by Arg-86 amino groups from the other side of the dA-2 phosphate group (Figure 4.15c). The other interactions were observed between the 5' overhang DNA strand-X and Gly-154, Asp-155, Tyr-175 and His-183 residues is similar to what have been determined with C1 structure (see section 4.1.1.3.3). These amino acids interacted directly with the DNA backbone or indirectly *via* water molecules found in the same positions in chain A of both complexes, C1 and C2, which are conserved interactions in the two T5FEN-D153K:DNA structures.

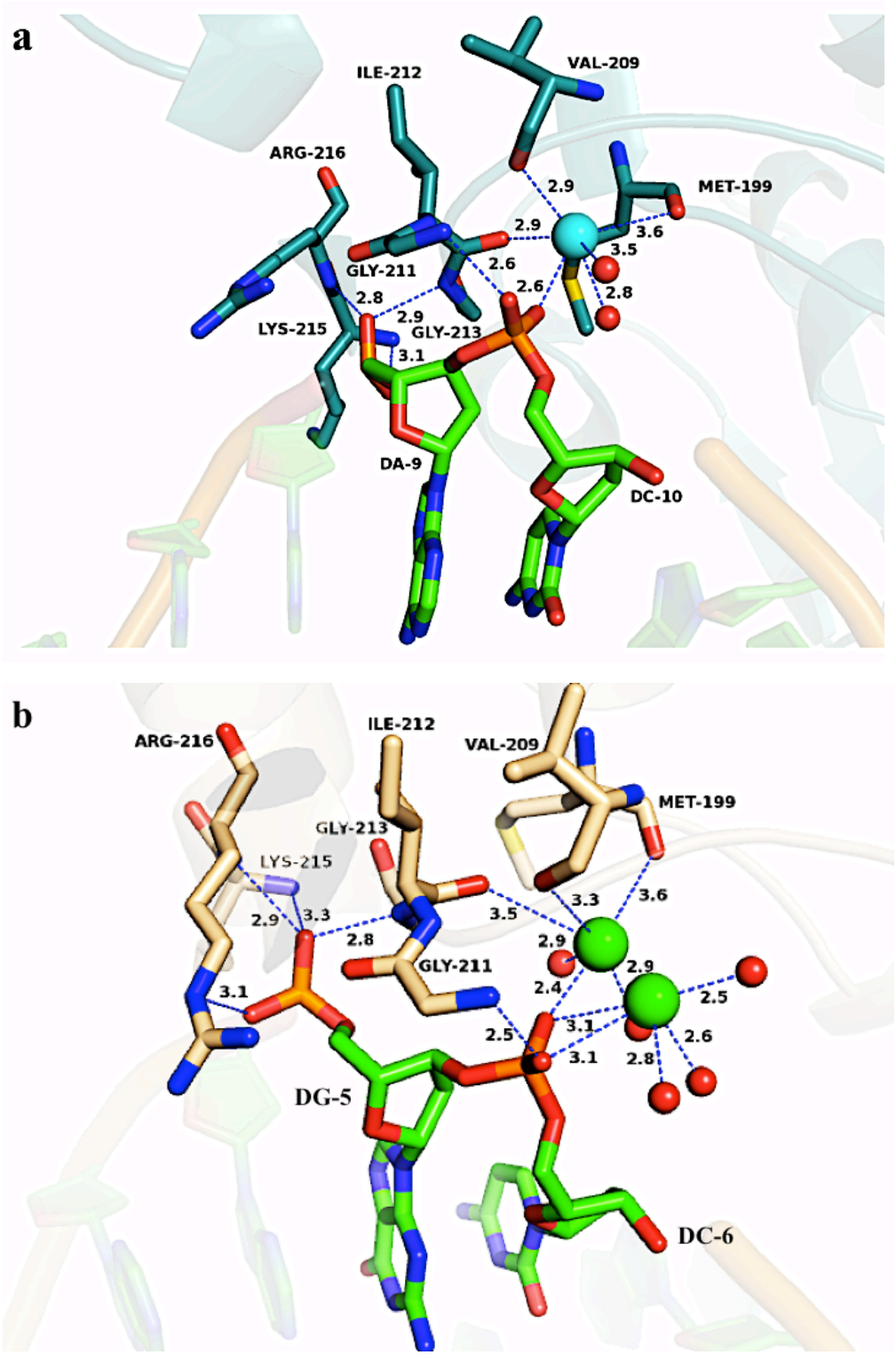



**Figure 4.50:** The C2-H3TH motif interactions with the DNA backbone.
**(a)** Shows the interactions of C2/A H3TH motif (deep teal sticks) with the DNA backbone. **(b)** Shows the same interactions in C2/B H3TH motif (wheat sticks) and DNA backbone. The potassium ions are present as cyan sphere in C2/A and green spheres in C2/B.

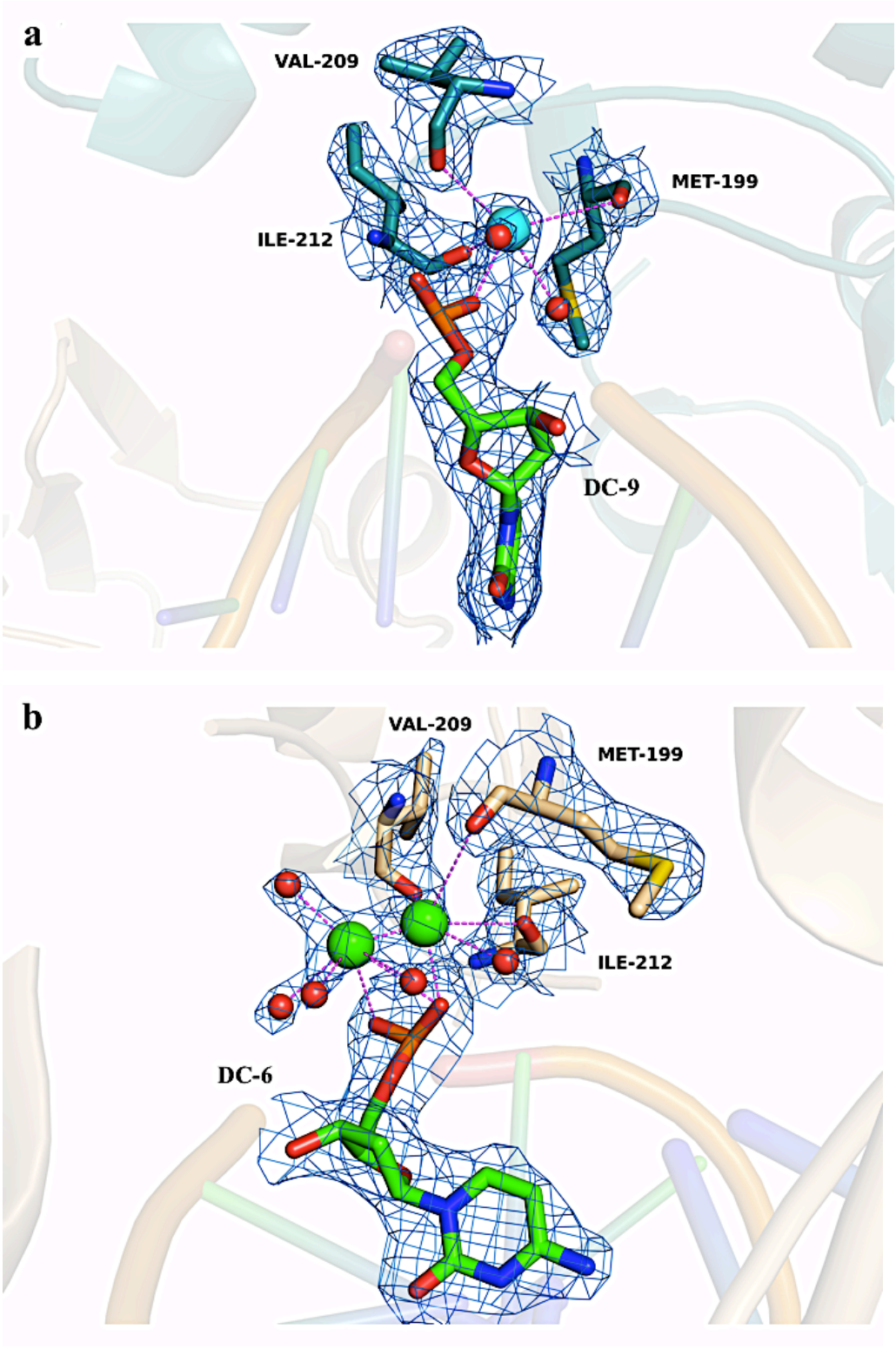



**Figure 4.51:** 2Fo-Fc electron density map for the potassium-binding site in C2 structure.
**(a)** Illustrates the $K^+$ ion in C2/A. **(b)** Shows the $K^+$ ion in C2/B in two positions.

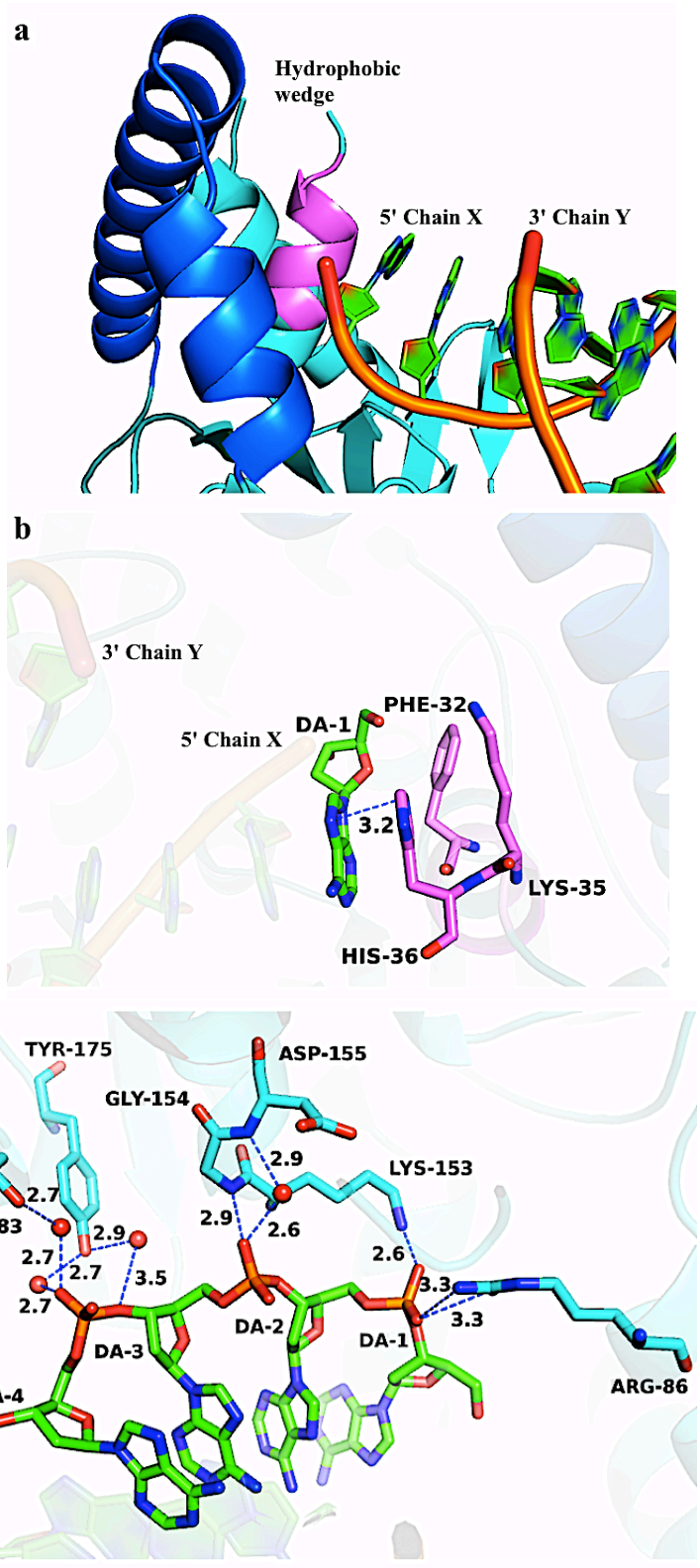



**Figure 4.52:** Diagrams for C2/A hydrophobic wedge (violet) and 5' overhang interactions.
**(a)** and **(b)** show the hydrophobic wedge in C2/A stacked against strand-X 5' overhang and its interactions with this strand respectively. **(c)** Shows the interactions between C2/A (cyan sticks) and the DNA strand-X 5' overhang.

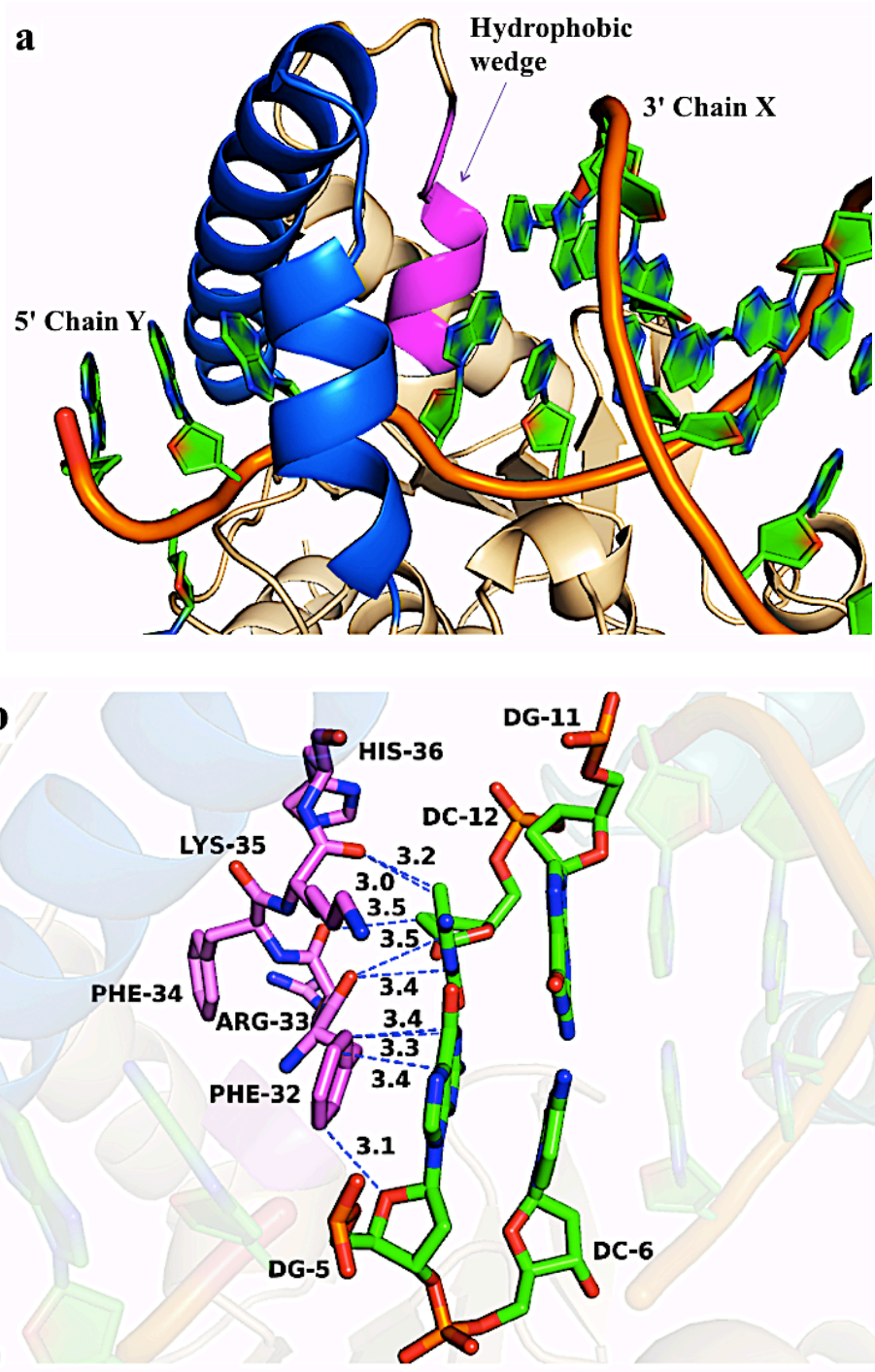



**Figure 4.53:** Diagrams illustrate the hydrophobic wedge in C2/B.
**(a)** Shows the hydrophobic wedge (violet) stacked against the DNA doublex region (orange backbone and green bases) while the 5' overhang thread through the arch. **(b)** Represents the interactions between the hydrophobic residues (violet sticks) and the DNA (orange backbone and green bases).

The 5' end overhang is directed towards the helical arch hole in chain A but it is too short to pass through it. In C2/B, the four single dAs in the 5' end overhang continue moving toward the helical arch and pass through it. These bases stack on the duplex nucleotides but are then twisted through 90º on the far side of the arch. Tyrosine-90 hangs down from the arch and is rotated by about 90º from its original position in the native T5FEN enzyme (1ut5.pdb) and coordinates the dG-5/Y phosphate oxygen *via* hydrogen bond formed by its hydroxyl group. This amino acid is close to the Phe-105 and together these residues form a gate-like structure which stacks against the 3' end (strand X) of the duplex DNA (Figure 4.17a). Arginine 86 which also hangs down from the arch interacts with the DNA substrate backbone. Bifurcated hydrogen bonds are formed between the Arg-86 guanidinium group and the DNA phosphate oxygens of the dG-5/Y and dA-3/Y (Figure 4.17b). The phosphate group in this end interacts with Gly-70 and Lys-71 which are part of the loop between $\beta$-2 and $\alpha$-3. These interactions are made through hydrogen bonds formed by main chain amide nitrogens in both amino acids; directly for Gly-70 and indirectly for Lys-71 *via* water molecule to the dA-4/Y and dA-3/Y phosphate oxygens (Figure 4.17c). Further contacts by Leu-76 side chain with dA-2/Y base *via* Van der Waals interactions have been noticed. A new distal DNA binding site was formed by these three residues which previously unrecognized.

#### *4.1.2.3 The Active Site and Metal Ions*

In the crystal structure presented here, C2/A active site is free from any metal ions. Interestingly, Lys-153 long side chain is rotated by approximately $\chi_1$ 120° (Figure 4.18a) and occupied which could be a new metal ion binding site and then interacts with the DNA single strand backbone. A hydrogen bond is formed between dA-2/X phosphate oxygen and the Lys-153 side chain. In C2/B, two magnesium ions ($Mg^{2+}$) are found in Cat1 (M1) and Cat2 (M3) separated by 10.30 Å (Figure 4.18b). These metal ions are present in the crystallization buffer components (see section 3.1.2.2). The first ion is coordinated by five water molecules and interacts with the DNA scissile phosphate dC-6/Y. These water molecules are

hydrogen bonded to four of the conserved active-site carboxyl residues: Asp-26, Asp-68, Glu-128 and Asp-130. Additionally, the mutated acidic residue D153K is ~ 4 Å from the first $Mg^{2+}$ (M1 site, Figure 4.18c). The second magnesium ion was found in the Cat2 site and interacts with the dG-7/Y DNA backbone. It was also coordinated by five water molecules, three of them being hydrogen bonded to Asp-201 (Figure 4.18b).

## 4.2 Discussion

There has been recent debate regarding the structure-specific 5'-nucleases and how they accommodate and process the 5' flap structure and other related DNA substrates (Grasby et al, 2012; Orans et al, 2011; Patel et al, 2012; Sobhy et al, 2013; Tsutakawa et al, 2011). Additionally, the role of the "helical arch"-like structure, which lies above the active site in most members of FEN family, in the 5' flap selectivity and the cleavage mechanism remains to be fully understood as to date, there are no published structures of a FEN with a full length DNA substrate. Many hypotheses and models have been suggested for the helical arch and its role in the interactions between the FEN enzymes and the DNA 5' end as described in section (1.2.1.3). Two structures of a short duplex DNA substrate possessing two 5' overhangs in complex with a catalytically inactive T5FEN (D153K variant) were obtained. The structures show for the first time in a FEN family members the looping-up of the single stranded DNA region of a 5' flap before threading through the helical arch. Two chains of the T5FEN-D153K were observed in the unit cell of these structures connected by the palindromic duplex DNA substrate (5ov4). In order to distinguish between these two complexes, the first complex called "C1-T5FEN-D153K:DNA" which shows the earlier process of threading model while the second one named "C2-T5FEN-D153K:DNA" with complete threading of the 5' overhang through the arch. The crystals of these structures were grown in quite different conditions (0.2 M NaCl, 0.1 M $(CH3)_2AsO_2Na$ pH 6.5, 2 M $(NH4)_2SO_4$ for C1 and (0.2 M $MgCl_2$, 0.1 M bis-Tris pH 5.5, 22 % PEG 3350) for C2 structures but despite this had the same space group.

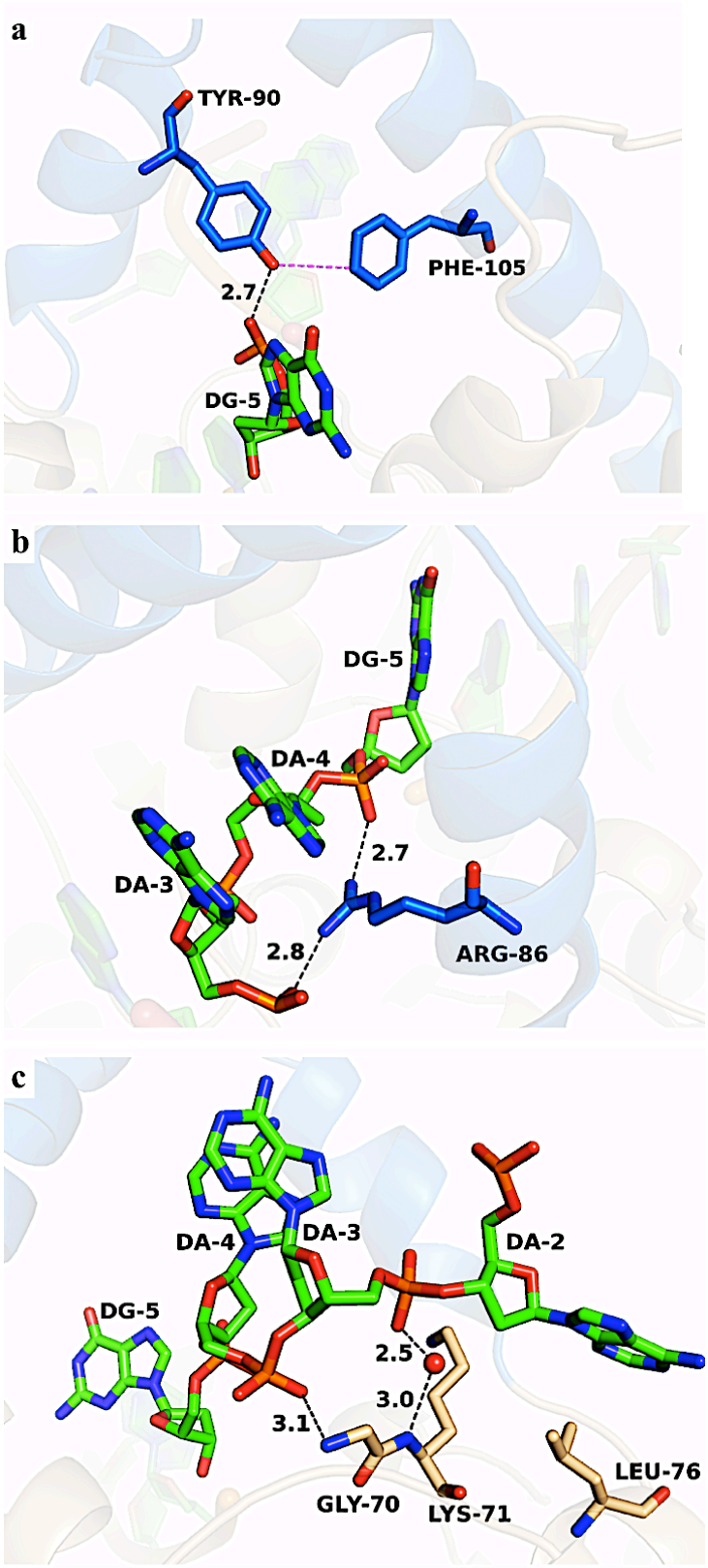



**Figure 4.54:** Interactions underneath the helical arch in C2/B.
**(a)** Shows the gate-like structure formed by Tyr-90 and Phe-105 residues (blue sticks) and the interactions between Tyr-90 and dG-5/Y phosphate oxygen. **(b)** Shows Arg-86 (blue sticks) from the arch (blue) interacting with the DNA backbone. **(c)** Illustrates the interactions between the distal DNA binding site composed of Gly-70, Lys-71 and Leu-76 (wheat sticks) and the DNA backbone.

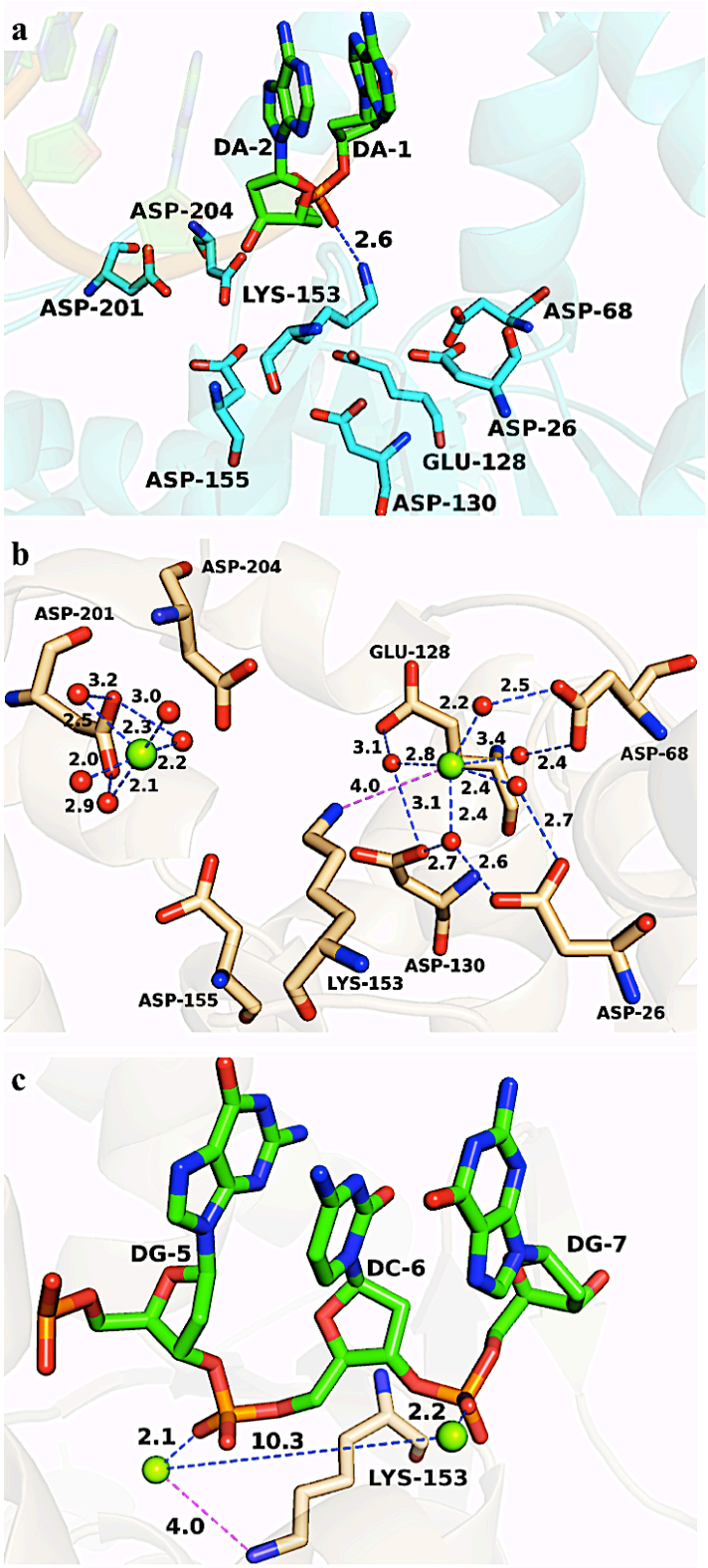



**Figure 4.55:** The active site residues and metal ions in C2-T5FEN-D153K:DNA. **(a)** C2/A active site residues (cyan sticks) and Lys-153 interacts with the DNA backbone. **(b)** Shows the conserved active site carboxylic amino acids (wheat sticks) in C2/B coordinated two $Mg^{2+}$ ions (green spheres) in Cat1 (M1) and Cat2 (M3) sites. **(c)** Illustrates the interactions in C2/B between the $Mg^{2+}$ ions and the DNA strand-Y backbone.

### 4.2.1 The Threading Process and 5' overhang Interactions

In the two complexes, the DNA substrate, 5ov4, which has an 8 bp duplex with two 5' overhangs consisting of 4-dA residues is placed between the two protein chains A and B and its strands called Y and X. The first interactions between the DNA substrate and T5FEN protein is assumed to be with chain B and the DNA duplex junction which is recognized firstly by the protein (Xu et al, 2001) and interacts with the hydrophobic wedge. The protein then makes intensive interactions with the H3TH:$K^+$ motif from the other duplex end (3' end/Y) in both complexes, C1 and C2. The 5'-ssDNA initially looped-up and then interacts with the arch residues as seen in C1 of T5FEN-D153K:DNA before start threading through the hole. The 5' overhang then pouched inside the arch electrostatically using the arch positive charged residues. However, this structure shows at least two nucleotides threaded through the arch hole while the other two nucleotides, dA-1 and dA-2, may have been removed by some contaminant or are simply disordered. Mass spectrometry of the nucleic acids recovered from crystals could be used to detect any degradation of the nucleotides. The continuation of the threading process can be obtained in C2 in which the DNA inserted in to the arch chamber makes a little movement by ~ 3.5 Å in the duplex region that interacted with the hydrophobic wedge in C2 in respect to the same area in C1. Due to this movement the DNA duplex region stacked with His-36 in the helix wedge which case a rotation by ~ 180 ° for the whole residues to open the way for the DNA substrate as seen in Figure 4.16b. The 5' overhang single strand threads through the hole in the arch and another gate-like structure is stacked against the 3' end/X of the duplex DNA that is formed by Phe-105 and Tyr-90 which hangs down from the arch and rotates by ~ 90° then interacts with dG5/Y phosphate group.

The phosphate groups of this 5' overhang are coordinated by Arg-86 side chain which is conserved along the FEN family members. The structures presented here, C1 and C2, show positively charged Arg-86 stabilizes the ssDNA before and during threading process by forming hydrogen bonds with the DNA backbone phosphate oxygens.

C1-T5FEN:DNA Arg-86 interacts with dC-6/Y backbone indirectly *via* water molecules while in C2 it forms hydrogen bonds to the phosphate oxygen of the first nucleotide in the duplex (dG-5/Y) and the third nucleotide in the ssDNA (dA-3/Y) which give an indication that this residue may be essential for the transmission of the single strand through the arch. These positively charged amino acids supposed to speed up the scissile phosphate hydrolysis by forming electrostatic field (Sengerova et al, 2010). During threading process other interactions with the DNA backbone are observed through trans-arch/distal phosphate site which is a DNA new binding site composed of Gly-70 and Lys-71 and recognized for the first time in T5FEN enzyme. In order to prevent sliding of the 5' overhang from the arch hole after threading, the dA-2/Y single-base is rotated by ~ 180° to form the ratchet/barb-like structure in the far side of the arch (Figure 4.17c).

The C1 and C2 structures strongly support the threading model through the arch-like feature that has been reported by a number of laboratories (Ceska et al, 1996; Lyamichev et al, 1993; Patel et al, 2012; Tsutakawa et al, 2011) as a required mechanism to process 5' flap ssDNA and they are clearly in disagreement with the tracking hypothesis (Gloor et al, 2010) and clamping model (Orans et al, 2011). In addition, the threading model can deal with flaps less than 5 nt as it can be seen from these two structures (4 nt) which is also in conflict with the hypothesis proposed by Gloor et al (2010). Gloor and his team measured the relationship between the length of the 5' flap and its ability to bind hFEN1 using electrophoretic mobility binding assay (EMSA). The result of this experiment suggested that the long 5' flaps bind hFEN1 with more affinity comparing to the 5' flaps less than 5 nt which is in disagreement with T5FEN:DNA complex explained here. However, the two enzymes though related are different that Gloor and co-worker were used eukaryotic FEN while T5FEN:DNA is prokaryotic one.

The observed structures, both chains A and B displaying ordered arches, suggests that the T5FEN helical arch is unlikely to be subjected to the disorder-to-order transmission mechanism proposed by Tsutakawa and coworkers (Grasby et al, 2012; Patel et al, 2012; Tsutakawa et al, 2011)

which appears to be a unique mechanism for FEN superfamily members. Several T5FEN structural studies reported an ordered arch in DNA-free enzyme (Ceska et al, 1996; Feng et al, 2004). Also a totally disordered arch was found in the T5FEN-K83A variant free of DNA (Garforth et al, 1999) while both ordered and distorted arches have been observed by Flemming (2011) for the two copies of native T5FEN molecules in the same unit cell under the identical crystallization conditions (see section 1.2.1.4 and Figure 1.18).

### 4.2.2 Chain A and DNA Interactions

The molecules represented by C1/A and C2/A of T5FEN-D153K:DNA structures are also recognized the DNA from the duple junction region but bound less extensively with the opposite end of the substrate (relative to chain B). In this case, the crystallized duplex DNA is not long enough to interact with chain A and chain B in a symmetrical manner. It is assumed that the protein:protein packing in the asymmetric unit does not allow the 5ov4 substrate DNA to bind equivalently to each end. Additionally, this interaction between the 5' end of DNA chain X and chain A imposes the observed ordered structure on this single stranded end of the substrate molecule. In C1-T5FEN:DNA the positive density that was noticed between the first two deoxyadenosines of the 5' overhang cannot be occupied with an identifiable molecule because the bond lengths and the geometry do not agree with any metal ions, waters or solvent molecules that are presented in the protein preparation and crystallisation buffers. Interestingly, the negative density that presents around the DNA backbone between these two bases may indicate missing of some atoms as a result of an exonuclease activity that was suggested to work through different mechanism from that for endonuclease activity (Garforth et al, 1999). This might be explained by very slow enzymatic cleavage carried out by the T5FEN-D153K, however, previous results indicated that this protein lacked any detectable enzymatic activity (Feng, 2002; Zhang, 2012).

### 4.2.3 The Potassium Binding Site

T5FEN:DNA structures explained above revealed a potassium binding site previously unobserved which could explain an earlier report that the enzyme activity is increased 1.4 fold in the presence of 75 mM KCl (Sayers & Eckstein, 1990). A binding site for the potassium ion was present in hPol $\beta$ and is thought to facilitate movement of the protein along the DNA substrate backbone (Pelletier & Sawaya, 1996) which could be the same for the other FENs. In C2-T5FEN:DNA presented here, chain B has $K^+$ ion in two positions that could make the protein movement easier during threading process. This site has been observed in DNA complexes with other FEN family members such as hFEN (Tsutakawa et al, 2011), hEXO1 (Orans et al, 2011) and *E.coli* ExoIX (Anstey-Gilbert et al, 2013). However, these published structures contain enzyme-product complexes rather than complete substrates. Additionally, Hemsworth (2009) has been observed another metal ion binding site within the H3TH motif of *E. coli* ExoIX which was referred to as a $Mg^{2+}$ or $Ca^{2+}$ binding site while it can be second position for $K^+$ ion as what is noticed in C2/B structure.

### 4.2.4 The Active Site and Metal Ions Interactions

Early crystal structures of native T5FEN, T4RNase H and *T. aquaticus* homologues (1ut5.pdb, 1tfr.pdb, 1taq.pdb) suggested that these enzymes possess two divalent-metal-ion binding sites. However, in recently published structures of hEXO1, hFEN1 and ExoIX, their similar Cat1 sites were found to contain two metal sites (M1 and M2) as described in section 1.2.1.4. Interestingly, the metal ions are absent in C1-T5FEN:DNA in spite of the presence of 300 mM sodium solutions in the crystallization and protein buffers which could be related to the presence of $(NH_4)_2SO_4$ at 2 M concentration in the crystallization buffer. The M2 position of the Cat1 binding site in C2/B is occupied by the mutated residue Lys-153 which explains the missing metal ion in this site of the structure and confirms the mutation's purpose. The scissile phosphate in this complex (dC-6/Y) is positioned in the active site and is coordinated by the first magnesium ion. The two metal ions ($Mg^{2+}$) at Cat1 and Cat2 form ionic bonds to the scissile

phosphate oxygen and also to the dG-7/Y phosphate oxygen respectively. If M1 and M2 in Cat1 are occupied by divalent metal ions the cleavage reaction can occur in the wild type enzymes and if this is not the situation, the metal ion in Cat2 (M3) could move to the M2 at Cat1 site and engage in the DNA backbone reaction. The movement of the metal ions within the active site in the metallonucleases is still as a point of debate (Dupureur, 2008).

In order to locate the scissile phosphate between the two metal ions in the active site of hFEN:DNA complex (Tsutakawa et al, 2011) and hEXO1:DNA (Orans et al, 2011) two bases of substrate would need to unpair and the importance of this step in the hydrolysis mechanism is unknown so far. The requirement of the di-nucleotide unpairing mechanism for hydrolysis of the scissile phosphate was not apparent in these T5FEN:DNA complexes which are catalytically inactive. This mechanism is likely to be a later step of the threading process depending on the nature of the substrate (nicked, pseudo-Y structures, flaps or overhangs).

### 4.2.5 Ammonium Sulphate in C1 Structure

Ammonium sulphate is one of the polyatomic ions which composed of more than two atoms. The crystal of C1 was grown in a condition has 2 M of $(NH4)_2SO_4$ and consequently the $NH4^+$ and $(SO_4)^{2-}$ are expected to present in the structure. A thirteen copy of $(SO_4)^{2-}$ ion are identified in this complex and they occupy similar places in both chains A and B as described above. The ammonium ions can form four hydrogen bonds with the surrounding atoms and 109.5° for H-N-H.

The next sections of this chapter deal with the analysis of these two complex structures through comparisons with native T5FEN in a complex with magnesium ions. The variations and similarities between chain A and B in each structure of C1 and C2 of T5FEN-D153K:DNA are determined by superimposing both chains together and then each chain in a complex with the DNA molecule with T5FEN:$Mg^{2+}$ structure. The superposition will be

obtained using the secondary structure matching in the LSQKAB program (CCP4, 1994). Comparisons between C2 and other FEN members in complexes with DNA substrates will be done in the next chapter. T4RNase H in a complex with pseudo-Y DNA will used as an example of bacteriophage FEN complex while hFEN:DNA and hEXO1:DNA will mentioned as examples for FEN superfamily complexes.

## 4.3 Comparisons

### 4.3.1 Comparison of T5FEN-D153K:DNA Chain A with Chain B

Superimposition of the two chains A and B in each complex, C1 and C2, of T5FEN:DNA shows a very good overlay of these chains on one another (Figure 4.19a-d) with an RMSD of 0.606 Å and 1.089 Å for C1 and C2 structures respectively over a core of 267 amino acid C$\alpha$ positions. The superposition analysis shows a high rate concordance between the two chains in individual complex but there are differences in some parts of them. The junction loop which connects $\alpha$1 helix to $\alpha$2 helix is partially built in C1/A and C1/B and in C2/A. The main clear difference is in the top part of the helical arch (a.a. 90-100) which appears in chain B from each complex to move down in the direction of the DNA (Figure 4.19a & c). The residues in this part have shifts ranging from 0.5 Å to 1.8 Å in C1 and from 1.55 Å to 3.63 Å in C2 chains A and B which could be a result for threading process. The second region of the residues movement has been noticed in C1 in the H3TH motif which are ranging from 1.0 Å as a minimum value to 2.0 Å as a maximum value for Gly-203 from its position in chain A (Figure 4.19b). Another clear difference between the C2 two chains A and B is in the orientation of the main chain and/or side chain in some residues in chain B from their position in chain A as seen for His-36 and Phe-169 (Figure 4.19e). Also some shifts are notable in some arch residues such as Arg-86, Tyr-90, Phe-105 and the trans-arch/distal DNA binding site (Gly-70 and Lys-71) as seen in Figure 4.19e.

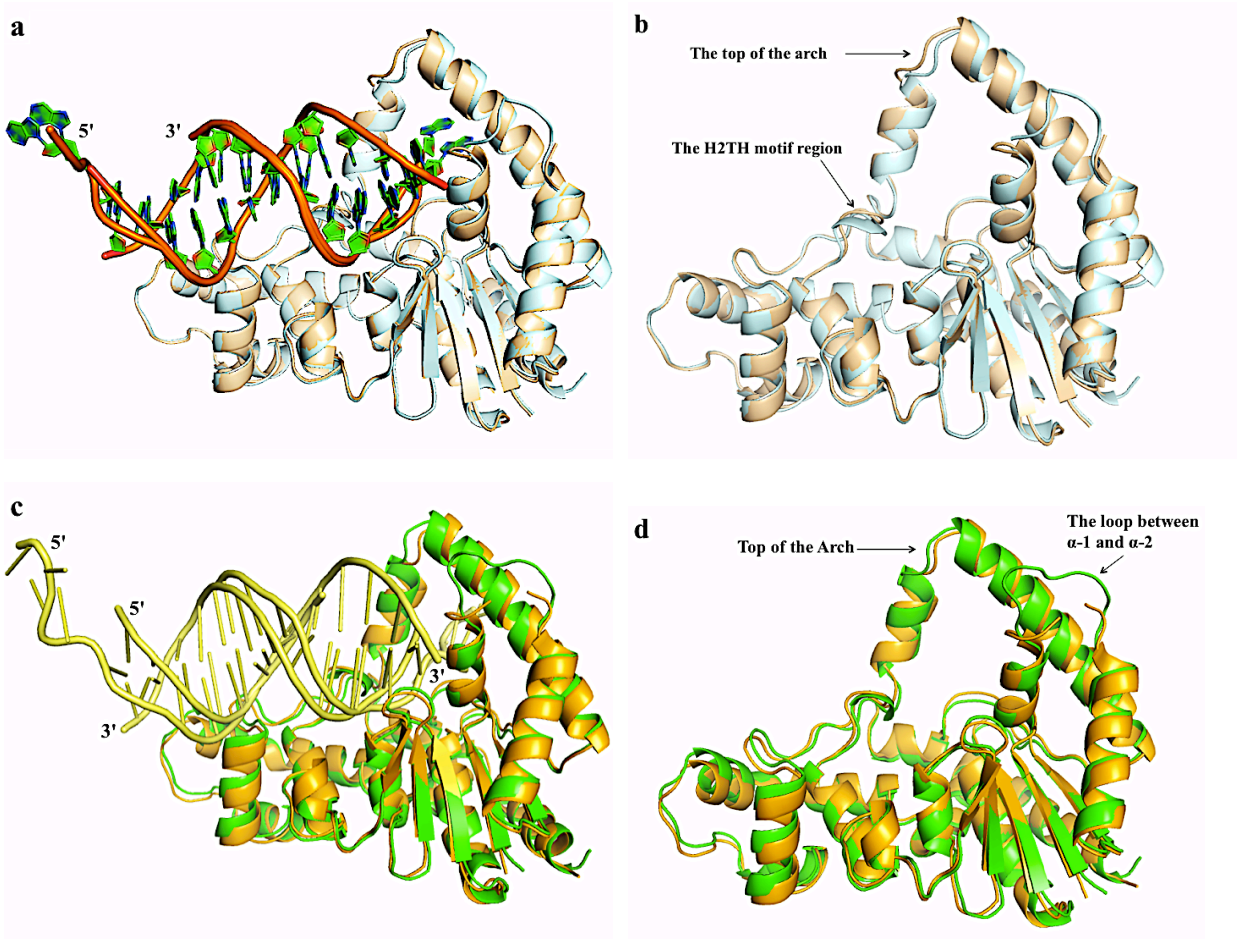
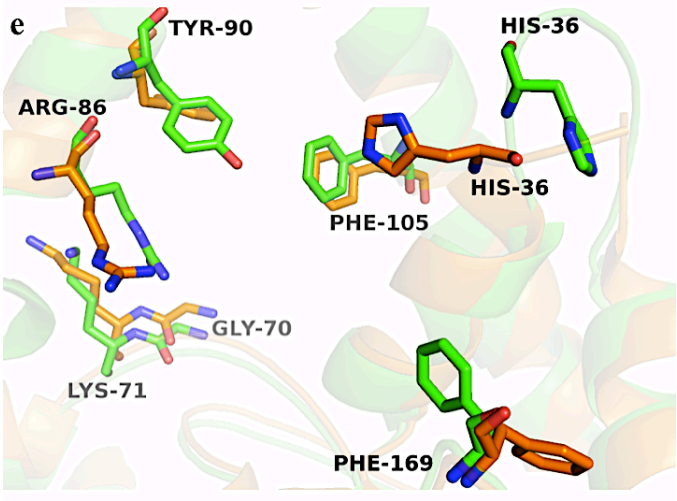



**Figure 4.56:** Superimposition of the two protein chains in the asymmetric unit for C1 and C2 of T5FEN-D153K:DNA.
**(a)** and **(b)** show superposition of C1/A (pale cyan) and C1/B (wheat) with DNA and without DNA respectively. **(c)** and **(d)** show superposition of C2/A (orange) and C2/B (green) with and without DNA respectively. **(e)** Shows the conformational changes between some residues in C2/A and C2/B.



#### *4.3.1.1 Superimposed Analysis of the Active Sites*

The residues in the two active sites overlay well on each other in C$\alpha$ positions in both complexes C1 and C2 with small location movements by about 1.0 Å for Asp-201 and Asp-204 main and side chains in C1/B from their positions in C1/A. The mutated Lys-153 presents in both active sites in the same place but the side chain NZ in chain B rotates by a 20° from its position in chain A (Figure 4.20a). As mentioned above the active sites in both C1/A and C1/B are metal ion-free. In C2 active site residues the maximum shift is 1.29 Å for Asp-155 in C2/B from its position in C2/A (Figure 4.20b). Interestingly, the mutated residue Lys-153 occupies M2 in Cat1 site of C2/B while its side chain is rotated by about $\chi_1$ 120° in C2/A (Figure 4.20b) to occupy which could be a new metal ion-binding site in the Cat1. A magnesium ion ($Mg^{2+}$) is present in both Cat1 and Cat2 sites in C2/B which interact directly with the DNA backbone while these metal ions are absent in the C2/A active site.

#### *4.3.1.2. Potassium Binding Site Superposition*

In C1 structure a $K^+$ ion is found in its binding site in C1/B but it is not observed in C1/A as described prevuosly. The potassium binding sites in both C2/A and C2/B superimpose well on each other (Figure 4.20c). In C2/B the $K^+$ ion has two places ~2.2 Å apart from each other. These $K^+$ ions coordinated with the main chain carbonyl oxygen of Met-199, Val-209 and Ile-212 residues and they are about 1.5 Å from each other (relative to the $K^+$ first position in C2/B).

### 4.3.2 Comparison of C1 and C2 on T5FEN:$Mg^{2+}$ Structure

The T5FEN:$Mg^{2+}$ complex is a crystal structure for the T5FEN wild type complexed with five magnesium ions which was determined by Flemming (2011) at a resolution of 1.5 Å. The structure was solved by molecular replacement with two molecules in the asymmetric unit.

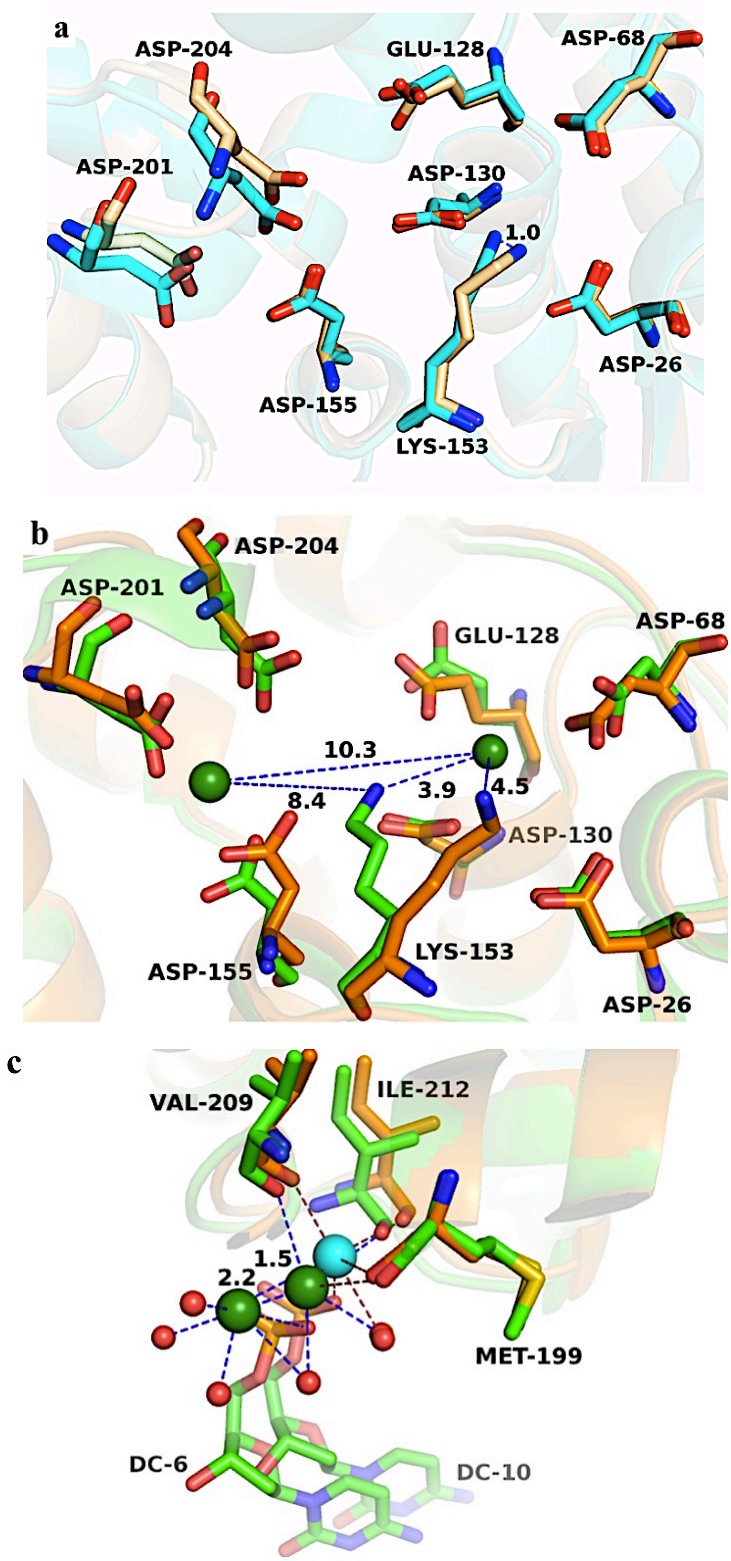



**Figure 4.57:** Superimposition of the active site residues and $K^+$ binding site in the two chains, A and B of C1 and C2 structures.
**(a)** Superposition of C1/A active site (cyan) on C1/B (wheat). **(b)** Shows C2/A active site residues (orange sticks) superposed on C2/B active site (green sticks) and the two $Mg^{2+}$ ions (green spheres) are present. **(c)** Shows the potassium ions in C2/A (cyan sphere) superimposed on the $K^+$ (green spheres) of C2/B.

One of these chains, A, has two $Mg^{2+}$ ions in its Cat1 site while the other chain, B, has three $Mg^{2+}$ ions two of them occupy Cat1 and Cat2 sites whilst the third ion located outside the active site close to the disordered arch. Superimposition of chain A on B showed four $Mg^{2+}$ ions in the superimposed active site. Two ions were found in M1 site and they occupied identical positions, one $Mg^{2+}$ within M2 and the other ion, M3, was located in the Cat2 site. In the comparison of T5FEN:DNA complexes with Flemming's native complex, the superposed chain A with the three $Mg^{2+}$ structure will be used.

#### *4.3.2.1 Superposition of C1 and C2 chains A on the Native Structure*

##### 4.3.2.1.1 Overall Superposed Structure

The two chains C1/A and C2/A overlay very well on T5FEN native (Figure 4.21a & b). The three chains together have a RMSD of 0.7 Å C$\alpha$s in an alignment length of 246 amino acids. The similarity between these chains is high and small shifts can be notable in the helical arch top area between them (Figure 4.21b). Some residues have been investigated such as Arg-86, His-36, Arg-33, Arg-125 and Phe-169. Arg-86 has the same conformation in all of these chains while Arg-33 side chain rotates by about $\chi_1$ 90° in C1/A comparing to the native one and C2/A (Figure 4.21c) which can be also explained by the presence of sulphate ion in Arg-33 position. His-36 is disordered in the native structure and C1/A while it presents in C2/A in the original place of T5FEN (1EXN, Figure 4.21c). Arg-125 and Phe-169 have similar conformations in C1/A and C2/A whilst in the native structure its side chain is rotated by $\chi_3$ 130° to have different position (Figure 4.21d). The 5ov4 DNA substrates overlay well together with some shifts in some region due to the DNA flexibility (Figure 4.22) In addition, the dA-1/X of C1/A 5ov4 is shifted by ~ 4 Å comparing to the one in the C2/A 5ov4. The 5ov4 DNA in both structures C1 and C2 is fairly overlay and the looping-up is clearly observed in the C1 overhang comparing to C2 (Figure 4.22). Additionally, the looping-up mark can be noticed in the 5' overhang of the C2/Y (Figure 4.22).

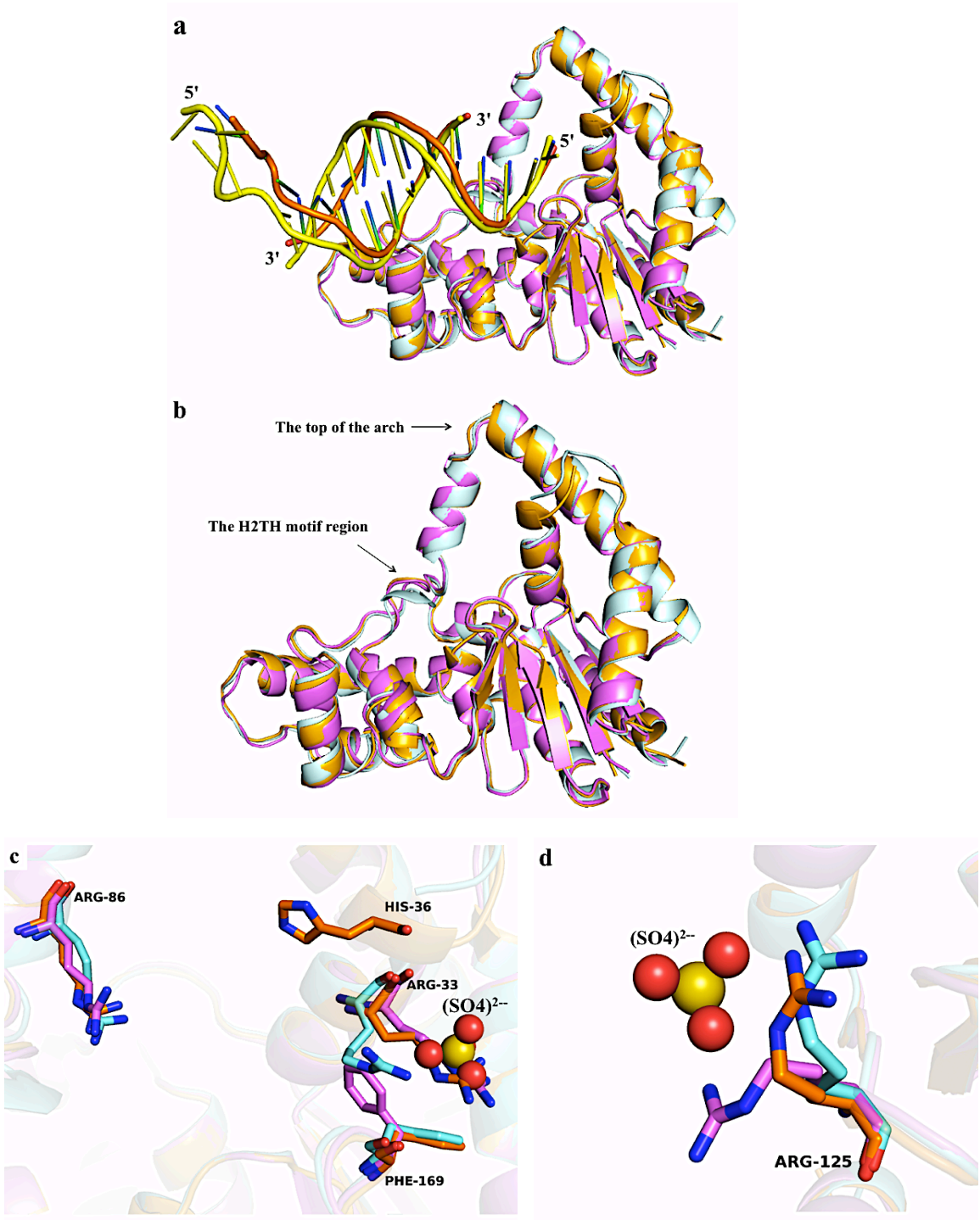



**Figure 4.58:** Superposition of C1/A and C2/A of T5FEN-D153K:DNA on T5FEN native structure.
**(a)** Shows the three protein chains overlay well on each other, C1/A (cyan), C2/A (orange) and T5FEN native in violet. **(b)** Shows the same three chains without DNA substrate and illustrates the differences between them. **(c)** Shows four amino acids, Agg-86, His-36, Arg-33 and Phe-169 from the three superimposed structures. **(d)** Shows Arg-125 residue from the three chains which has a different conformation in the native structure (violet stick). It can be seen the presence of one of the $(SO4)^{2-}$ ions close to Arg-125 location.

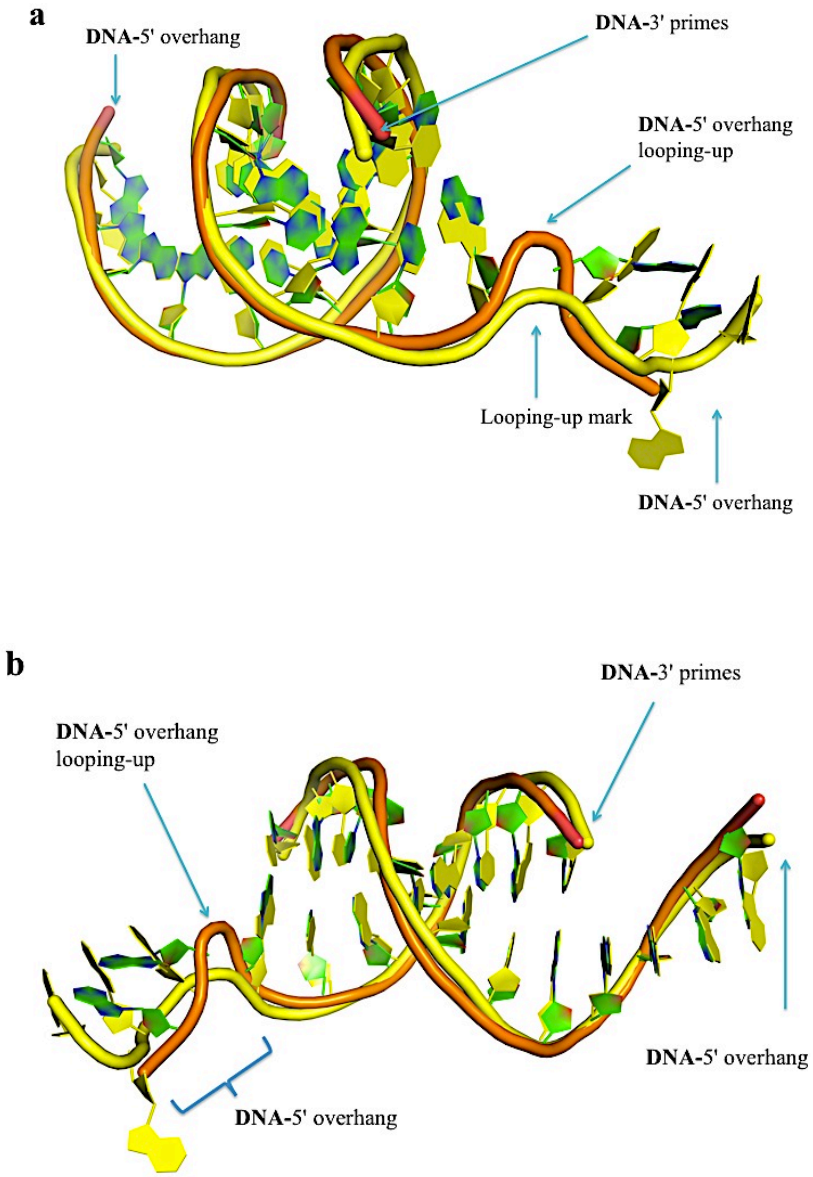



**Figure 4.59:** Superposition of DNA 5ov4 of C1structure (orange backbone and green bases) on C2 structure (yello).
**(b)** Rotation of **(a)** by 180°.

#### 4.3.2.1.2 Superposition of the Active Site Residues

When the active site residues of C1/A and C2/A superimposed on T5FEN native active site residues they overlay well in C$\alpha$ positions (Figure 4.23a). Some location movements are noticed for Asp-201 and Asp-204 in C1/A structure comparing to the other structures, the native and C2/A (Figure 4.23a). In this superposed active site three magnesium ions are present which belongs to T5FEN native structure and occupy M1 and M2 in Cat1 and M3 in Cat2 site while C1/A and C2/A do not have any metal ions (Figure 4.23a). Additionally, the mutated residue, D153K, in this superimposed active site has similar conformations. One of them is found in C1/A while the other one belongs to C2/A and its side chain rotates by $\chi_4$ 180° comparing to C1/A-Lys153 and became 0.9 Å apart from the $Mg^{2+}$ ion in M1 of Cat1 site (Figure 4.23a). The NZ of these two positively charged residues is 3.7 Å from each other (Figure 4.23a).

#### 4.3.2.1.3 Potassium Binding Site

A potassium binding site is determined for C2/A (see section 4.1.2.2.2) while for the native enzyme and C1/A it is not noticed. Superposition of these three chains together in potassium binding site area in C$\alpha$s shows the same arrangement of the amino acids which coordinate this ion in all of them (Figure 4.23b).

### *4.3.2.2 Superposition of C1 and C2 chains B on the Native Structure*

#### 4.3.2.2.1 Overall Superposed Structure

The two chains B from C1 and C2 of T5FEN-D153K in complex with 5ov4 and T5FEN native are superimposed together and they overlay well on one other (Figure 4.24). They have a RMSD of 1.2 Å in C$\alpha$ positions over 257 residues with some location movements. These movements are noticed in the arch motif, helices ($\alpha$-4 & $\alpha$-5) and the top regions of C1/B and C2/B comparing to the native chain (Figure 4.24b & c). Addition shifts appear in the H3TH motif second helix, $\alpha$-10 and extend to $\alpha$-11, $\alpha$-12 and the flexible loops between them (Figure 4.24b & c).

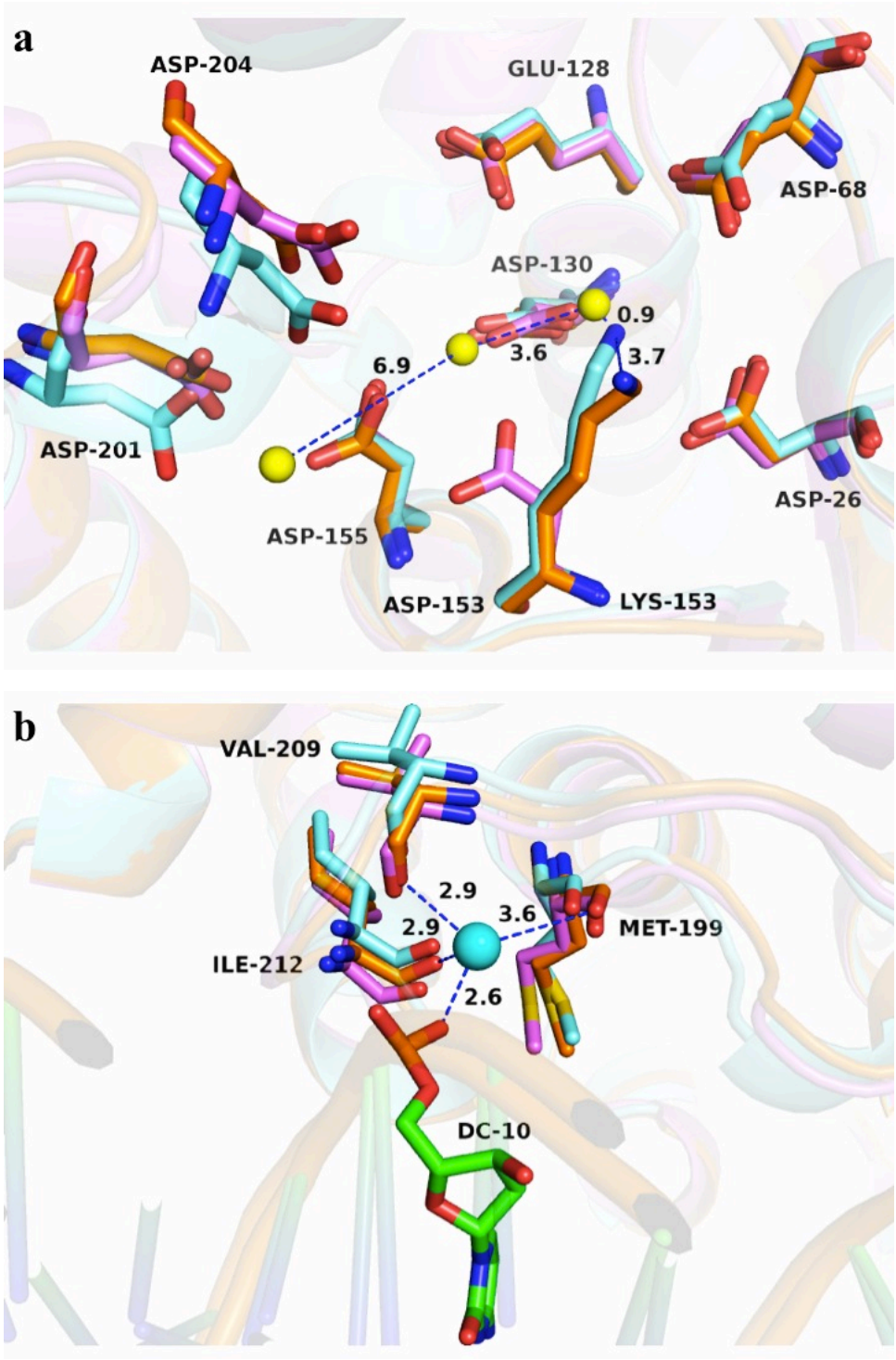



**Figure 4.60:** Superposition of the active site residues and $K^+$ site of C1/A and C2/A on T5FEN native.
**(a)** Active site residues of C1/A (cyan sticks) and C2/A (orang sticks) on T5FEN native (violet sticks). Three magnesium ions are present in the superposed active site (yellow spheres) belongs to the native structure while the other two chains A active sites do not have any metal ions. **(b)** Superposed of C1/A and native potassium ion binding site on C2/A $K^+$ ion (cyan spheres). The potassium ion presents in its site in C2/A only and the geometry of coordinations have the same arrangement.

Some shifts in some conserved and important amino acids have been noticed between these three protein chains. Arg-86 which interacts with the DNA backbone has the same conformation in T5FEN native and C1/B while its side chain rotates by $\chi_1$ 120° in C2/B (Figure 4.25a). In addition, His-36 which is disordered in the native structure and has the original conformation in C1/B as that seen in T5FEN (1EXN) but is completely rotated by 180° in C2/B (Figure 4.25a). Moreover, Arg-33, Arg-125 and Phe-169 (Figure 4.24a & b) are occupied similar places in T5FEN native and C2/B while their side chains are rotated by $\chi_1$ 130°, $\chi_2$ 90° and $\chi_1$ 90° respectively in C1/B. Arg-33 side chain rotation can be refereed to the presence of sulphate ion in the residue position comparing to the other chains (native and C2/B).

#### 4.3.2.2.2 Superposition of the Active Site Residues

The active site residues overlay well on one other in all of the three chains (Figure 4.26a). Five magnesium ions are present in the superimposed active site, three of them belongs to T5FEN native structure and the other two belongs to C2/B while the C1/B active site is metal ion-free. Furthermore, the mutated Lys-153 presents in the superimposed active site in two positions that are ~ 5 Å apart from each other. One of them for C1/B while the other one for C2/B which occupies M2 of Cat1 site. C1/B-Lys-153 side chain rotates by $\chi_1$ 120° (relative to C2/B-Lys153) to occupy a place close to M1 site. This residue is about 1.8 Å and 2.4 Å from the $Mg^{2+}$ ions in M1 of T5FEN native and C2/B respectively (Figure 4.26a).

#### 4.3.2.2.3 Potassium Binding Site

A binding site for the $K^+$ ion is observed in C1/B and C2/B but not for T5FEN native in a complex with $Mg^{2+}$ ions. Superposition of C1/B on C2/B shows well overlay of the $K^+$ ion coordination residues and DNA base as seen in Figure 4.26b. As explained previously in section 4.1.2.2.2 the potassium ion has two positions in C2/B and the $K^+$ ion of C1/B is too close from the first position which is about 1.1 Å while it is about 2.9 Å from the $K^+$ ion in the second binding site of C2/B (Figure 4.26b).

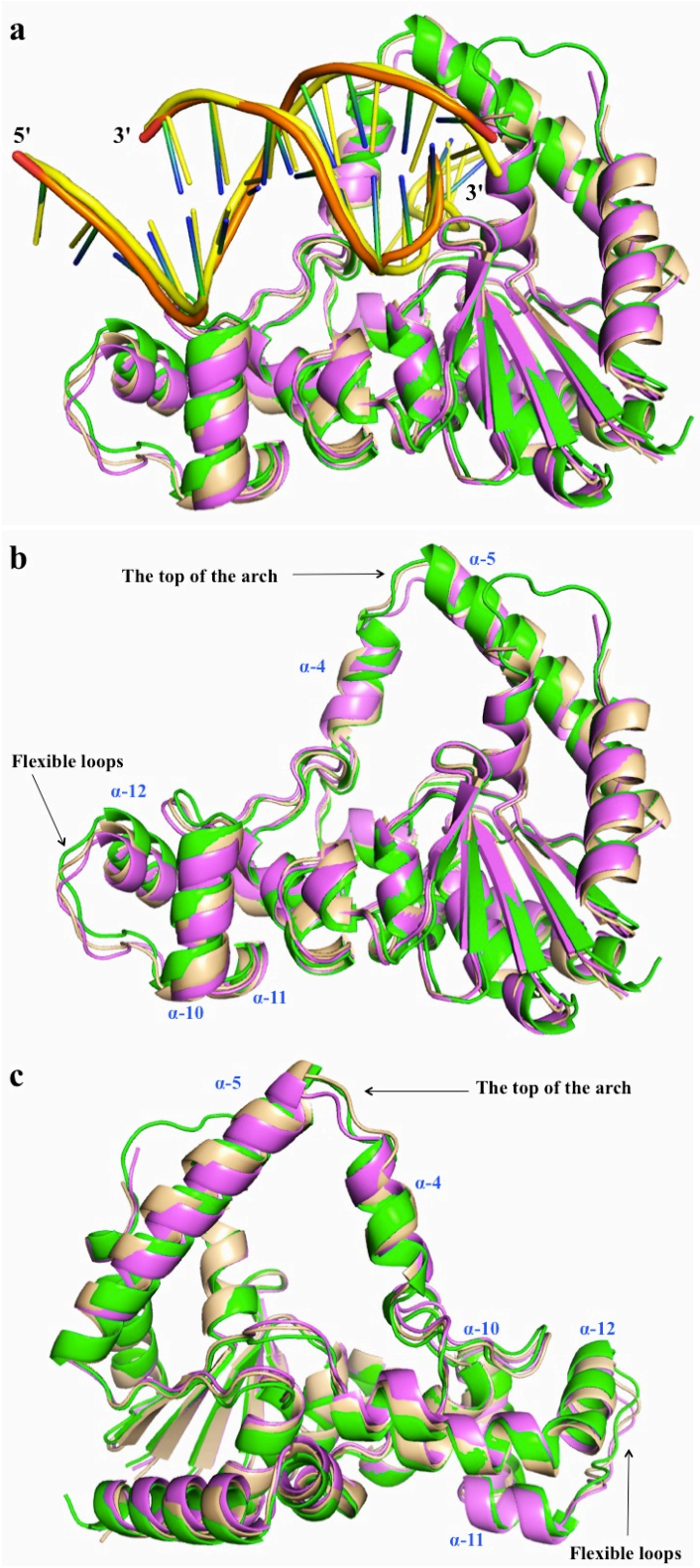



**Figure 4.61:** Superposition of C1/B and C2/B on T5FEN native structure.
**(a)** Shows the superposition structure of C1/B (wheat protein and orange DNA backbone with green bases), C2/B (green protein and yellow DNA) on T5FEN native (violet protein). **(b)** and **(c)** show the location movement regions between the three chains in front and back views respectively.

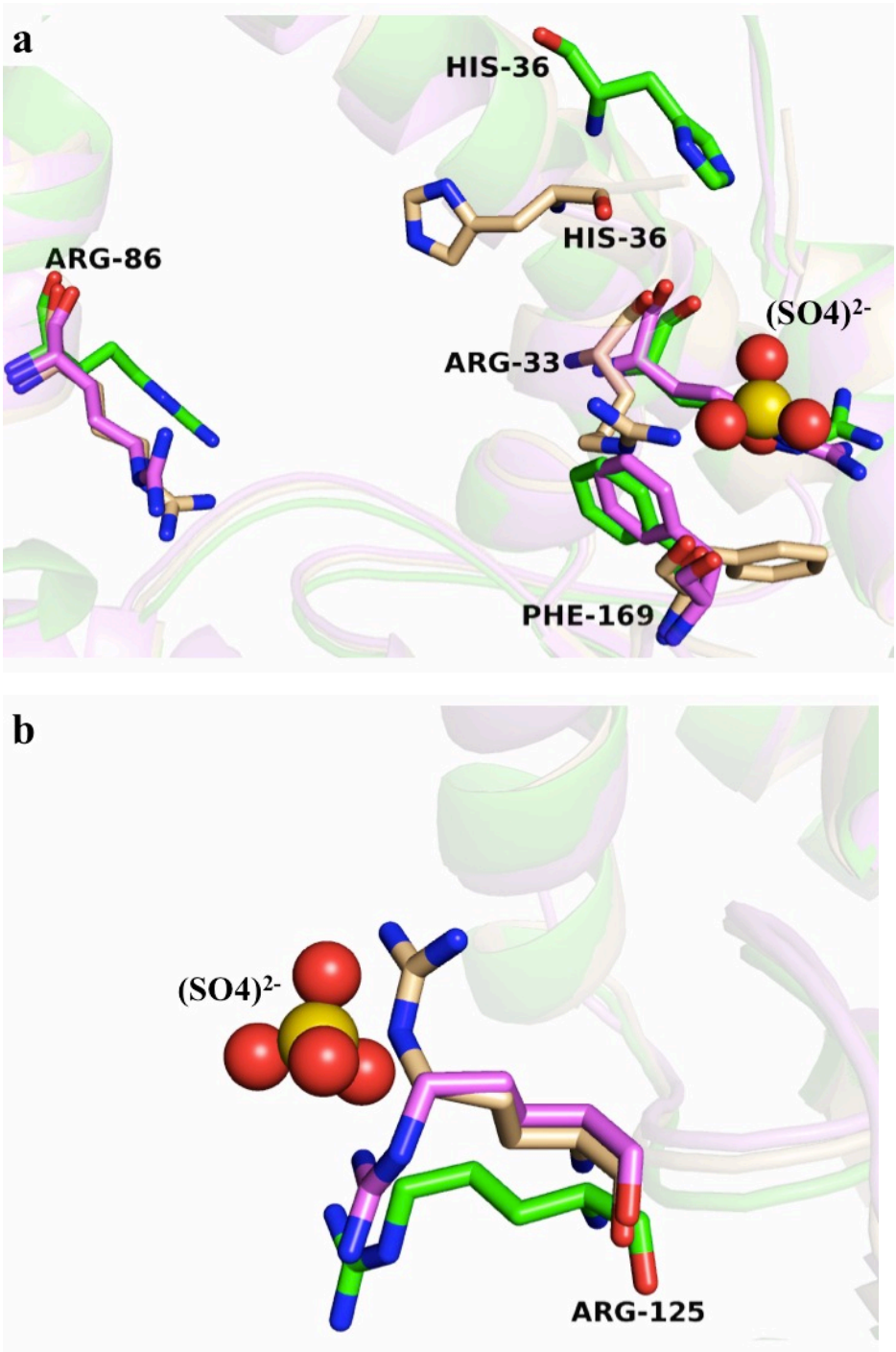



**Figure 4.62:** Superposition of some residues in the C1/B (wheat), C2/B (green) on T5FEN native (violet).
**(a)** Shows four amino acids, Agg-86, His-36, Arg-33 and Phe-169 from the three superimposed structures. **(b)** Shows Arg-125 from each T5FEN chain and here it takes similar conformations in C2/B and the native structure. One of the $(SO4)^{2-}$ ions presents close to the Arg-125 location in C1/B.

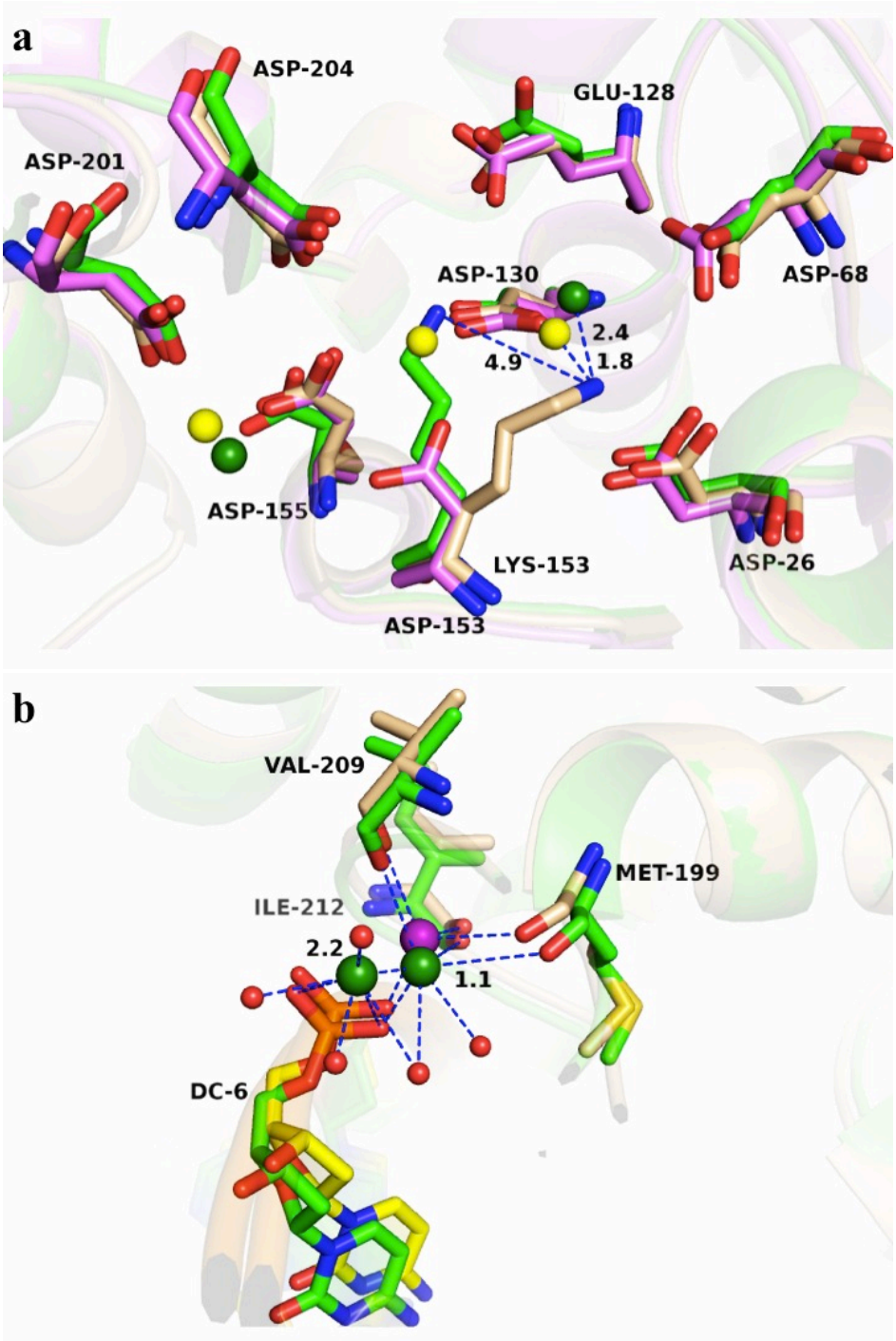



**Figure 4.63:** Superposition of C1/B, C2/B on T5FEN native, active site and $K^+$ ion. **(a)** Superposition of the active site residues in C1/B (wheat sticks), C2/B (green sticks) on T5FEN native (violet sticks). Five magnesium ions are present in the superposed active site, three of them (yellow spheres) belongs to T5FEN native and the other two (green spheres) are found in C2/B while C1/B active site does not has any metal ion. **(b)** Superposed of C1/B potassium ion (purple sphere) in its binding site on C2/B $K^+$ ion (green spheres). Each $K^+$ ion in C1/B and C2/B interacts with the same DNA backbone dC-6, green in C1 structure and yellow in C2.

# Chapter 5: Comparisons between C2-T5FEN-D153K:DNA and other FEN Family Members

Comparisons of C2-T5FEN-D153K:DNA structure with other published FENs including T4RNase H in complex with pseudo-Y, hFEN in complex with duple flap and hEXO1 in complex with 3' flap will be discussed in this chapter.

## 5.1 Comparison of C2 with T4RNase H:DNA complex

The T4RNase H is a member of the FEN family which has 5' flap exo- or endo-nuclease activity (Bhagwat et al, 1997). Its structure has been solved with a DNA substrate containing of 18 and 24 bases to produce a forked DNA with both 5' and 3' arms. This structure was obtained by Devos et al 2007 (Devos et al, 2007) at a resolution of 3 Å using molecular replacement. One protein chain was complexed to one pseudo-Y molecule of the DNA free of metal ions. This structure is mentioned in this thesis as an example of a bacteriophage FEN complexed with pseudo-Y DNA.

### 5.1.1 Superposition of C2/A and C2/B chains on T4RNase H:DNA

Chain A and B of C2-T5FEN:DNA structure overlay well with the T4RNase H:DNA complex (Figure 5.1) with a RMSD of 3.30 Å and 2.42 Å respectively in a core of 206 C$\alpha$s. The two structures are strikingly similar in overall topology but the four DNA chains look less similar as seen in Figure 5.1

#### *5.1.1.1 Differences between C2 and the T4RNase H structures*

Figure 5.2a is a ribbon diagram for the superposition of the two structures and the differences between them. The $\beta$-strands overlay very well except for the last one in each structure (a.a. 172-175) for the T5FEN and (a.a. 222-224) for the T4RNase H. Another obvious difference is in the N-terminal which is visible in the T4RNase H protein as a long loop (Figure 5.2a) while it is disordered and could not be built in the wild type T5FEN that was

solved for the first time by Ceska et al (1996). The first helices in the H3TH motif for both enzymes fit well one on each other but the second one in C2/B does not (Figure 5.2b). Additionally, the loop joining these two helices together is extended in T4RNase H by insertion of about 13 amino acids comparing to T5FEN. After this motif the helices do not fit well together and additional helix and loop conformations are found in T5FEN (residues 235-247) which are missing in the T4RNase H, while another helix and loop are presented in the T4RNase H C-terminal composed of about 14 residues (Figure 5.2a) which are absent from the T5FEN protein.

In addition, the hydrophobic wedge and the $\alpha$2 helix present in both proteins superimpose well on each other with partial disorder in the connecting loop in C2/A while it is fully ordered in C2/B and T4RNase H (Figure 5.2a). Moreover, the $\alpha$-5 helix which is a part of the arch structure fits well but it is smaller than in T4RNase H comparing to T5FEN. The $\alpha$-4 helix which forms the rest of the arch is disordered in T4RNase H whilst it is well ordered in T5FEN (Figure 5.2a).

#### *5.1.1.2 Superposition of the Active Site Residues*

The conserved active site in C2/A and C2/B superimpose very well with the equivalent in T4RNase H. The two mutated residues which made in order to abolish or reduce these enzymes activity are present in the superimposed active sites. Asp-132 was mutated to Asn-132 in T4RNase H (Devos et al, 2007) while Asp-153 was mutated to Lys-153 in T5FEN (Zhang, 2012) (Figure 5.3). The large movements can be seen for both Asp-197 and Asp-200 in T4RNase H by 1.92 Å and 1.81 Å respectively from the equivalent residues in C2/A and 1.7 Å and 1.8 Å respectively in C2/B (Figure 5.3). Moreover, C2/A active site and T4RNase H complexed with the fork DNA have no metal ions while two metal ions are present in the superimposed active site of C2/B and T4RNase H complexes (Figure 5.3b) which are $Mg^{2+}$ ions that occupy M1 in the Cat1 and M3 in the Cat2 sites.

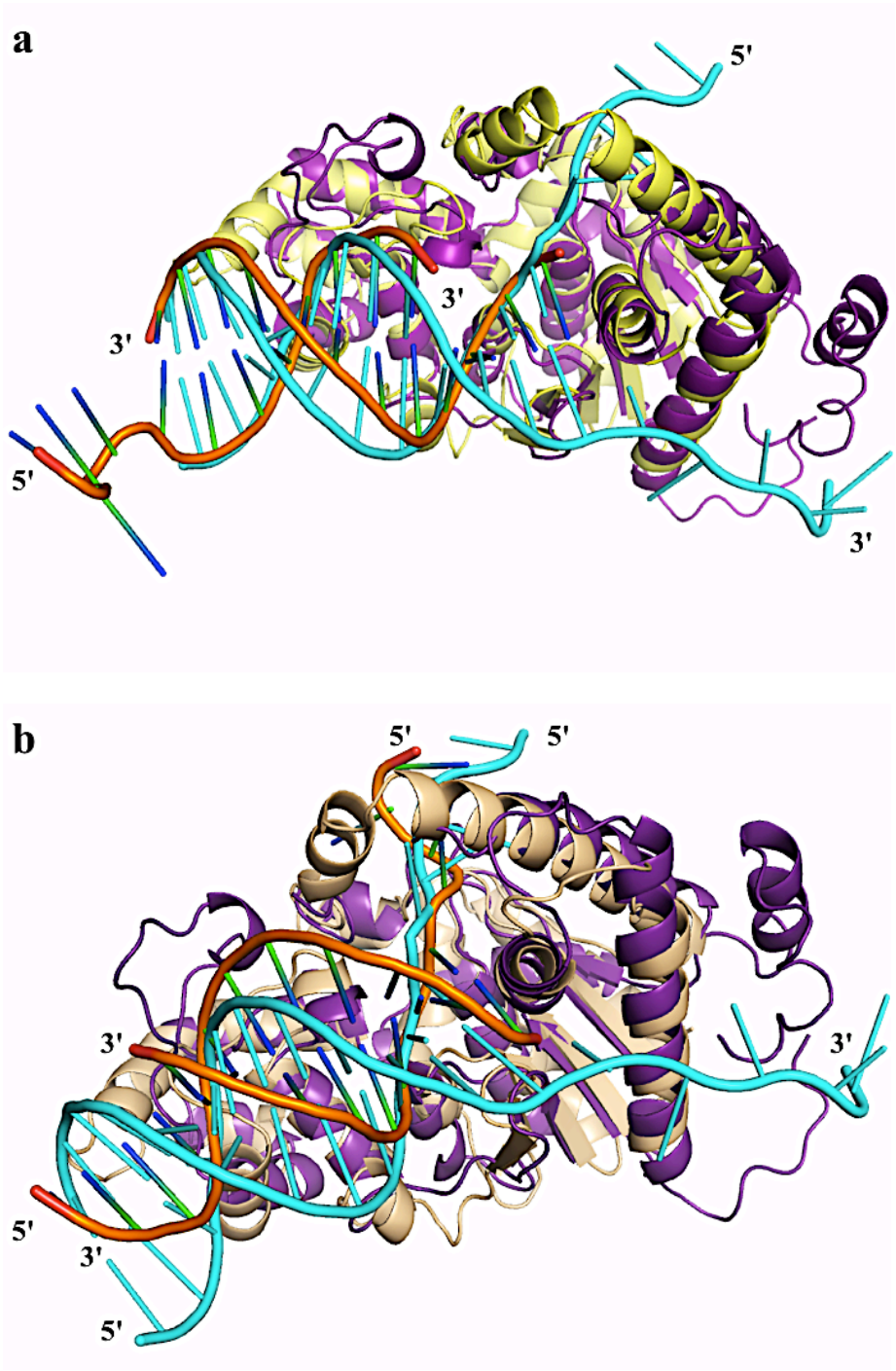



**Figure 5.64:** Superposition of T4RNase H (purple) in complex with pseudo-Y DNA (cyan) on C2-T5FEN-D153K:DNA.
**(a)** and **(b)** show "overhead" views for superposition of C2/A (yellow) and C2/B (wheat) complexes to 5ov4 DNA (green bases and orange backbone) on T4RNase H respectively.

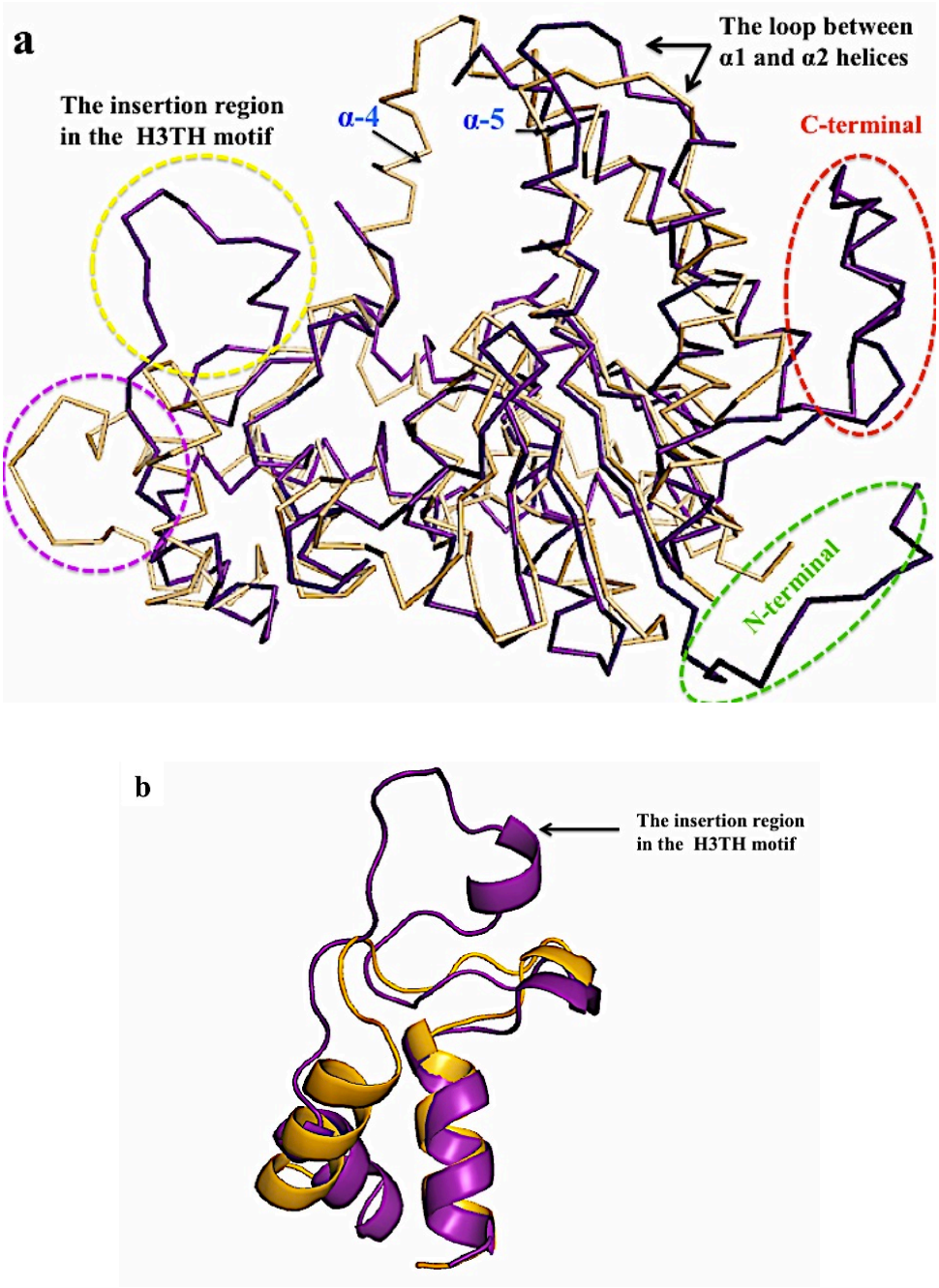



**Figure 5.65:** The major differences between the two proteins, T5FEN (wheat) and T4RNase H (purple) structures.
**(a)** Ribbon diagram shows the N-terminal (green dashes circul), the C-terminal (red dashes circul), the insertion in the H3TH loop (yellow dashes circul) in the T4RNase H and the insertion part (magenta dashes circul) in T5FEN. **(b)** Shows the superposition of T4RNase H H3TH motif (purple) on T5FEN (wheat).

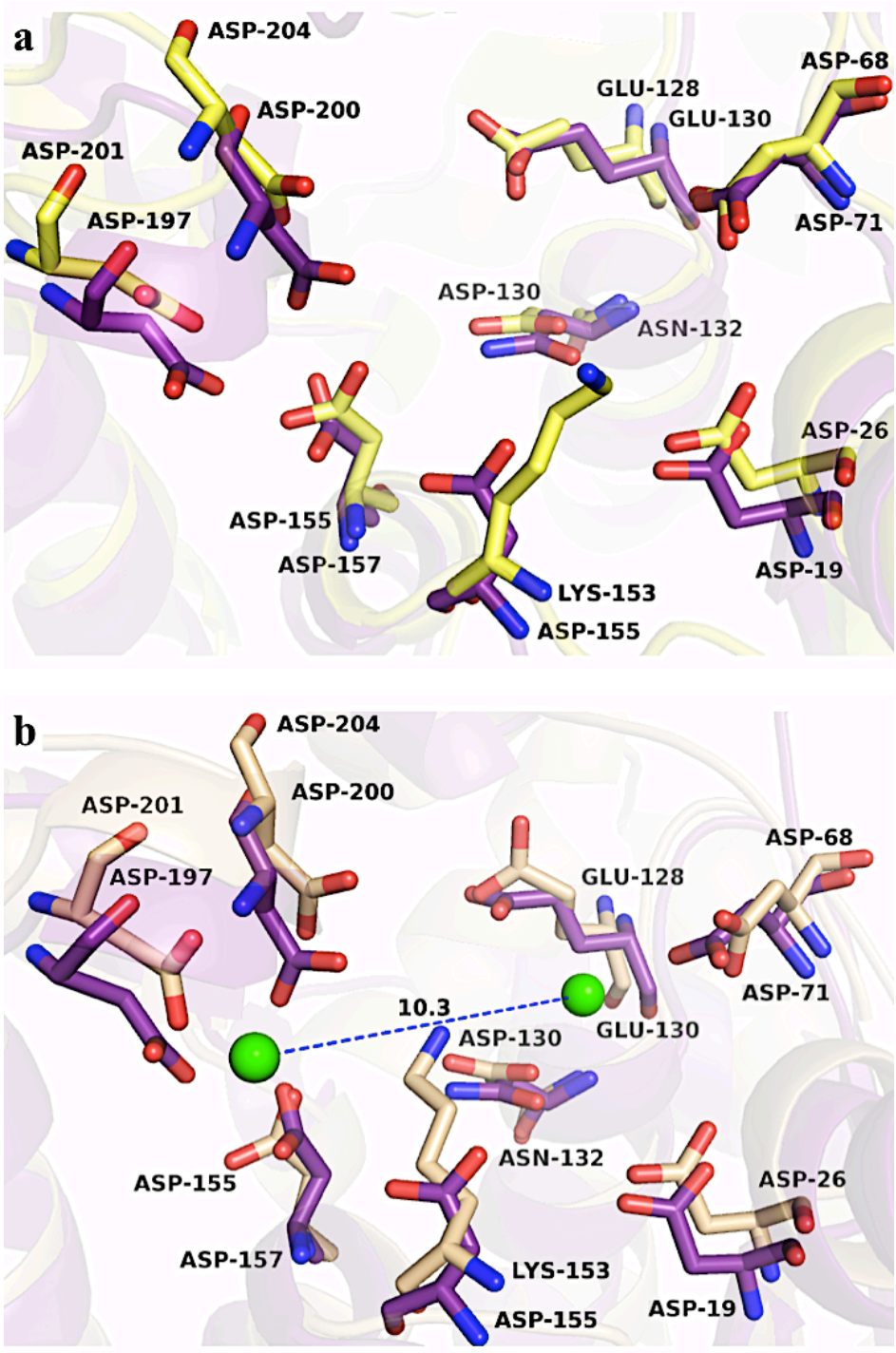



**Figure 5.66:** Superposition of the active site residues of C2-T5FEN-D153K:DNA on T4RNase H:DNA (purple).
**(a)** Shows the superposition with C2/A active site residues (yellow sticks). **(b)** Illustrate the superposition of C2/B active site residues (wheat sticks). The two $Mg^{2+}$ ions are shown as green spheres.

### *5.1.1.3 Interactions with the DNA Substrates*

#### 5.1.1.3.1 Interactions with the H3TH:$K^+$ Motif

In both complexes, C2 chain A and B and T4RNase H:DNA, the 3' end of the duplex DNA substrates interacts with the H3TH motif by forming hydrogen bonds between the motif residues and the DNA backbone as described in detail in sections (4.1.2.2.1) and (1.2.1.1.1) for C2 and T4RNase H:DNA respectively. The potassium binding site observed in C2/A and C2/B (see section 4.1.2.2.1) was not reported in the T4RNase H:DNA structure.

#### 5.1.1.3.2 Interactions with the Hydrophobic Wedge

The junction of the bifurcated substrate in each enzyme complex interacts with the hydrophobic wedge which is structurally conserved in FEN family members. These interactions are explained in section (4.1.2.2.3) for the C2/B and in section (1.2.1.1.2) for the T4RNase H:DNA while in C2/A the end duplex of the 5ov4 DNA did not reach the helix wedge and so could not make any interactions with this part of the enzyme as mentioned previously.

#### 5.1.1.3.3 Interactions of the 5' End

The 5' flap of the 5ov4 DNA in C2/A extends above the active site but does not reach the arch hole and it is too short to pass through it. Consequently, no significant interactions were observed here (see section 4.1.2.2.4). However, the 5' arm of the forked DNA which is a part of the T4RNase H:DNA complex (DNA/C) can be modeled onto the C2/A structure where it penetrates through the ordered arch (Figure 5.1a) and binds similarly to the 5' end of 5ov4 (DNA/Y) which is threaded through the arch hole in C2/B. The conserved Arg-86 in C2/A is 3.8 Å and 4 Å from dT-7/C and dT-8/C phosphate oxygens of the fork DNA (Figure 5.4a) in this model. The importance of this residue, Arg-86, was investigated in 1997 by Nossal et al. (Bhagwat et al, 1997) in T4RNase H when they mutated the equivalent residue, Arg-90, to Ala. This study demonstrated that Arg-90 is important

for the DNA binding and the enzyme activity which were reduced relative to the wild type enzyme.

In superposition of C2/B on T4RNase H:DNA the 5' arms are followed the very similar path and can be seen the threading through the helical arch which is well ordered in the C2 but it is only partially ordered in the T4RNase H:DNA structures (Figure 5.4b). The charged side chain of T4RNae H Arg-90 points away from the DNA backbone, ~12 Å, and it is too far to interact with the DNA phosphate groups (Figure 5.4b). In contrast, Arg-86 in C2/B could form bifurcated hydrogen bonds between its terminal amino group hydrogens and the dT-6/C and dC-5/C phosphate oxygens in the 5' arm of the fork DNA (Figure 5.4b).

The gate-like structure that was formed in front of the arch by Tyr-90 and Phe-105 in C2/B (Figure 4.17a) as one of the suggested mechanisms to prevent the ssDNA moving back from the arch after threading through (see section 4.1.2.2.4) is not seen in the T4RNase H:DNA structure as this region is disordered. Phe-105 presents in both proteins: the T5FEN and the homolog T4RNas H at the same position (Figure 5.4b) but Tyr-90 in the T5FEN is replaced by the disordered Arg-94 residue in the T4RNase H.

#### 5.1.1.3.4 The trans-arch/distal DNA Binding Site

The new trans-arch/DNA binding site does not interact with the DNA backbone in C2/A because the ssDNA does not reach the arch and thread through it. Superposition of T5FEN:DNA on T4RNase H shows Gly-70 and Lys-71 main chains in C2/A far from the pseudo-Y backbone (5.9 Å and 4.8 Å respectively) and cannot interact with it (Figure 5.5a). This site is clearly observed in C2/B as mentioned above. The equivalent amino acids in the T4RNase H (Lys-74 and Ser-75) cannot interact with the DNA backbone by their main chains as observed in C2/B but the side chain of Lys-74 can connect to the dA-3 phosphate oxygens of the fork DNA (Figure 5.5b) which gives an indication that this site could be formed in the T4RNase H enzyme as it is in homolog T5FEN. Additionally, the equivalent phosphate

group in the 5' arm of the pseudo-Y substrate, dC-5/C, is shifted by 4 Å from the dA-4/Y phosphate group position in the 5ov4 substrate and could not interact with the T5FEN trans-arch/distal site (Figure 5.5b).

#### 5.1.1.3.5 Interactions of the DNA 3' Arm

The 5ov4 DNA complexed with the T5FEN lacks a 3' arm as described before and so no interactions were observed for this part of the DNA substrate in this enzyme. The interactions between the T4RNase H enzyme and the 3' end are explained in section 1.2.12. Most of these interactions were made between the 3' arm of the DNA and the extended region of the C-terminal in the T4RNase H protein which is structurally not a part of the T5FEN enzyme.

## 5.2 Comparison of C2 structure with the hFEN:DNA complex

Human flap endonuclease (hFEN1) is a member of FEN superfamily and like other FENs it is important in processing Okazaki fragments during DNA replication. It is presented in this analysis as an example of a FEN superfamily member complexed to a duple flap DNA composed of 11 base pairs in the downstream region and another six pairs as an upstream region with a single 3'-nucleutide and 4 nt 5' flap.

The crystal structure was solved by Tsutakawa et al (2011) at a resolution of 2.2 Å. In addition to presence of the DNA molecule, this wild type structure contained two metal ions ($Sm^{3+}$) in its active site and a $K^{+}$ ion within the H2TH motif. The structure crystallized with $Mg^{2+}$ ions in addition to $Sm^{3+}$ ions which allowed the hydrolysis of the scissile phosphate to be occurred 1 nt into the DNA duplex and so the crystal structure called hFEN:DNA product. Another complex structure for hFEN:DNA has the same components of $Sm^{3+}$ ions and $K^{+}$ ion in the active site and H2TH motif respectively and has 1 nt DNA flap instead of 4 nt in the previous one was solved at 2.3 Å and called hFEN:DNA substrate (Tsutakawa et al, 2011).

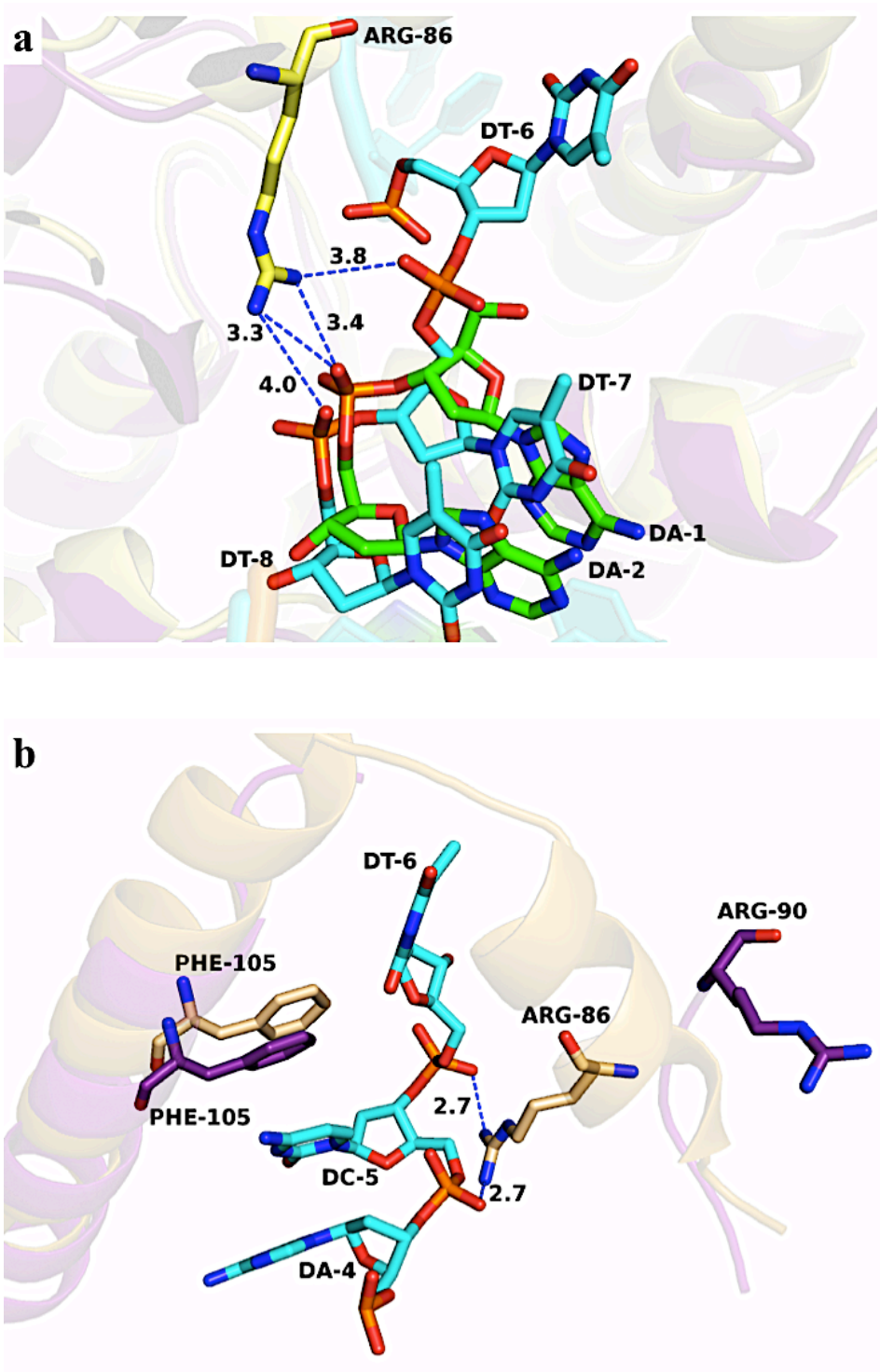



**Figure 5.67:** DNA interactions of superimposed C2-5' overhang on T4FEN (purple protein & cyan DNA).
**(a)** and **(b)** show the possible interactions of Arg-86 in C2/A (yellow sticks) and C2/B (wheat sticks) with pseudo-Y DNA (orange backbone and cyan bases) respectively.

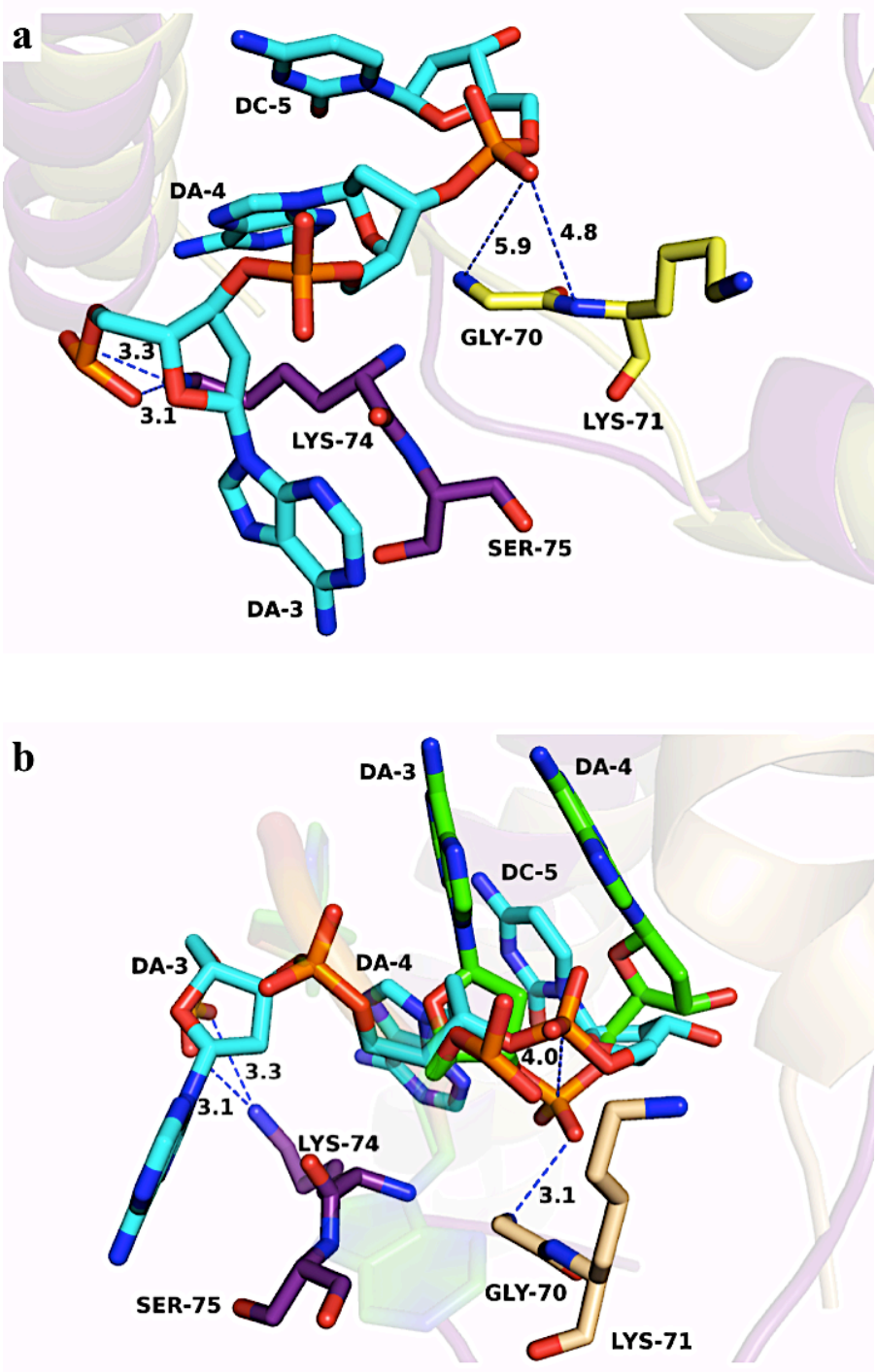



**Figure 5.68:** Superposition of the new trans-arch/DNA binding site in C2 on T4RNase H:DNA.
**(a)** Shows the trans/arch distal DNA binding site of C2/A (yellow sticks) far from the modeled pseudo-Y DNA backbone and could not interact with it. **(b)** Shows the same DNA binding site in C2/B (wheat sticks) and the equivalent residues in T4RNase H:DNA.

Theses two complexes will used in C2 structure comparison while the third hFEND181A:DNA mutant complex (Tsutakawa et al, 2011) wont be used because it was crystallized with 1 nt DNA flap and $K^+$ ion only and the D181A mutant did not add more sufficient information to the complex.

### 5.2.1 Superposition of C2/A and C2/B on hFEN:DNA

The overall topology for the two FEN members, C2 and hFEN, is preserved (Figure 5.6). The main features such as the helical arch, the hydrophobic wedge, the H3/2TH motif and the active site are structurally conserved and overlay well on each other. The superposition of these two enzymes together had an RMSD of Å 3.1 Å and 2.986 on C$\alpha$s for chain A and B respectively and some differences are noticed between them.

The two DNA substrate superimposed good on each other (Figure 5.7) chain X and Y of 5ov4 in C2 on chain D and H respectively of the duple flap in hFEN:DNA complex.

#### *5.2.1.1 Differences Between the Two Structures*

The helical arch is well ordered in C2 and in hFEN enzymes which is a result for the DNA substrate binding in the latter one (Sakurai et al, 2005; Tsutakawa et al, 2011) (Figure 5.6). The two proteins have similar ordered helices in the arch-like structure but it looks longer in hFEN enzyme (Figure 5.8a & b). The next difference between these FENs is in the extended region of two helices and a loop at the C-terminal of the hFEN which is clearly absent from the T5FEN protein (Figure 5.8a, b).

In the hFEN the second helix in the H2TH motif superimoses very well with its equivalent in the T5FEN than the first one (Figure 5.8c). The first $\beta$-sheet ($\beta$-4 in T5FEN) in the $\beta$-pin motif fits better than the second one with an extended loop in hFEN (about 13 residues) compared to the $\beta$-pin loop in the T5FEN (about five residues) (Figure 5.8d).

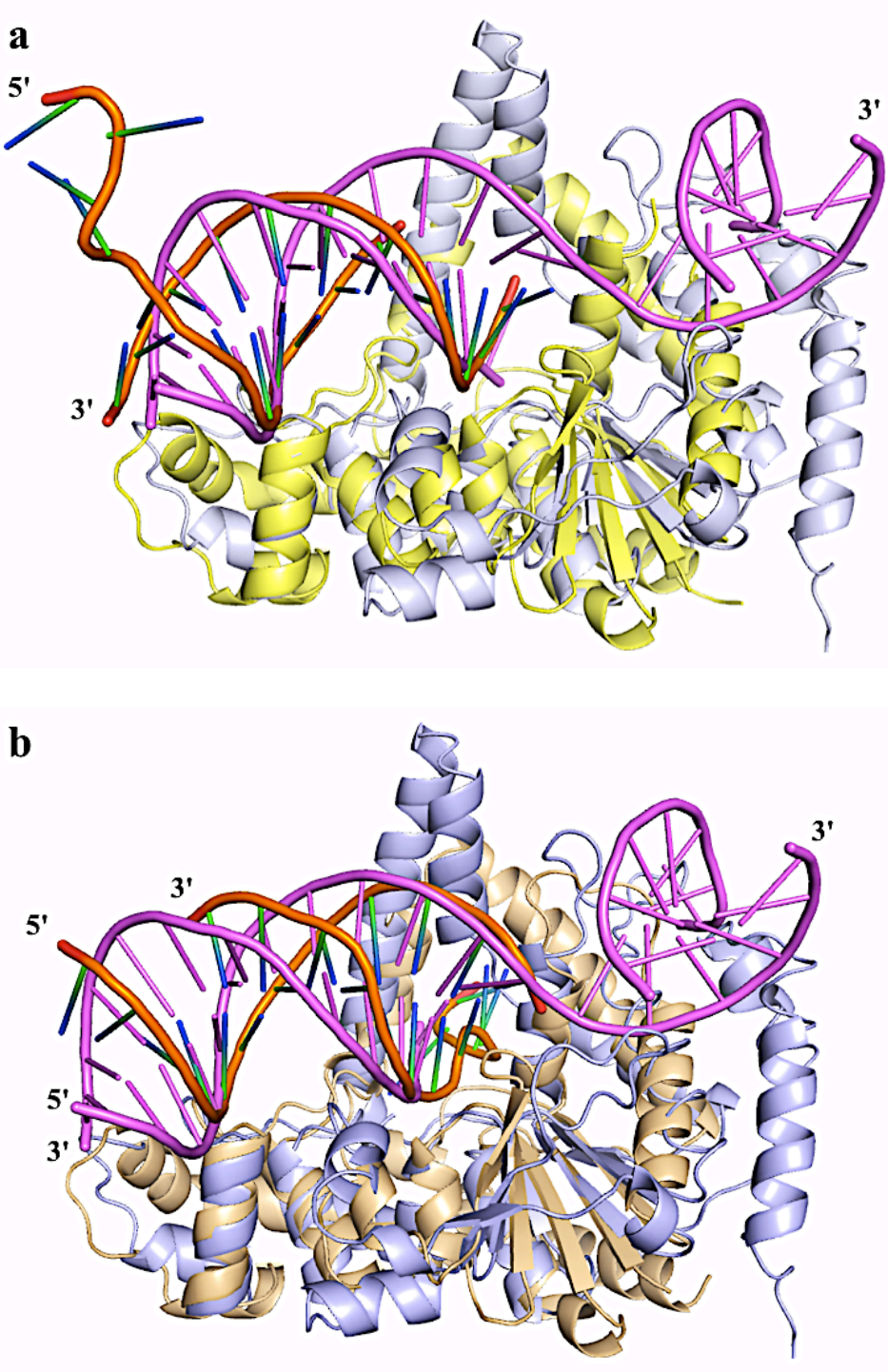



**Figure 5.69:** Superposition of C2-T5FEN-D153K:DNA on hFEN (white blue) complexed to the double flap product DNA (magenta).
**(a)** and **(b)** show the overall superposition of C2/A (yellow) and C2/B (wheat) on hFEN:DNA respectively.

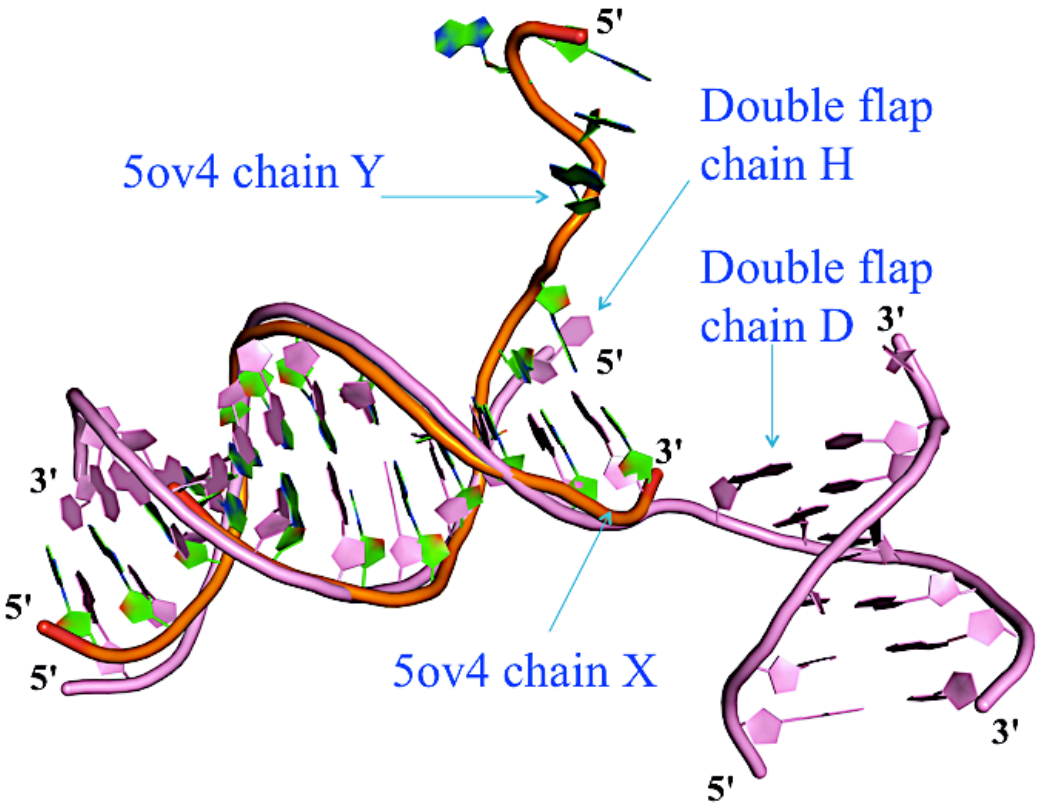



**Figure 5.70:** Superposition of the C2 5ov4 DNA (orange backbone and green bases) on the double flap product DNA (pink) of hFEN:DNA complex.
The figure shows matches between the two strands Y and X of the 5ov4 with H and D of the double flap respectively and well overlay in the duplex region.

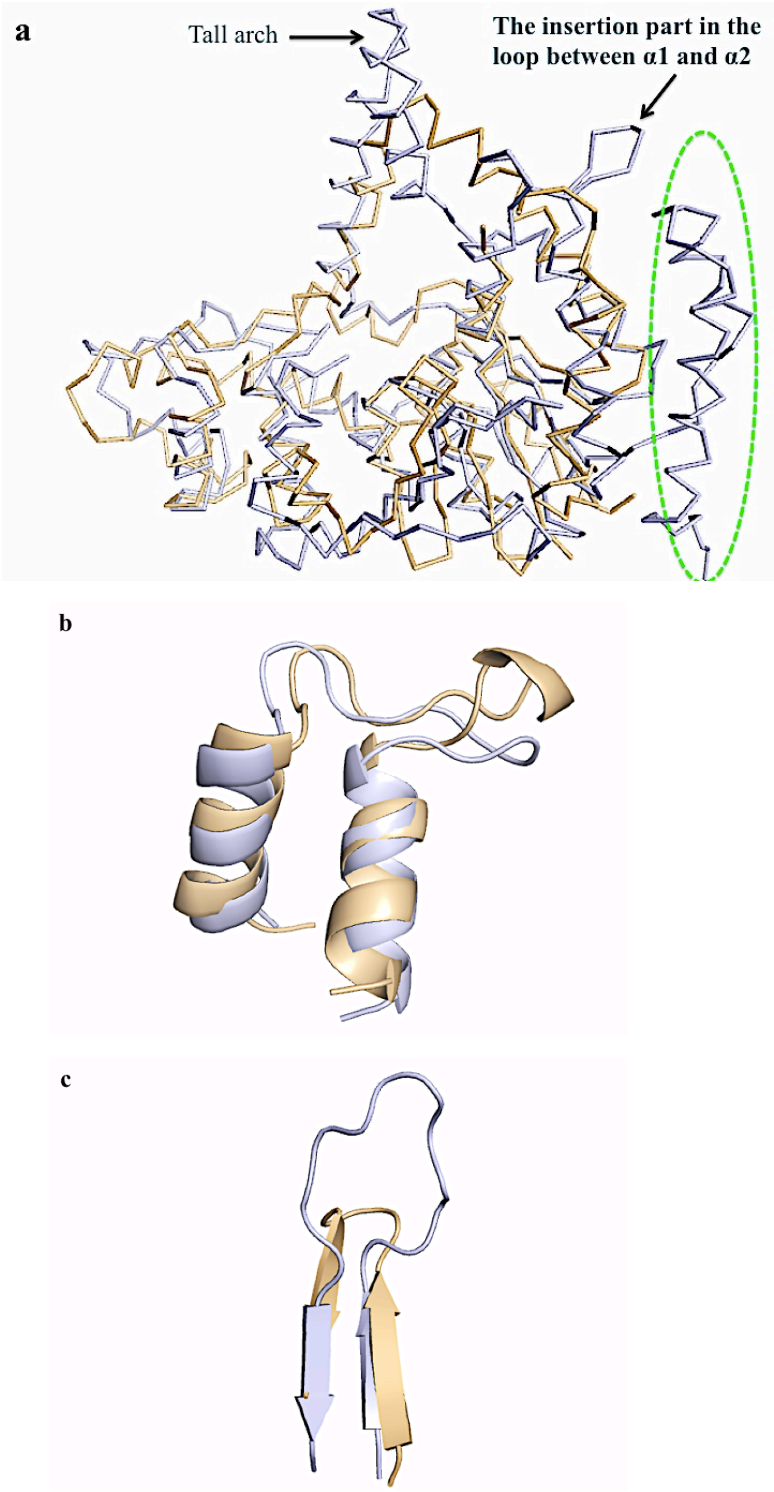



**Figure 5.71:** The differences between hFEN (white blue) complexed to the double flap DNA (magenta) T5FEN:DNA (wheat).
**(a)** Ribbon diagram for the two superimposed enzymes. **(b)** Shows the two H3/2TH motifs overly well to each other. **(c)** Shows the $\beta$-pin motif in the two complexes and the extended loop between the two $\beta$-sheets in the hFEN enzyme (white blue) is compared to that in T5FEN (wheat).

In both proteins the hydrophobic wedge is superimpose well and the loop lies between the $\alpha 1$ ad $\alpha 2$ helices in the hFEN (Figure 5.8a) which engage in formation of the 3'-single nucleotide pocket with the C-terminal, is absent in T5FEN enzyme.

#### *5.2.1.2 Superimposed Analysis of the Active Sites*

The active site in the hFEN:DNA complex structure is similar to that in the C2 structure. From Figure 5.9a and b it is clear that the Cat1 site is structurally well conserved between these two FEN proteins. Cat2 in hFEN is not present in the same position of T5FEN which is occupied by the N-terminal (Gly-2) in hFEN active site. Strikingly, Asp-130 in the T5FEN Cat1 site is replaced by Glu-160 with its longer side chain in the hFEN Cat1 site (Figure 5.9). Some movement places in the active site amino acids in C$\alpha$s are seen and they range from less than 0.5 Å for Asp-34 in hFEN from its equivalent Asp-26 in T5FEN to 3.11 Å for Asp-233 in hFEN and its equivalent Asp-204 in the T5FEN chain A (Figure 5.9a). In addition, the two $Sm^{3+}$ metal ions that are present in the hFEN:DNA active site of the substrate and product complexes could occupy M1 and M2 in Cat1 site of the C2/A and C2/B (Figure 5.9a & b). The distance between Lys-153 in C2/A and the second $Sm^{3+}$ in the M2 site is 5.2 Å (Figure 5.9a). The two $Mg^{2+}$ ions in C2/B are present in the superposed active site Cat1 (M1) and Cat2 (M3) sites as mention previously (Figure 5.9b). The distance between the superimposed $Mg^{2+}$ ion in M1 site and $Sm^{3+}$ in M2 site is 4.6 Å and 4 Å in the hFEN:DNA substrate and product complexes respectively (Figure 5.9b). In contrast, the second superposed $Mg^{2+}$ ion in M3 of the Cat2 site is 7.2 Å and 7.6 Å away from $Sm^{3+}$ ion in M2 of hFEN:DNA substrate and product structure respectively (Figure 5.9b).

These differences in the distances between the metal ions in hFEN:DNA substrate and product complexes are due to the movement of the $Sm^{3+}$ in M2 of Cat1 in the substrate complex structure toward the appropriate place for the two metal ions mechanism allowing the reaction to occur in the product complex (Tsutakawa et al, 2011).

The mutated residue Lys-153 that presents in the Cat1 of C2 structure is clearly extended inside the superimposed active sites. In C2/A the Lys-153 long side chain rotates by about $\chi_1$ 120° and interacts with the 5ov4 DNA backbone, dA-2/X, and is 5.4 Å away from the equivalent phosphate oxygen ,dT-2/E, in hFEN:DNA substrate complex (Figure 5.10a) while it is stacked with this base (dT-2/E) in the product complex and it is ~ 1.4 Å from it (Figure 5.10b). The scissile phosphate (dT-2/E) in the hFEN:DNA product complex is 2.2 Å and 2.3 Å away from the first $Mg^{2+}$ ion in M1 of Cat1 site which is expected because this ion is similar in its location for the first $Sm^{3+}$ ion (Figure 5.10c). Interestingly, the second $Mg^{2+}$ in C2/B Cat2 site is 3.5 Å and 2.0 Å from the dG-3/E phosphate oxygens in hFEN:DNA (Figure 5.10c). Superposition of the active site for the two complexes C2/B and hFEN:DNA product shows the dC-6/Y scissile phosphate of the 5ov4 DNA is placed in the active site and it is 2.2 Å and 3.7 Å apart from the first and the second $Sm^{3+}$ ions respectively in the hFEN enzyme (Figure 5.10d).

#### *5.2.1.3 The Potassium Binding Site*

The potassium binding site is observed for the first time in the T5FEN:DNA complex as well as for the hFEN:DNA complex (Tsutakawa et al, 2011). This site could present only in the presence of the DNA substrate in FEN enzyme complexes. Superposition of the potassium binding site in C2/A and C2/B on hFEN:DNA (Figure 5.11a & b) shows location similarity between them. Additionally, the $K^+$ ion in hFEN:DNA is ~ 1.0 Å and 2.0 Å away from its position in C2/A and C2B, (first position), respectively (Figure 5..11a & b). The $K^+$ ion in T5FEN:DNA structure is coordinated by the main chain carbonyl oxygens of Met-199, Val-209 and Ile-212 which are equivalent to Leu-230, Ile-238 and Ile-241 respectively in hFEN:DNA. Leu-230 in hFEN is far away from the $K^+$ site and cannot coordinate the potassium ion but instead Ser-237 side chain is engage in the $K^+$ coordination (Figure 5.11). The hFEN Ser-237 is equivalent to Gly-208 in T5FEN which is also cannot interact with the $K^+$ ion in C2 structure but instead Val-209 coordinates the $K^+$ ion (Figure 5.11).

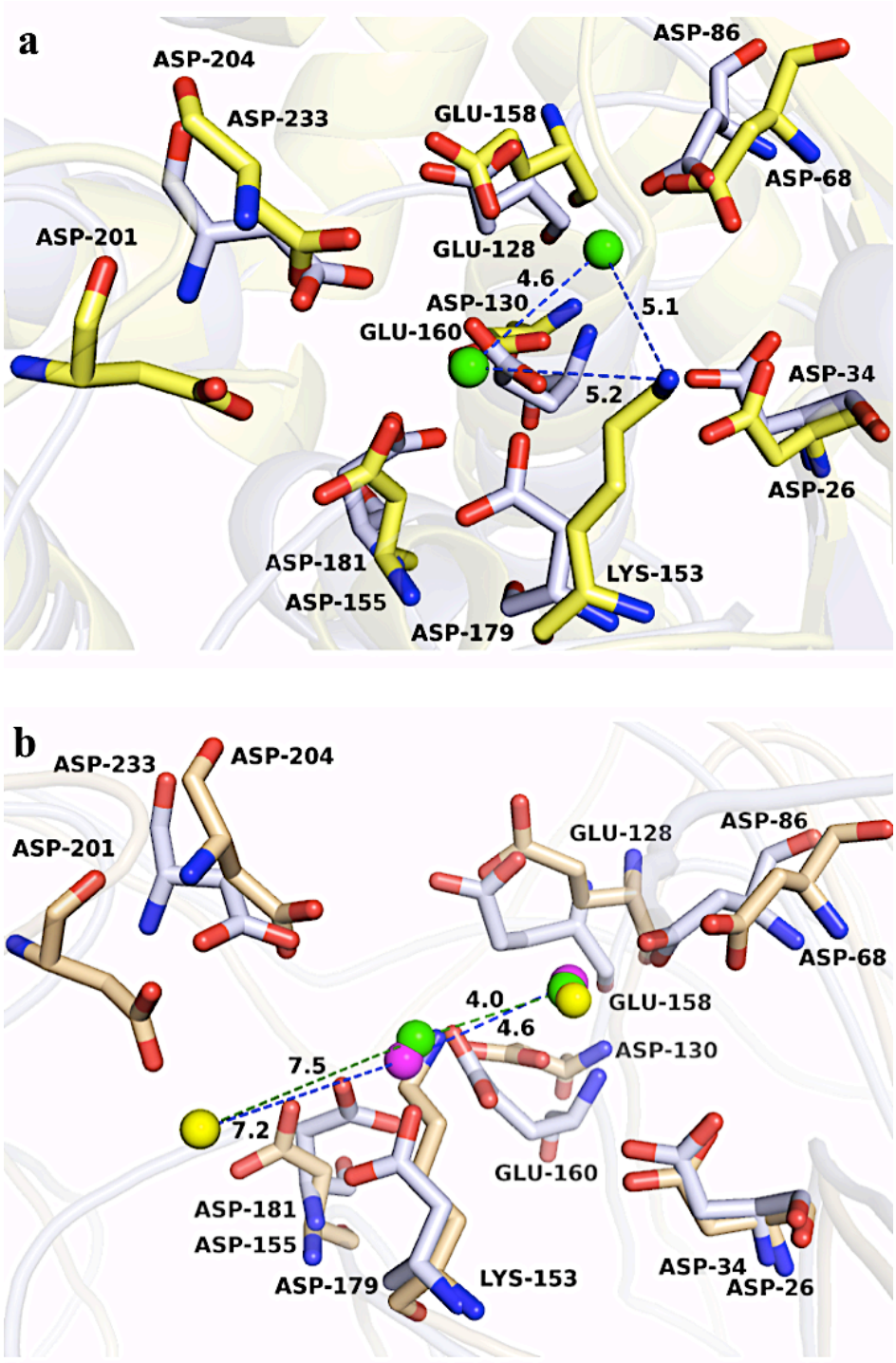



**Figure 5.72:** Superposition of the active site residues in hFEN:DNA (white blue) on C2/A (yellow) and C2/B (wheat) shown in **(a)** and **(b)** respectively.
It shows the four metal ions: $2Mg^{2+}$ (yellow spheres) for C2/B and $2Sm^{3+}$ (magenta and green spheres) for hFEN:DNA substrate and product complexes respectively.

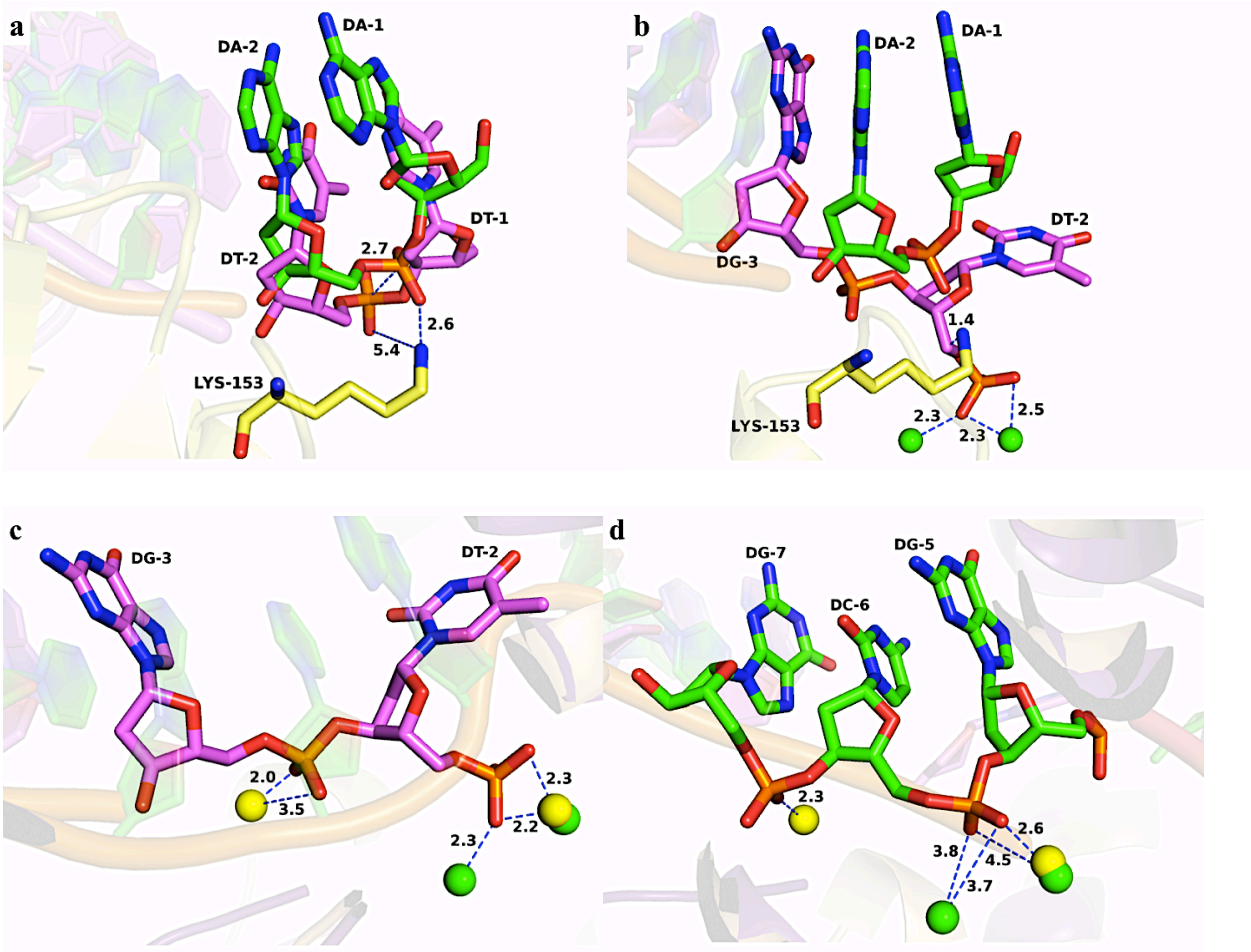



**Figure 5.73:** The possible interactions between the DNA and the metal ions in the superposed active site of C2 structure and hFEN:DNA.
**(a)** and **(b)** The distances between the mutated Lys-153 in C2/A and hFEN substrate and product respectively. **(c)** Shows the interactions that could be made between the two $Mg^{2+}$ ions (yellow spheres) in Cat1 and Cat2 with the 5' flap DNA backbone (magenta bases and orange backbone) of hFEN:DNA product. **(d)** Shows the interactions between the two $Sm^{3+}$ (green spheres) in the hFEN Cat1 site that could be made with the 5ov4 DNA backbone (green bases and orange backbone).

The interactions of $K^+$ ion in both proteins with DNA substrate are explained in section 1.2.1.1.1 and 4.1.2.2.2 for the hFEN:DNA and the T5FEN:DNA respectively.

### *5.2.1.4 Interactions of the DNA*

#### 5.2.1.4.1 Interactions of the Duplex Regions

The interactions of the double strand DNA in both FEN members presented here are observed in two regions of the proteins: the H3/2TH:$K^+$ motif and the helix wedge region. Superposition of these two enzymes together confirms the same interactions in these structurally conserved motifs. The duplex region of 5ov4 overhang in C2 structure and the double flap substrate in hFEN:DNA is connected to the H3/2TH:$K^+$ motif through hydrogen bonds formed between the DNA and the protein residues which are described in more detail in sections 4.1.2.2.1 and 1.2.1.1.1 for the C2 and hFEN:DNA respectively. The GhG sequence within the H3/2TH in T5FEN and hFEN are present in these two proteins. The 3' end of the duplex in the two enzymes complex structures are stacked against the hydrophobic wedge motif which also explained in section 4.1.2.2.3 for the C2/B and in section 1.2.1.1.2 for the hFEN:DNA.

#### 5.2.1.4.2 Interactions of the 5' Flap

The superimposition of hFEN:DNA substrate and product complexes on C2/A shows similarity in the 5' flap between 5ov4 DNA in C2/A and DNA substrate in hFEN complex in which both of them are placed above the active sites (Figure 5.12a) while in hFEN:DNA product this 5' flap is placed in the active site (Figure 5.12b). The 5' flap in the C2/B passes through the arch and makes some interactions with the amino acids that are built this motif. These interactions are described in section 4.1.2.2.4. In the hFEN:DNA product the 5' flap was cleaved and some interactions were seen between the monophosphate of the cleaved product and the hFEN enzyme which are explained in section 1.2.1.3.

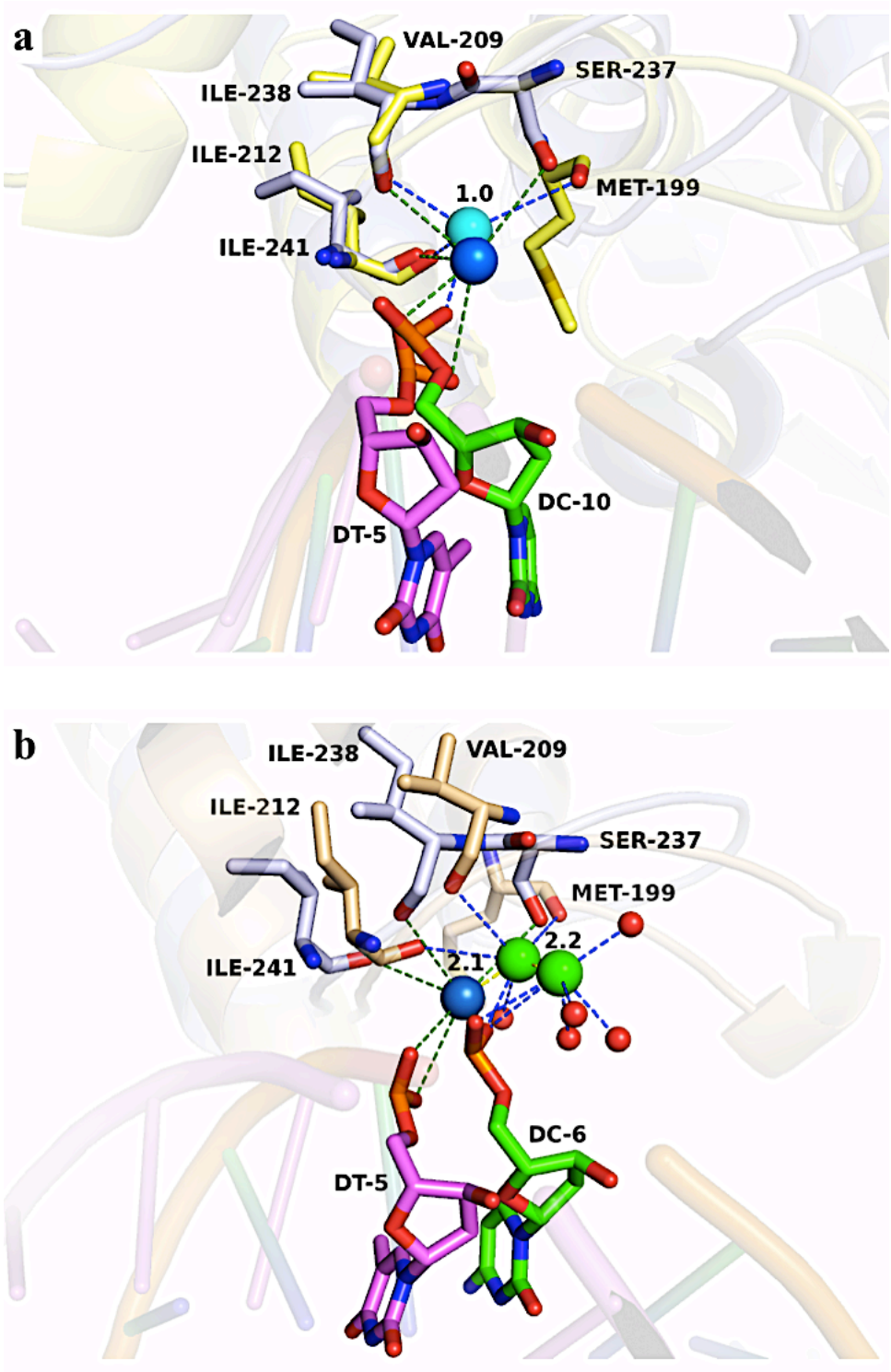



**Figure 5.74:** Superposition of the potassium binding site in hFEN:DNA (white blue protein and magenta DNA) on C2 structure.
**(a)** and **(b)** show the superposition on C2/A (yellow protein and green DNA) and C2/B (wheat protein and green DNA) respectively. The $K^+$ ion is present in hFEN:DNA as blue sphere while it is colored in cyan and green in C2/A and C2/B respectively.

Interestingly, the conserved Arg-86 in C2/A and its equivalent Arg-100 in the hFEN:DNA substrate and product complexes fit very well with each other (Figure 5.12c & d respectively). Arg-86 in C2/A could interacts with the DNA backbone in hFEN:DNA substrate complex (Figure 5.12c) by forming hydrogen bond between its pair ions and dT-2/E phosphate oxygens and it could form another hydrogen bond from its amino group to dT-1/E oxygen (Figure 5.12c) while it is about 3.7 Å and 4.5 Å from the DNA in hFEN product complex (Figure 5.13d). In the superposition with C2/B, Arg-86 shifts by 2.93 Å and 2.77 Å from the Arg-100 C$\alpha$ position in hFEN substrate and product complexes respectively (Figure 5.13a & b) and it could not interacts with the DNA backbone in these complexes (Figure 5.13a & b) while Arg-100 in hFEN:DNA product could form hydrogen bond to the dG-5/Y phosphate oxygen in the 5ov4 DNA (Figure 5.13b).

The new trans-arch/distal DNA binding site that is found in T5FEN and composed of Gly-70 and Lys-71 residues is absent in hFEN. The equivalent residues Lys-88 and ser-89 in hFEN are too far to interact with the 5ov4 DNA backbone in C2 structure. They are 4 Å and 2.5 Å respectively from the interaction points in the trans-arch/distal site (main chain amide nitrogens) in C2/B (Figure 5.13c) which give an indication that this site could be absent in hFEN enzyme or may interact through alternative method or have some conformational changes after threading of the 5' flap through the arch in this enzyme.

#### 5.2.1.4.3 Interactions of the 3' Arm

C2-5ov4 substrate has the 5' overhang but lacks the 3' arm. In hFEN:DNA complex this arm presents and interacts with the C-terminal residues which are structurally missing in the T5FEN protein. These interactions are explained in section 1.2.1.2. Superposition of C2 on the hFEN:DNA shows the absent of the 3' flap pocket from its structure (Figure 5.14a) which is conserved in FEN superfamily members from arch to human (Friedrich-Heineken et al, 2003) but not in bacteriophage FENs. In hFEN Gln-54 and Thr-61 interact with the 3' single nucleotide hydroxyl group. The equivalent

residues Lys-41 and Pro-42 in the C2 are able to interact with the same nucleotide (dC-7/F) (Figure 5.14b). Furthermore, Thr-50 in C2 which is equivalent to Arg-70 in hFEN is able to interact with the dA-14/D phosphate oxygen (Figure 5.14b) while Ser-49, Ser-53 and Thr-170 in C2, which are equivalent to Tyr-90, Arg-73 and Thr-195 in hFEN, are so far from the 3' arm DNA substrate and as so these residues could not interact with it (Figure 5.14b).

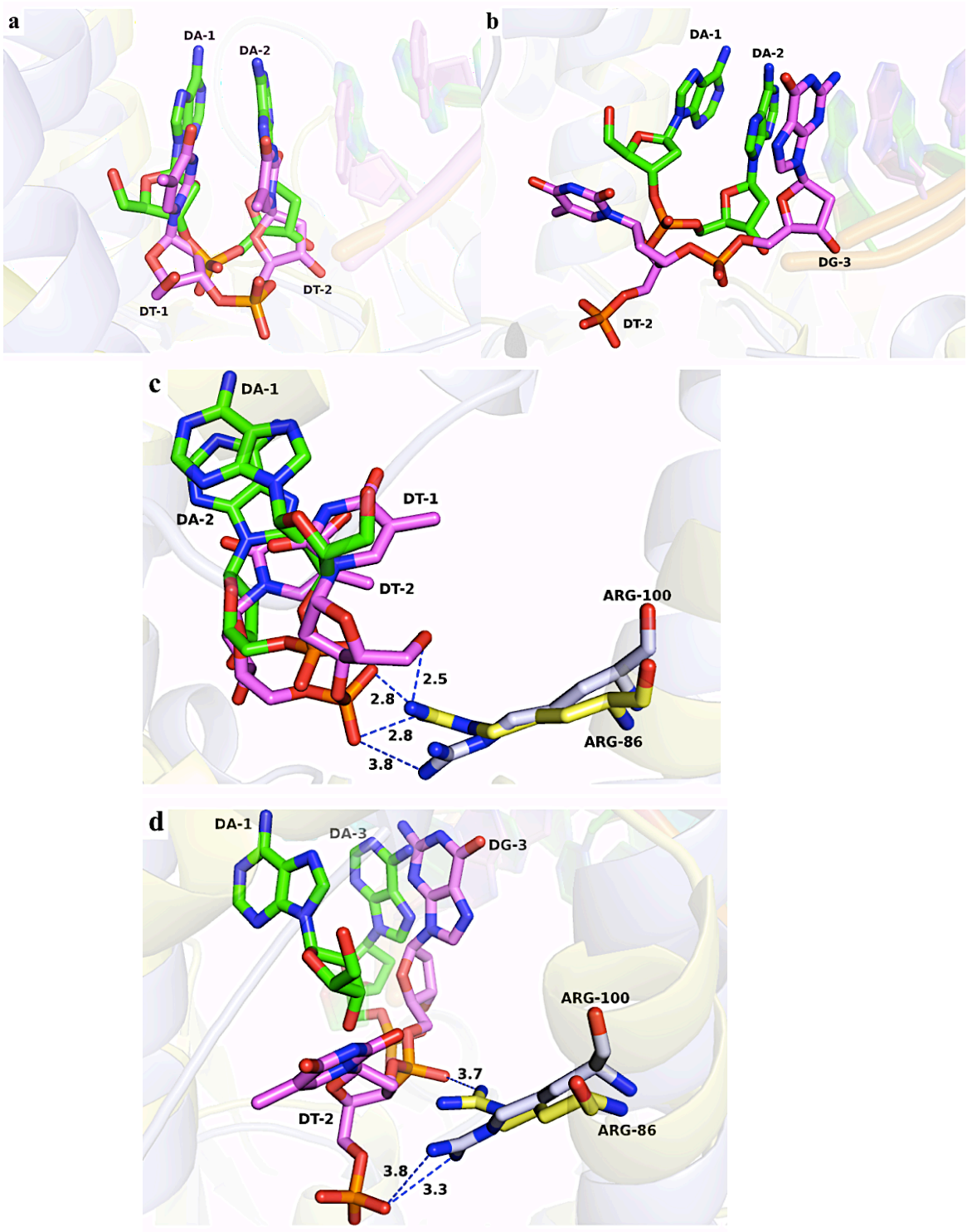



**Figure 5.75:** The DNA 5' overhang of C2/A (yellow protein and green DNA bases) superposed on hFEN:DNA 5' flap (white blue protein and magenta DNA bases) substrate and product.
**(a)** and **(b)** show the C2/A 5' overhang superposed on hFEN:DNA 5' flap substrate and product respectively. **(c)** and **(d)** illustrate Arg-86 in C2/A and Arg-100 in hFEN:DNA substrate and product respectively and their possible interactions with the DNA.

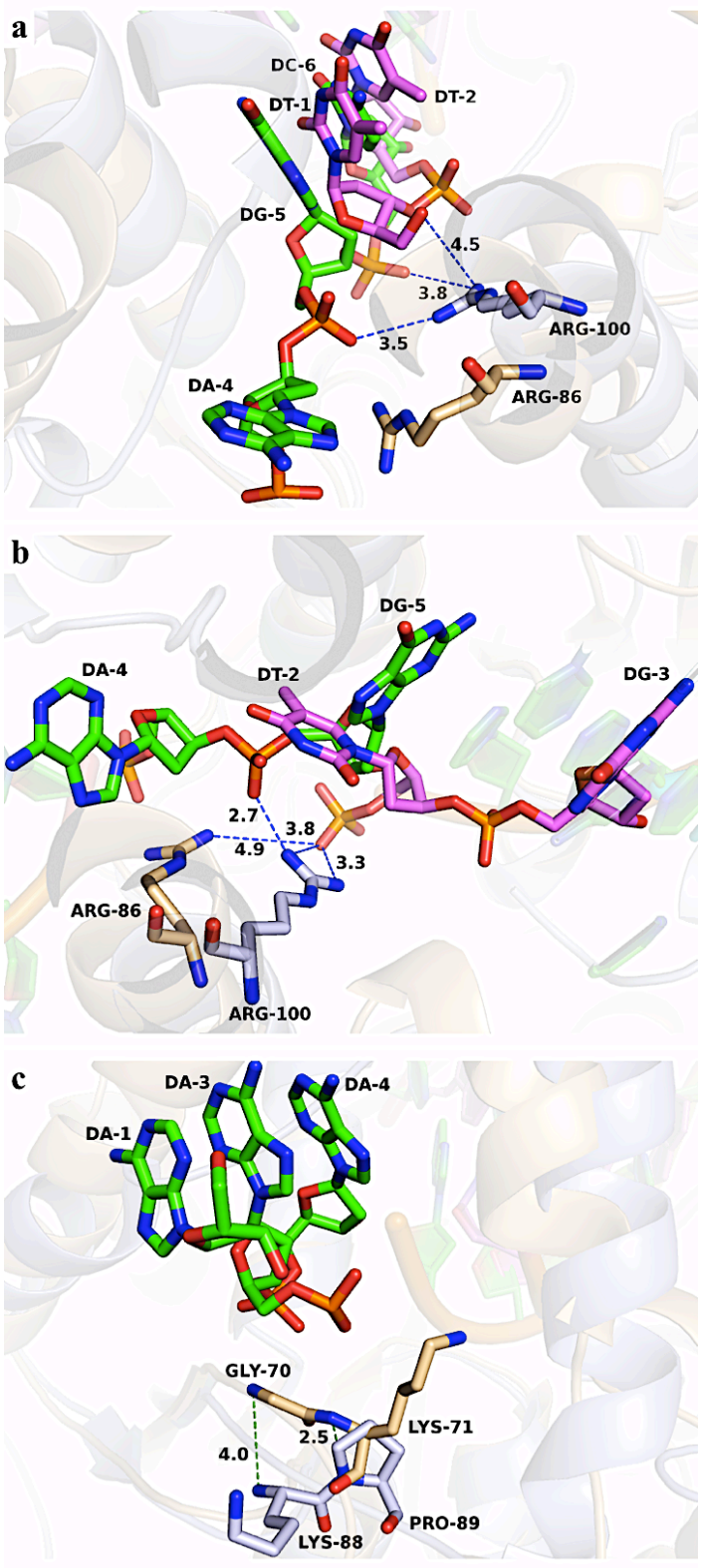



**Figure 5.76:** The arch-like structure of C2/B (wheat protein and green DNA bases) superposed on hFEN:DNA substrate and product structures (white blue protein and magenta DNA bases).
**(a)** and **(b)** show the possible interactions between C2/B-Arg-86 and hFEN Arg-100 with the DNA in substrate and product complexes respectively. **(c)** Shows superposition of the new trans-arch/distal DNA binding site of C2/B on hFEN:DNA

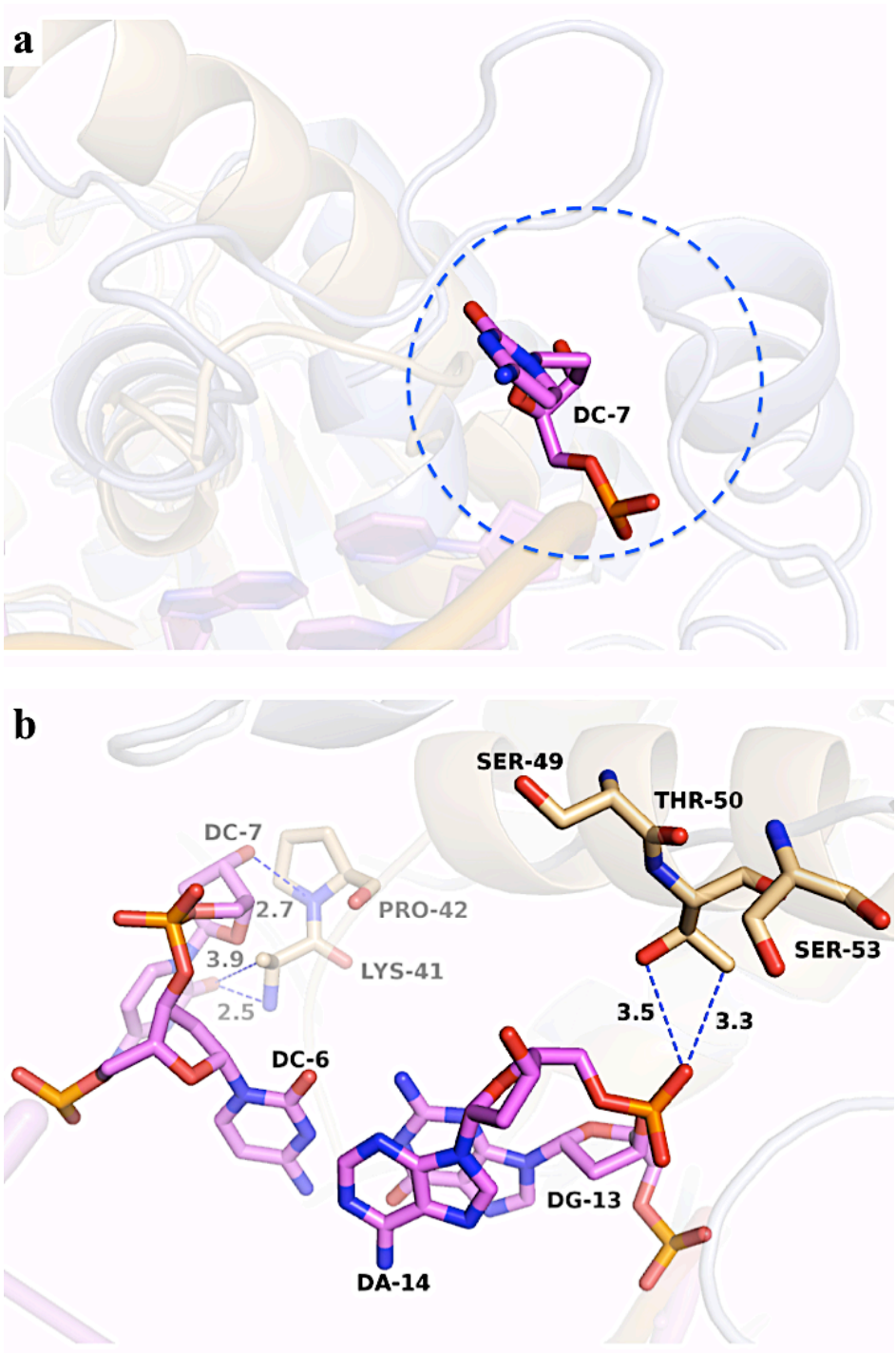



**Figure 5.77:** The 3' flap region in the superimposed structure of C2/B and hFEN:DNA.
**(a)** Shows the 3'-single nucleotide inserted in the 3' flap pocket in hFEN (blue dashed circle) which is absent from T5FEN. **(b)** Illustrates some possible interactions between C2/B protein and the 3' arm in hFEN:DNA.

## 5.3 Comparison of C2 with hEXO1:DNA Structure

Human Exonuclease-1 (hEXO1) is a second member of the FEN superfamily. It is also a metallonuclease that can presents both endo- and exo-nuclease reactions. It is included in this comparison as another example of a FEN superfamily complexed to a DNA substrate. The crystal structure of hEXO1 in a complex with ten base pairs duplex and 3' flap containing three-single bases was determined by Orans et al (2011) at a resolution of 3 Å. Two $Ba^{2+}$ metal ions were observed in the active site region and a potassium ion occupied a site within the H2TH motif (Orans et al, 2011).

### 5.3.1 Differences Between C2 and hEXO1 Structures

The two complexes: C2 and hEXO1:DNA superimpose well one on each other (Figure 5.15) with an average RMSD of 2.7 Å and 2.74 Å for C2/A and C2/B respectively on C$\alpha$s over a core of 194 amino acids. In general, there is again a high level of conservation at the main motifs or features which are a hallmark for the FEN family members including the H3/2TH:$K^+$ motif, the helix wedge, the helical arch and the active site. The superposition analysis shows good overlay between the two enzymes, and the DNA substrates in both complexes overlay fairly well one on each other (Figure 5.15). The majority of variations between these two proteins is a result from differences in the loop lengths, some insertions in some motifs and some extended parts in these proteins (Figure 5.16a). As it can be seen from this figure, the variations included the visible N-terminal with its $\alpha$-helix and $\beta$-sheet in the hEXO1 wich is removed from C2-T5FEN-D153K variant. In addition, about eight amino acids are added to the hydrophobic wedge in hEXO1 but are missing in C2. The next point at which these two structures significantly diverge is in the arch region which appears taller in hEXO1 while in C2 is likely to be more extended (Figure 5.16a).

Another dissimilarity between them concerns on an insertion region in the hEXO1 C-terminal (a.a. 289-346) composed of two $\beta$-strands, a helix and a loop region (Figure 5.16c) while the C-terminal presents as a small helix in the C2 protein (a.a. 272-291) (Figure 5.16c). Most of the residues in the

hEXO1 C-terminal are acidic amino acids as observed by the electrostatic surface map (Figure 5.16d) which may play a role in the processing of ssDNA nucleotides after threading through the arch if it will be the same mechanism in this enzyme as it is in T5FEN. The superposition of C2/B and hEXO1:DNA complexes showed the loop in the hEXO1 C-terminal approach from the threaded 5ov4 ssDNA. Furthermore, the H3/2TH motif fit very well in the two proteins (Figure 5.16e) as well as the $\beta$-pin motif which can be seen in Figure 5.16f with a small insertion in the loop region in hEXO1.

### 5.3.2 Superposed Analysis of the Active Sites

The superposed conserved active site amino acids overlay well in both enzymes (Figure 5.17a & 5.18a). The residues in hEXO1 active site present some shifts from the C$\alpha$s locations of the equivalent residues in C2 active site. Additionally, the two $Ba^{2+}$ ions from hEXO1 are shown in the Cat1 site of the two superimposed proteins: C2/A and C2/B and hEXO1. The first $Ba^{2+}$ ion is only 1.1 Å away from the mutated Lys-153 in C2/A active site (Figure 5.17a) which supports what we mentioned before about the presence of a new metal ion binding sit in C2/A active site (see section 4.2). In addition to the $2Ba^{2+}$ ions in the superimposed active site of C2/B and hEXO1:DNA it has also $2Mg^{2+}$ ions belongs to C2/B (Figure 5.18a). The first two metal ions ($Ba^{2+}$ and $Mg^{2+}$) in the hEXO1 and the C2/B are occupied M1 in Cat1 site. Unlike in hFEN the $Ba^{2+}$ ion is shifted by 2.7 Å from the $Mg^{2+}$ ion position to occupy a new metal ion site in Cat1 while the second $Ba^{2+}$ is in M2 of the Cat1 site (Figure 5.18a). The latter metal ion is 4.6 Å apart from the first $Mg^{2+}$ and 2.4 Å away from the mutated amino acid Lys-153 which is also varied from what have been seen in the native T5FEN complexed with $Mg^{2+}$ ions and in the hFEN:DNA complex (Figure 4.23a and Figure 5.9 respectively). C2/A Lys-153 that was formed a hydrogen bond to the 5ov4 DNA backbone, dA-2/X, as mentioned before is also show possibility to interact with the equivalent dC-2/B phosphate oxygen in hEXO1 complex through hydrogen bond (Figure 5.17b).

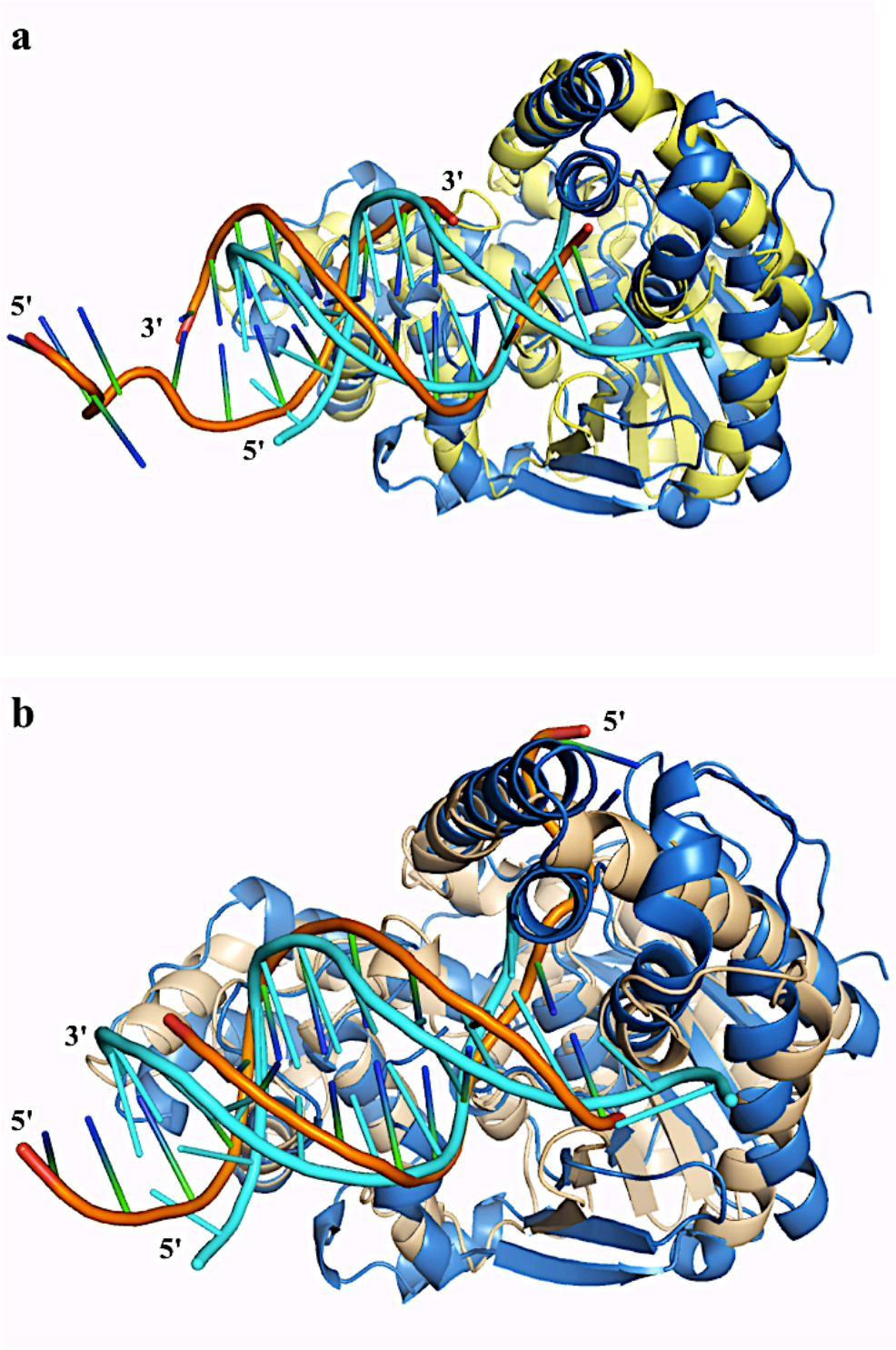



**Figure 5.78:** Superposition of C2 structure on hEXO1 (blue) complexed to the DNA substrate (cyan).
**(a)** and **(b)** show the overall fold of C2/A (yellow protein, orange DNA backbone and green bases) and C2/B (wheat protein, , orange DNA backbone and green bases) respectively on hEXO1:DNA.

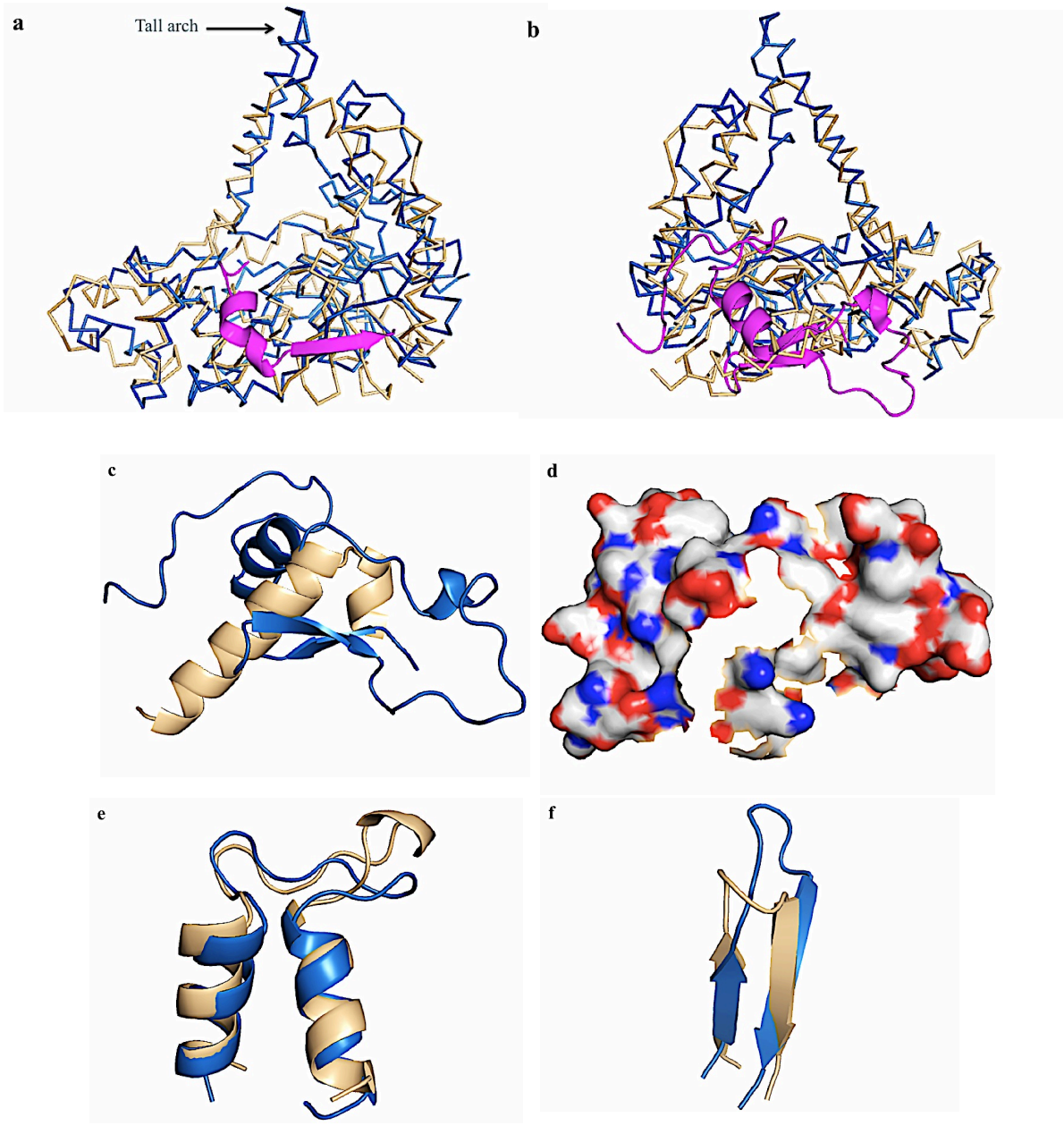



**Figure 5.79:** Differences between C2-T5FEN (wheat) and hEXO1 (blue).
**(a)** and **(b)** ribbon diagrams illustrate the visible N-terminal (magenta) in the hEXO1 which is removed from T5FEN and cannot be seen and the extended C-terminal (magenta) in the hEXO1 that contains two $\beta$-strands, a helix and a loop respectively. **(c)** Shows the T5FEN C-terminal superimposed on hEXO1 C-terminal. **(d)** The electrostatic surface for hEXO1 C-terminal. **(e)** and **(f)** show the H3/2TH and the $\beta$-pin motifs respectively.

The second $Mg^{2+}$ ion is found in the C2/B Cat2 site and separated by 6.3 Å and 9.1 Å from the first and the second $Ba^{2+}$ ions respectively in hEXO1 (Figure 5.18a). In contrast, the scissile phosphate in the 5ov4 (dC-6/Y) of C2/B appears to be too close to the hEXO1 metal ions which is about 1.7 Å and 2.8 Å apart from the first and the second $Ba^{2+}$ ions respectively. The second $Mg^{2+}$ ion which occupies the C2/B Cat2 site is able to interact with the dG-3/B phosphate group in hEXO:DNA complex which is 1.8 Å away from it and also interacts with the equivalent superposition dG-7/Y backbone in C2/B through hydrogen bonds (Figure 5.18b). Furthermore, the phosphate group in the DNA substrate complexed with hEXO1, dC-2/B, is 4.1 Å and 4.2 Å from the first and the second $Ba^{2+}$ respectively and it is further far (4.4 Å) from the first $Mg^{2+}$ ion (Figure 5.18b).

## 5.3.3 Interactions of the DNA

### *5.3.3.1 Interactions of the H3/2TH Motif and $K^+$ Binding Site*

In both enzymes the H3/2TH motif interacts with the duplex regions and these interactions are made by the equivalent amino acids in the two proteins to the DNA substrate backbone as described numerously in sections 4.1.2.2.1 and 1.2.1.1.1 for the C2/A and C2/B and the hEXO1:DNA complexes respectively. A potassium ion ($K^+$) site is present in all structures, C2/A, C2/B and hEXO1:DNA. hEXO1 $K^+$ ion is about 4.3 Å apart from the one in C2/A (Figure 5.19a) and 3 Å away from C2/B $K^+$ ion in the first position (Figure 5.19b) when they superposed. The main chain carbonyl of Met-199, Val-209 and Ile-212 coordinate the $K^+$ ion in C2-T5FEN:DNA complex which are equivalent to Ser-222, Leu-230 and Ile-233 respectively in hEXO1:DNA. The same situation in hFEN:DNA, Leu-230 in hEXO1 which is equivalent to Val-209 in C2 is far away from the $K^+$ ion and cannot interact with it but Ser-229 side chain form hydrogen bond to $K^+$ ion (Figure 5.19). Gly-208 in C2 structure which is equivalent to Leu-229 in hEXO1 is far away from the potassium ion in C2 and cannot interact with it and again Val-209 engage in $K^+$ ion coordination in C2-T5FEN-D153K:DNA structure (Figure 5.19).

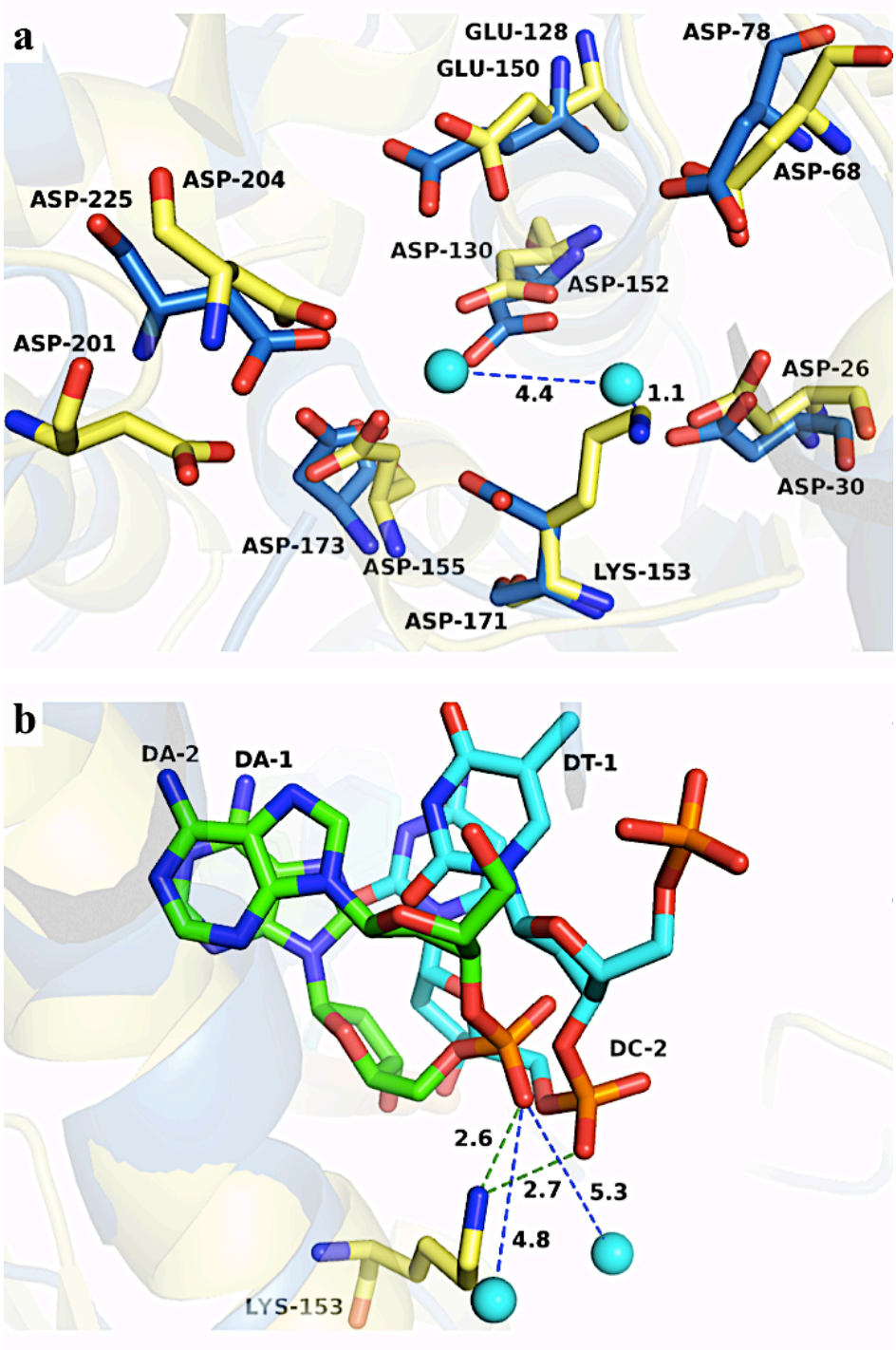



**Figure 5.80:** Superposition of C2/A active site on hEXO1:DNA.
**(a)** Shows the active site residues (yellow sticks) with the equivalent residues in hEXO1 (blue sticks). Two $Ba^{2+}$ ions (cyan spheres) in hEXO1 are present in the superimposed active sites. **(b)** Shows Lys-153 in C2/A (yellow stick) is able to interact with the DNA phosphate group of hEXO1:DNA (cyan bases, green dashes) while the two $Ba^{2+}$ are far away from 5ov4 DNA (green bases, blue dashes).

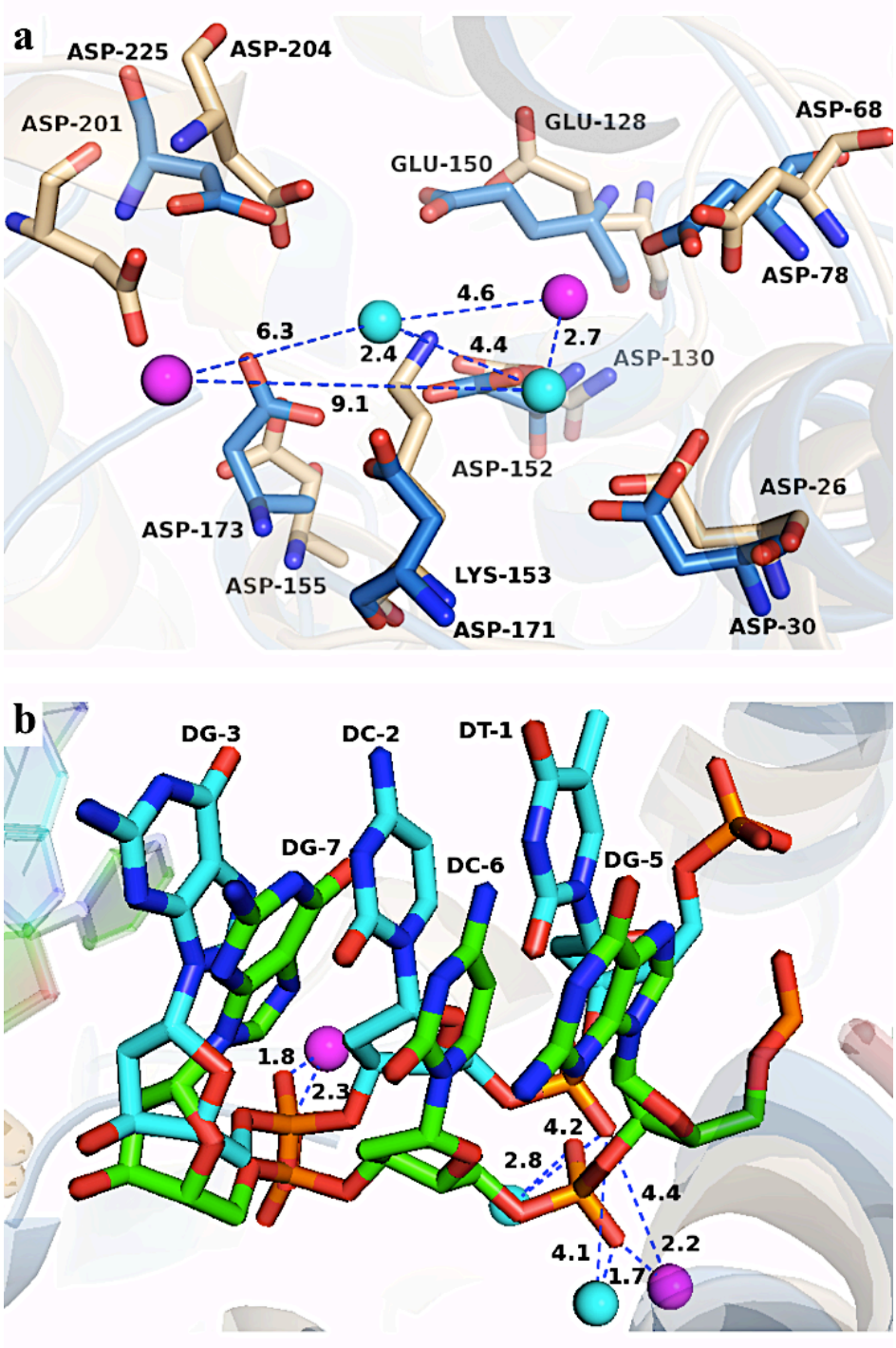



**Figure 5.81:** Superposition of C2/B active site on hEXO1:DNA.
**(a)** Shows the C2/B active site residues (wheat sticks) with the equivalent residues in hEXO1 (blue sticks). Four metal ions are present in the superimposed active sites: $2Ba^{2+}$ ions (cyan spheres) for hEXO1 and $2Mg^{2+}$ (magenta spheres) for C2/B. **(b)** Shows the distances and the possible interacts of the metal ions with the DNA phosphate group of hEXO1:DNA (cyan bases) and 5ov4 DNA (green bases).

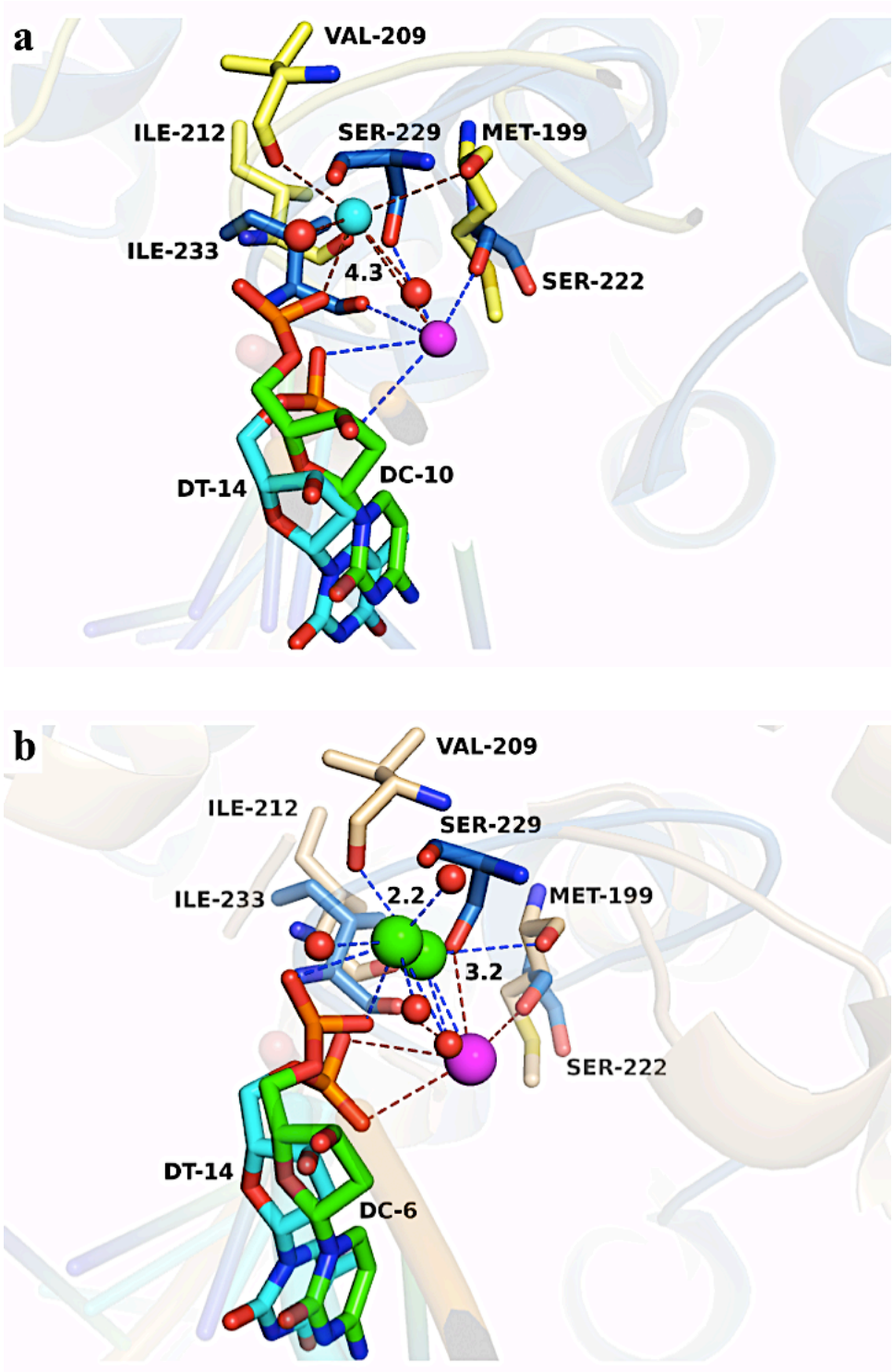



**Figure 5.82:** Superposition of the potassium binding site of C2 on hEXO1:DNA (blue).
**(a)** Shows the two $K^+$ ions (cyan sphere in C2/A and magenta sphere in hEXO1) as present in the superposed structures and they are ~ 4.3 Å apart from each other. **(b)** Illustrate the $K^+$ ions in C2/B (green spheres) in its two positions and the hEXO1:DNA potassium ion (magenta sphere).

#### *5.3.3.2 Interactions With the Helix Wedge*

The majority of interactions of the C2/A lie in the H3TH:$K^+$ motif (see section 4.1.2.2). The 3' end (strand-Y) duplex of the 5ov4 in C2/A does not reach to the hydrophobic wedge and does not interact with this part of the enzyme. In C2/B and hEXO1:DNA the 3' terminal base pairs of the duplex in the two complexes are stacked against the hydrophobic wedge. More explanation for the interactions that are made in the hydrophobic wedge between C2/B and the 5ov4 3' duplex end (strand X) can be found in section 4.1.2.2.3 while section 1.2.1.2 described the interactions of hEXO1:DNA substrate in the same area.

#### *5.3.3.3 Interactions of the 5' Flap*

The DNA substrate complexed to hEXO1 lacks the 5' flap but has three-single nucleotides in its 3' flap region. On the basis of this fact, the interactions of the 5' flap in hEXO1:DNA complex with the arch amino acids cannot be determined. However, some important interactions with the arch residues can be analyzed and investigated by comparison with C2 structure. The superposition of C2 and hEXO1:DNA shows Arg-92 in hEXO1 which interacted with the DNA backbone is close to the equivalent Arg-86 in C2/A (Figure 5.20a) and Arg-86 could interact with the dC-1/B phosphate oxygen in hEXO1:DNA complex through its guanidinum group.

In contrast, this residues in C2/B is seen far away from the superimposed DNA in hEXO1 complex as well as its equivalent Arg-92 in the hEXO1:DNA which is proposed to interact with the DNA phosphate group is found away from the superimposed 5ov4 5' overhang ( Figure 5.20b) which could be referred to the absence of the real 5' flap in the hEXO1:DNA structure and the movement of C2/B Arg-86 after threading process from its original position to a new conformation .

#### *5.3.3.4 The trans-arch/ distal DNA Binding Site*

In the C2 structure this site lies in a loop behind the arch and interacts with the DNA 5' overhang backbone in C2/B (see 4.1.2.2.4) while in the hEXO1:DNA crystal structure the equivalent amino acids to generate this site are cys-80 and Thr-81. These residues are shifted by 1.94 Å and 1.06 Å respectively from their C$\alpha$s equivalents in C2/B and the distance between the main chain amide nitrogen of hEXO1 Cys-80 and Thr-81 are 2.4 Å and 1.0 Å from their equivalent in the T5FEN (Figure 5.20c). The absence or the presence of this site in the hEXO1 can be explained as above with the hFEN (see section 5.2.1.4.2).

#### *5.3.3.5 Interactions of the 3' Arm*

As described above the C2-T5FEN-D153K:DNA does not have a 3' flap, but one is present in the hEXO1:DNA structure. The single 3'-nucleotide binding pocket that consists of helix-hairpin-helix (HhH) and is present in FEN family members except for the bacteriophage FENs is also structurally missing in the hEXO1 structure (Figure 5.21a). Additionally, C2/B His-36 could form hydrogen bonds from its main and side chains to the dC-21/A oxygen-2 and the dA-22/A phosphate oxygen in hEXO1 complex (Figure 5.21b).

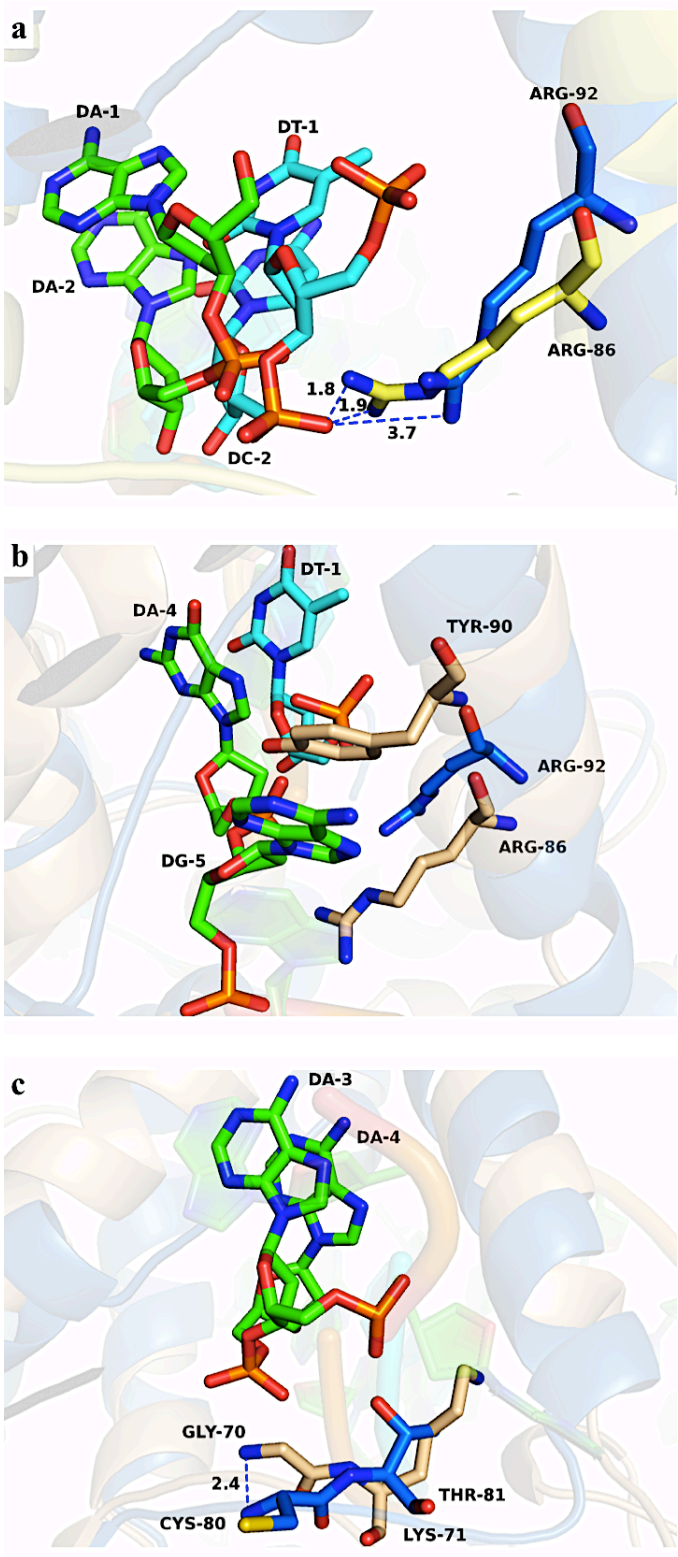



**Figure 5.83:** Superposition of C2 5' overhang region on hEXO1:DNA.
**(a)** and **(b)** show Arg-86 in C2/A (yellow stick) and C2/B (wheat stick) respectively with their equivalent Arg-92 (blue stick) in hEXO1 and the possible interactions between them and the DNA substrates. **(c)** Illustrates the new DNA binding site in C2/B (wheat sticks) and its equivalent residues in hEXO1 complex.

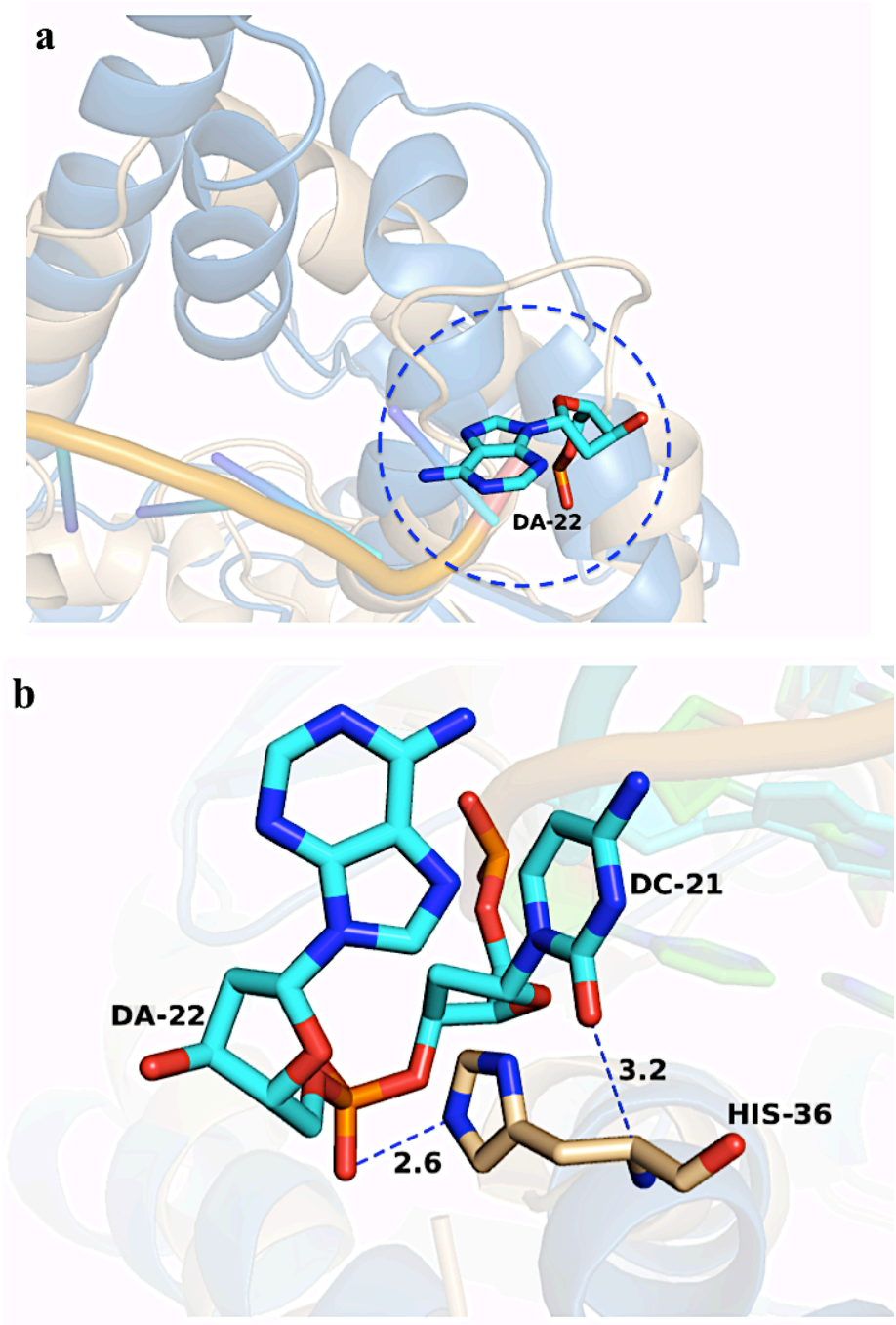



**Figure 5.84:** The 3'-arm superposition and possible interactions in C2 and hEXOIX:DNA.
**(a)** Shows the absence of the 3'-single nucleotide pocket from hEXO1 and C2 enzymes (blue dashed circle). **(b)** Illustrate some interactions could happen between the 3' arm substrate (cyan) of hEXO1:DNA and C2/B His-36 (wheat stick).

# Chapter 6: T5FEN-D155K and WT in Complexes With 3' overhang DNA

T5FEN-D155K and the wild type enzyme were successfully crystallized in complexes with a duplex DNA oligonucleotide containing 3' overhang ends. During the analysis of these structures, T5FEN-WT enzyme will be referred to, as CW and the other two T5FEN-D155K variants as C3 and C4 as mentioned in chapter three. The last two structures differed in one was obtained with calcium ions and also had a potassium ion in its complex while the other one was devoid of metals in the active site but still has a $K^+$ ion in its binding site. The latter structure, C4, showed a $Na^+$ ion interacting with the DNA nucleotides but away from the active site. A structure comparison between T5FEN-D155K complexed with 3ov6 DNA and other FEN members in complexes with DNA also are discussed.

## 6.1 Results

### 6.1.1 Overall Structures

The three structures of T5FEN enzymes in complexes with 3ov6 overhang DNA (C3, C4 and CW) were solved in the space group $P4_32_12$ at 1.86 Å, 2.2 Å and 2.9 Å resolution respectively. There is one chain of the T5FEN enzyme joined to one copy of the 3ov6 DNA in the asymmetric unit (Figure 6.1). Symmetrically, a palindromic duplex is generated by two-fold axis binding to two molecules of T5FEN which reveal the assembly system of the crystal contents (Figure 6.2). The single molecule of the T5FEN protein is present with the familiar architecture for the T5FEN-WT as described in section 1.3.2. Each one of these molecules is built by amino acids 20-291 with some missing of side-chains in each of which due to the poor or totally absent of the density in these locations. In C3 the fully truncated side-chains are such as Arg-20, Glu-96, Glu-102 and partially like Asn-37, Glu-97, Lys-97, Lys-109, Glu-113, Arg-207 while for C4 the side-chains of some amino acids could not be readily traced such as those for Arg-20, Asn-37, Lys-79 and Leu-202 or were only partially visible such as in Glu-102, Glu-251 and

Lys-273. Additionally, the number of the side-chains that could not be built was higher in CW structure which is likely due to the low resolution of this crystal structure (2.9 Å) as mentioned above. These residues with untraceable side chains include Lue-22, Asn-37, Lys-40, Gln-52, Lys-56, Gln-92, Glu-95, Glu-96, Lys-97, Leu-100, Glu-106, Lys-109, Glu-113, Leu-114, Lys-116, Ile-140, Leu-143, Asp-162, Ile-193, Lys-196, Leu-202, Ile-221, Leu-253, Phe-254, Val-277, Lys-273, Glu-277 and Gln-291 while the amino acids with partially defined side chains include Lys-73, Lys-79, Lys-137, Leu-159, Arg-207, Glu-223, Ile-230, Gln-233, Lys-241 and Leu-257.

### 6.1.2 3ov6 DNA Interactions

The sequence of the 3ov6 oligonucleotide was meant to be derived from that of the 5ov4 overhang DNA and was designed to have a long 5' overhang. It consists of ten base pairs with some insertions (2dGs) and some mismatches (2 T-C, Figure 3.13c), symmetrically, as described in section 3.3.3.2.4. During preparation process of this substrate for mixing with the T5FEN protein it can create a 3' overhang instead of 5' overhang and crystalized in a complex with T5FEN as 3' overhang DNA but it is unclear whether the 3' overhang forms prior to complex assembly with T5FEN or if it is formed during the crystallization process. The duplex DNA interacts with the T5FEN chains in three sites similar to what have been observed with C1 and C2 of T5FEN-D153K in complexes with 5ov4 DNA (see chapter 4).

#### *6.1.2.1 Interactions with the H3TH:$K^+$ Motif*

Four amino acids interact with the DNA substrate in the H3TH motif directly by hydrogen bonds while other three residues form indirect interactions through a potassium ion. The main-chain amide nitrogen of Gly-211, Gly-213, Lys-215 and Arg-216 forms hydrogen bonds with the DNA backbone, dC-4 and dT-5 (Figure 6.3).

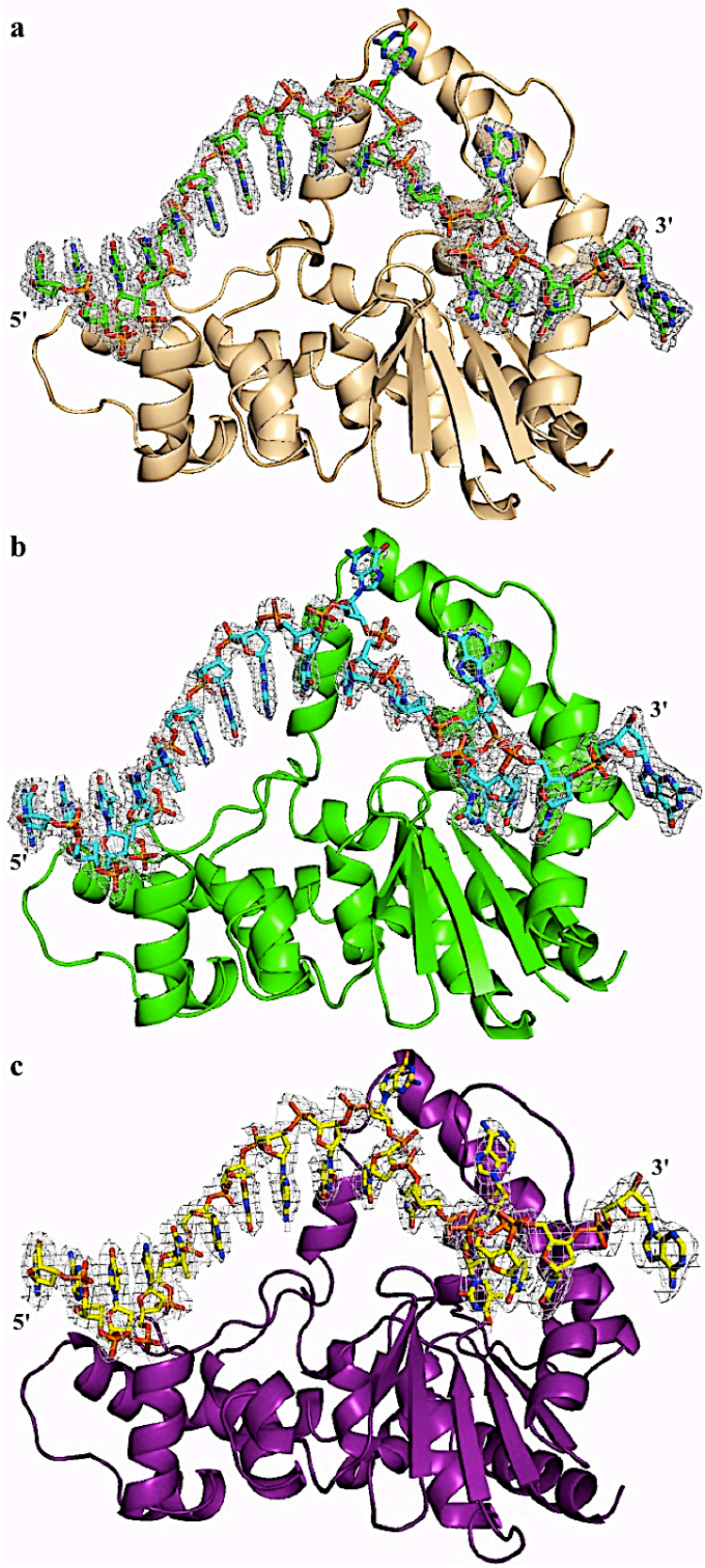



**Figure 6.85:** The crystal structures of T5FEN in complexes with 3ov6 DNA. **(a)** and **(b)** are C3 and C4 of D155K variant while **(c)** is CW. The DNA substrate presents in 2Fo-Fc electron density map in all of the three structures.

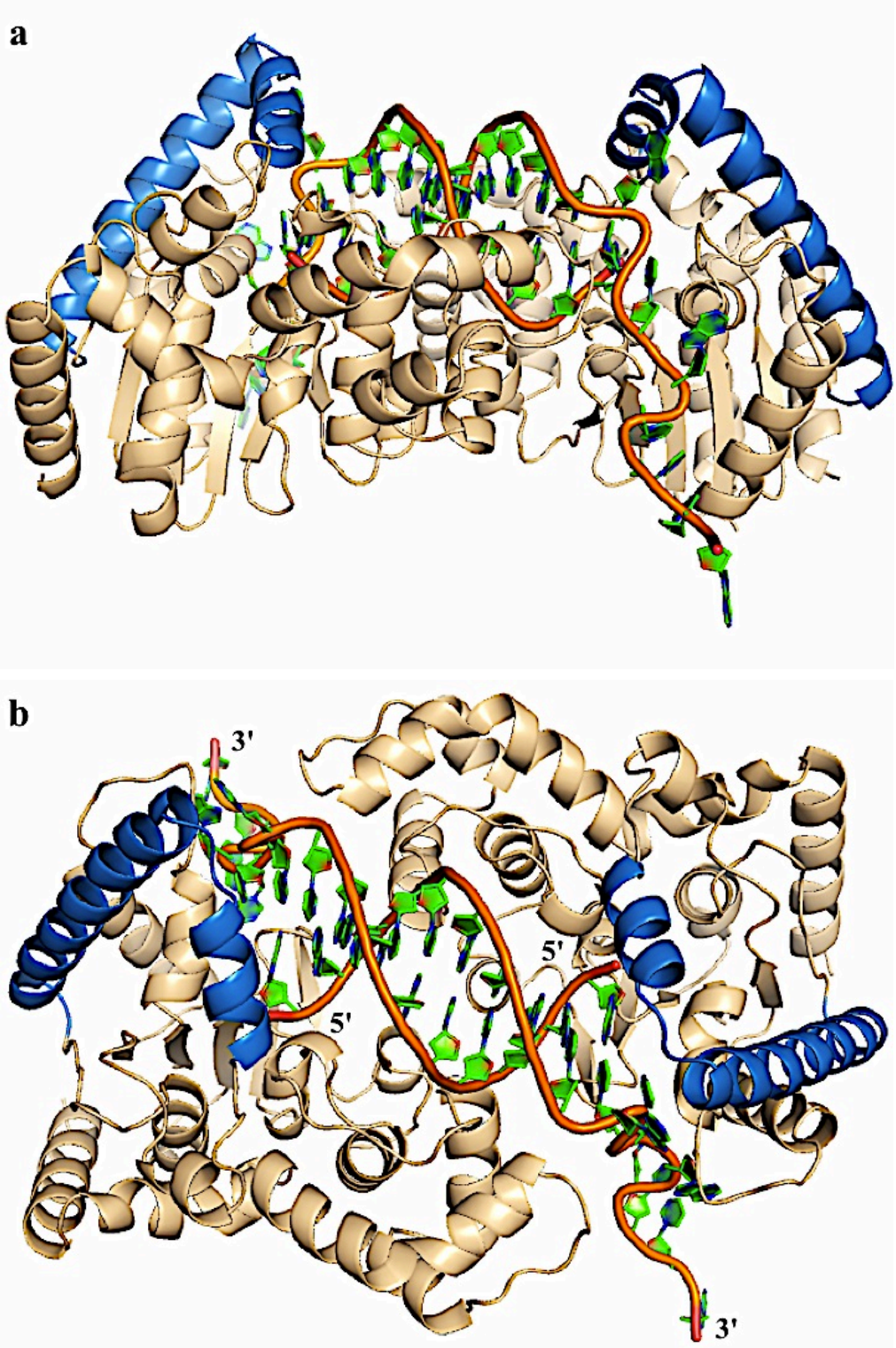



**Figure 6.86:** Shows the two-fold axis crystallography to generate a palindromic duplex of DNA between two molecules of T5FEN.
**(a)** Is a side view for the complex while **(b)** is an "overhead" view. The helical arch presents in blue while the other protein parts in wheat.

Additionally, the $K^+$ ion, which is revealed in C3 and C4 structures but not in CW, forms indirect interactions between the Met-199, Val-209 and Ile-212 main-chain carbonyl oxygens and dT-5 phosphate oxygens (Figure 6.3a & b and 6.4a & b).

#### *6.1.2.2 Interactions with the Hydrophobic Wedge*

The 3ov6 substrate lacks a 5' overhang and consequently its path tracked away from the active site. The end of the symmetric palindromic duplex stacks against the hydrophobic wedge (Figure 6.5a). Tyrosine 90 is hanging down again from the arch in this structures and stacks against dT-16. Additionally, Tyr-90 moves closer to Phe-105 (~ 3.7 Å) relative to its position in DNA-free enzyme to create a gate-like structure (Figure 6.5b).

#### *6.1.2.3 The 3' Overhang Interactions*

Indirectly, a network of interactions is formed between each enzyme molecule and the 3' overhang of the DNA oligo by hydrogen bonds *via* water molecules. Three hydrogen bonds are formed by three residues, Arg-33, Arg-172 guanidinium groups and Phe-169 main chain carbonyl oxygen, to dA-13 phosphate oxygens using three water molecules. (Figure 6.6a). Another notable interaction is composed of a hydrogen bond between Phe-169 main-chain carbonyl oxygen and the dT-14 deoxyribose ring while the hydroxyl group of Tyr-58 interacts with dT-14 oxygen-4 *via* a water (Figure 6.6a). Further interactions are formed by hydrogen bonds through five waters to dT-14, dC-15 and dT-16 phosphate oxygens (Figure 6.6a). In this DNA substrate, 3ov6, guanosine-10 is clamped by the arch and interacts by its base nitrogen-2 with hydroxyl group of Thr-94 (Figure 6.6b) while dA-13 loops-up and forms hydrogen bonds to the hydrophobic wedge residues side-chains. His-36 which is rotated by about 180° from its original position in T5FEN structure (1EXN) form hydrogen bonds between its side-chain and the dA-13 deoxyribose oxygen-4 and nitrogen-9 (Figure 6.6b) while Asn-38 side chain interact with dA-13 nitrogen-6 (Figure 6.6b).

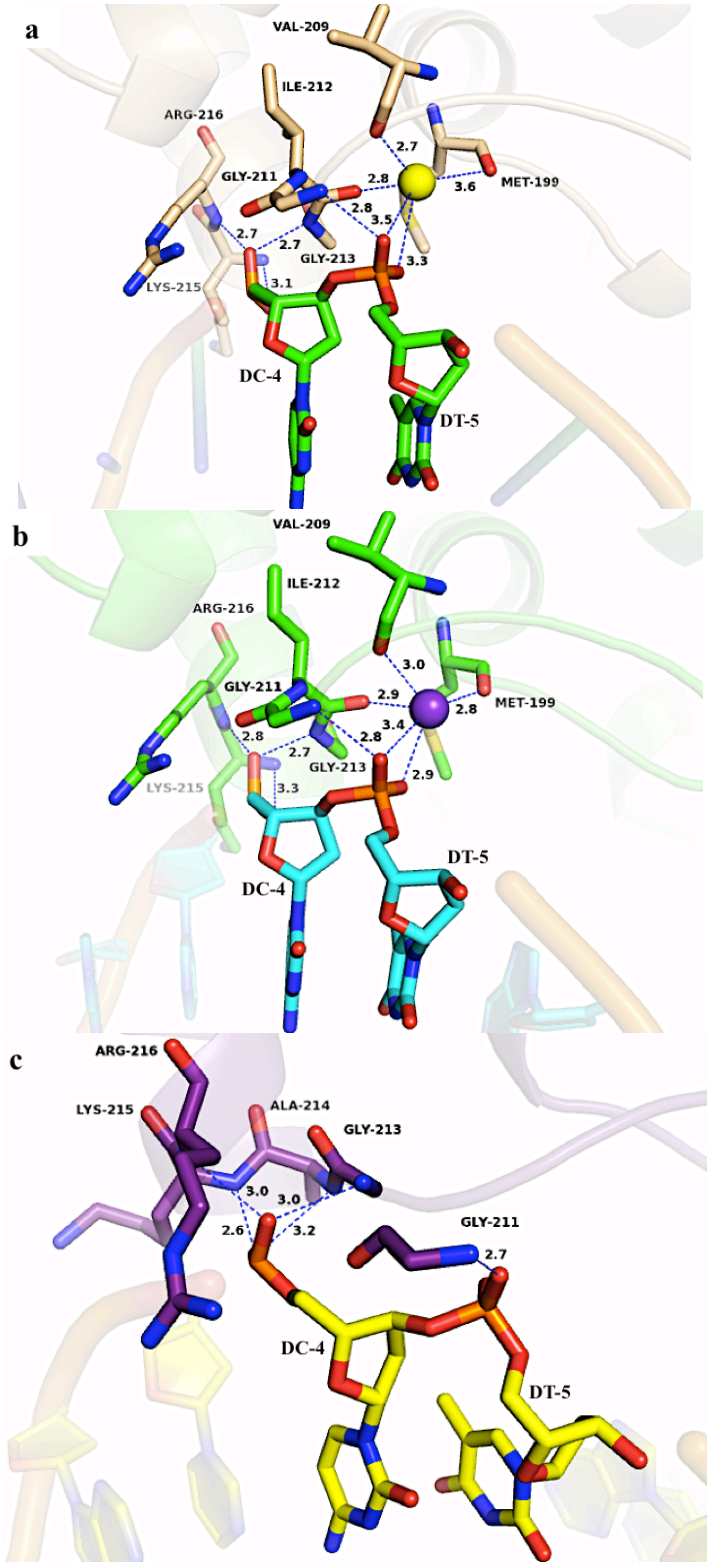



**Figure 6.87:** The H3TH motif interactions.
**(a), (b)** and **(c)** show these interactions in C3, C4 and CW structures respectively. A potassium ion presents in C3 (yellow sphere) and C4 (purple sphere) while absent from CW structure.

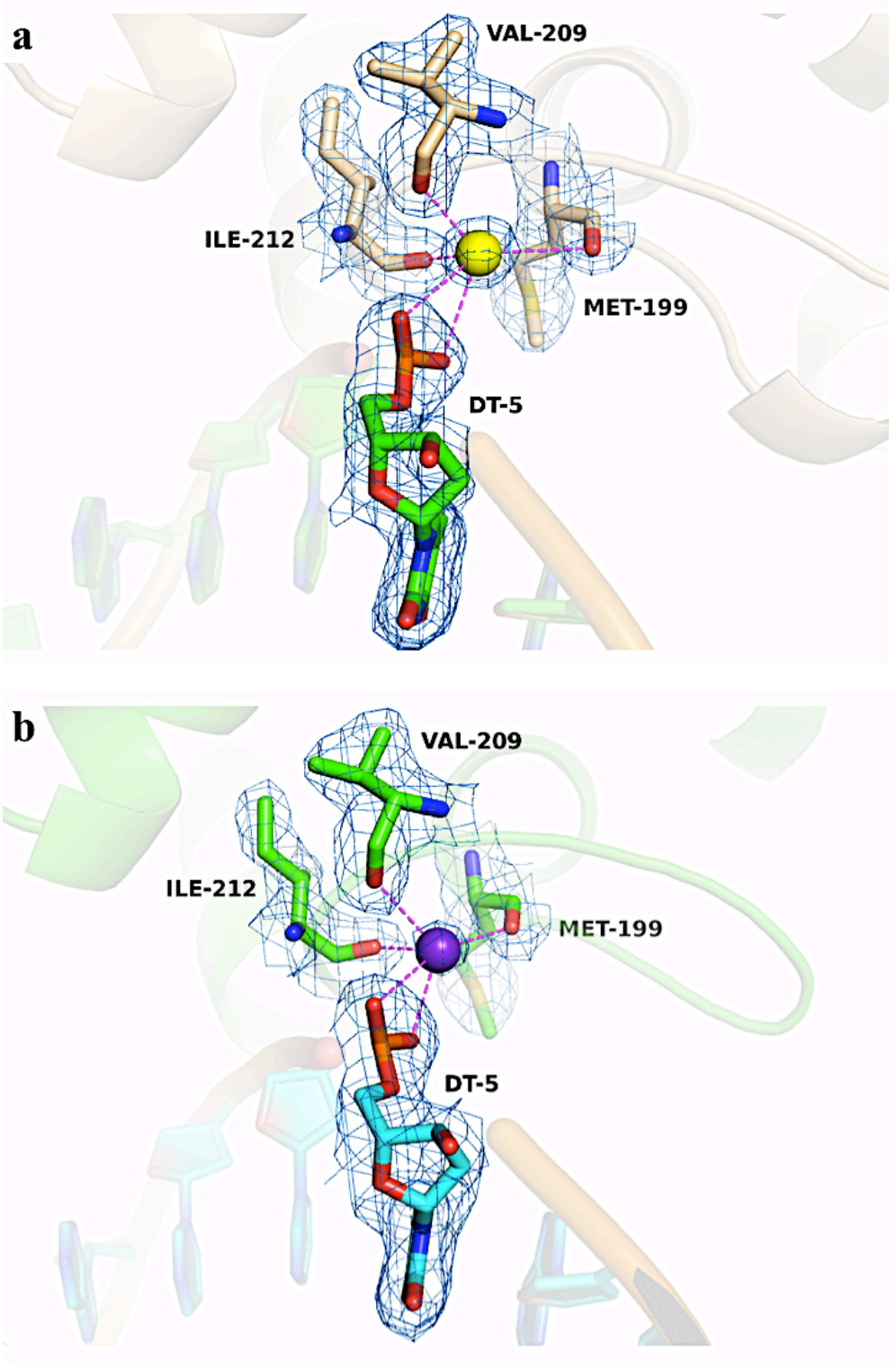



**Figure 6.88:** The potassium-binding site in C3 and C4 structures.
**(a)** and **(b)** show 2Fo-Fc electron density map for the $K^+$ ion in its binding site in C3 and C4 structures respectively. This ion coordinates with Met-199, Val-209, Ile-212 and dT-5 in each complex.

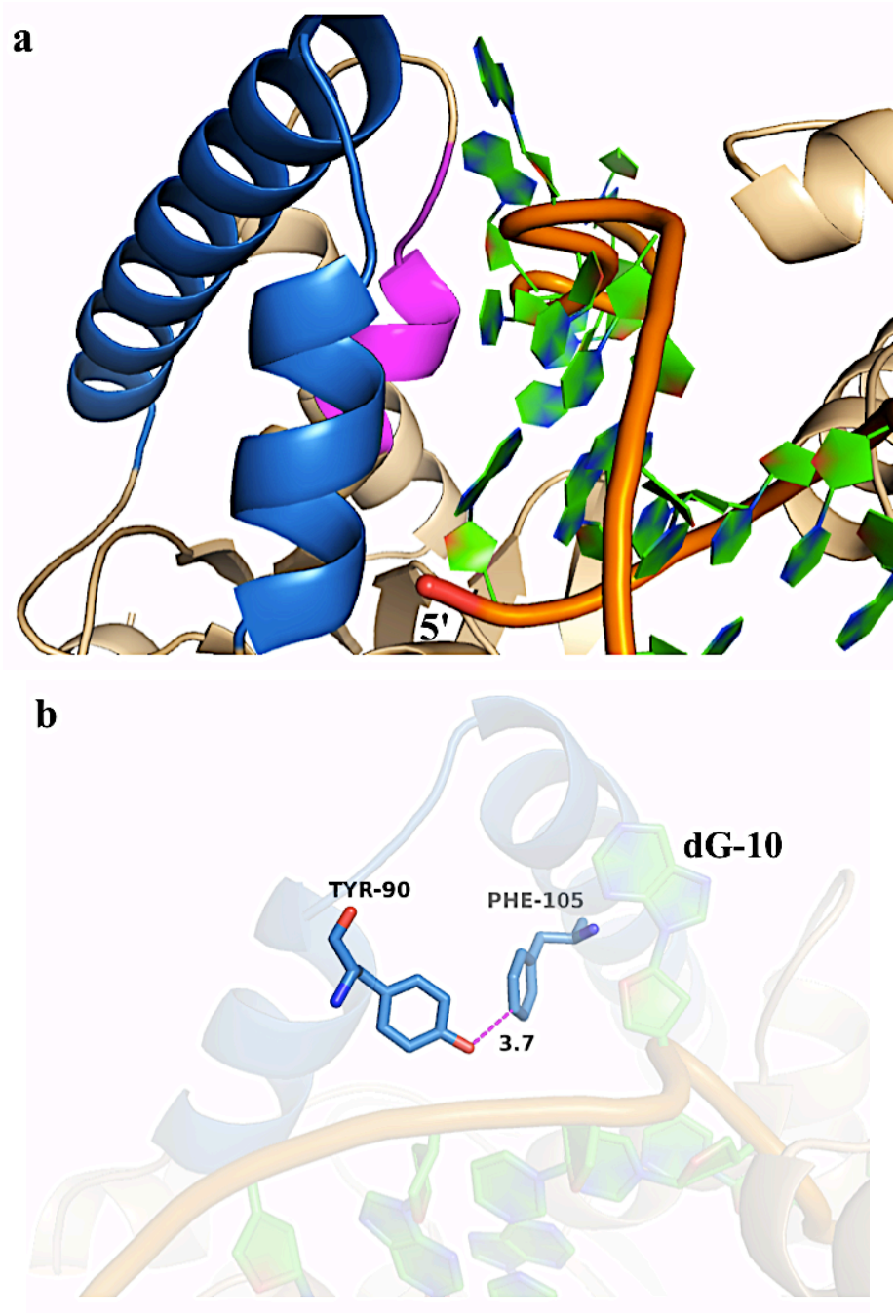



**Figure 6.89:** The hydrophobic wedge and the gate-like structure in T5FEN-D155K:DNA structures.
**(a)** Illustrates the hydrophobic wedge (magenta) stack against the DNA duplex. **(b)** Shows the gate-like structure formed by the arch residues Tyr-90 and Phe-105 (blue sticks). The helical arch presents in blue while the other protein parts in wheat.

*Via* crystal symmetry, the 3' overhang threads through the arch of a neighboring molecule and its dG-17 faces the dG-1 of the 5' in this duplex (Figure 6.7a & b). Interestingly, these two nucleotides pass above the active site of the neighboring chain and dG-17 deoxyribose is 4.1 Å apart from the $Ca^{2+}$ ion in this active site (Figure 6.8a). Lys-83, Arg-86 and Arg-125 of the neighboring protein molecule interacts with the DNA backbone by forming hydrogen bond from their amino groups to the dG-17 and dC-15 phosphate oxygen (Figure 6.8b). The dG-17 phosphate group at the end of the 3' overhang is able to form crystal contacts with the distal site of a neighboring D155K molecule (Figure 6.8b). Hydrogen bonds are formed from Gly-70 main chain amide nitrogen to a phosphate oxygen on dT-16 and from Lys-71 side chain to dT-14 backbone (Figure 6.8b).

### 6.1.3 The Active Site and Metal Ions

The T5FEN-D155K protein was prepared by Jing Zheng to prevent cations from binding to the M2 site (recently called M3 in Cat2, (Anstey-Gilbert et al, 2013)) by replacing Asp-155 with Lys-155. This results in a protein that can bind DNA but is catalytically inactive (Zhang, 2012). Theoretically, this version of the protein cannot bind to metal ions at M3 of Cat2 site which should be occupied by the lysine side chain. Structurally, Lys-155 does not occupy M3 but instead its long side-chain is rotated by $\chi_1$ 120 and presents in which could be a new metal binding site in M2 as that one determined previously in C2/A (see section 4.1.2.3). However, two calcium ions were found in C3 of T5FEN-D155K:DNA. The first ion occupies M1 of Cat1 site and is coordinated by four water molecules and two conserved active-site carboxyl residues: Asp-130 and Glu-128 (Figure 6.9a). The water molecules are hydrogen bonded to three active site residues: Asp-26, Asp-68 and Asp-153. The second $Ca^{2+}$ ion is far away from the active site and interacts with dC-15 in the 3' arm overhang and is coordinated by the base oxygen-2 and seven water molecules (Figure 6.10a & b). The metal ions are absent from the active site of C4 and CW structures (Figure 6.9c & b) but a metal ion is observed close to the dC-4 and dT-5 bases in C4 structure (Figure 6.10c &d).

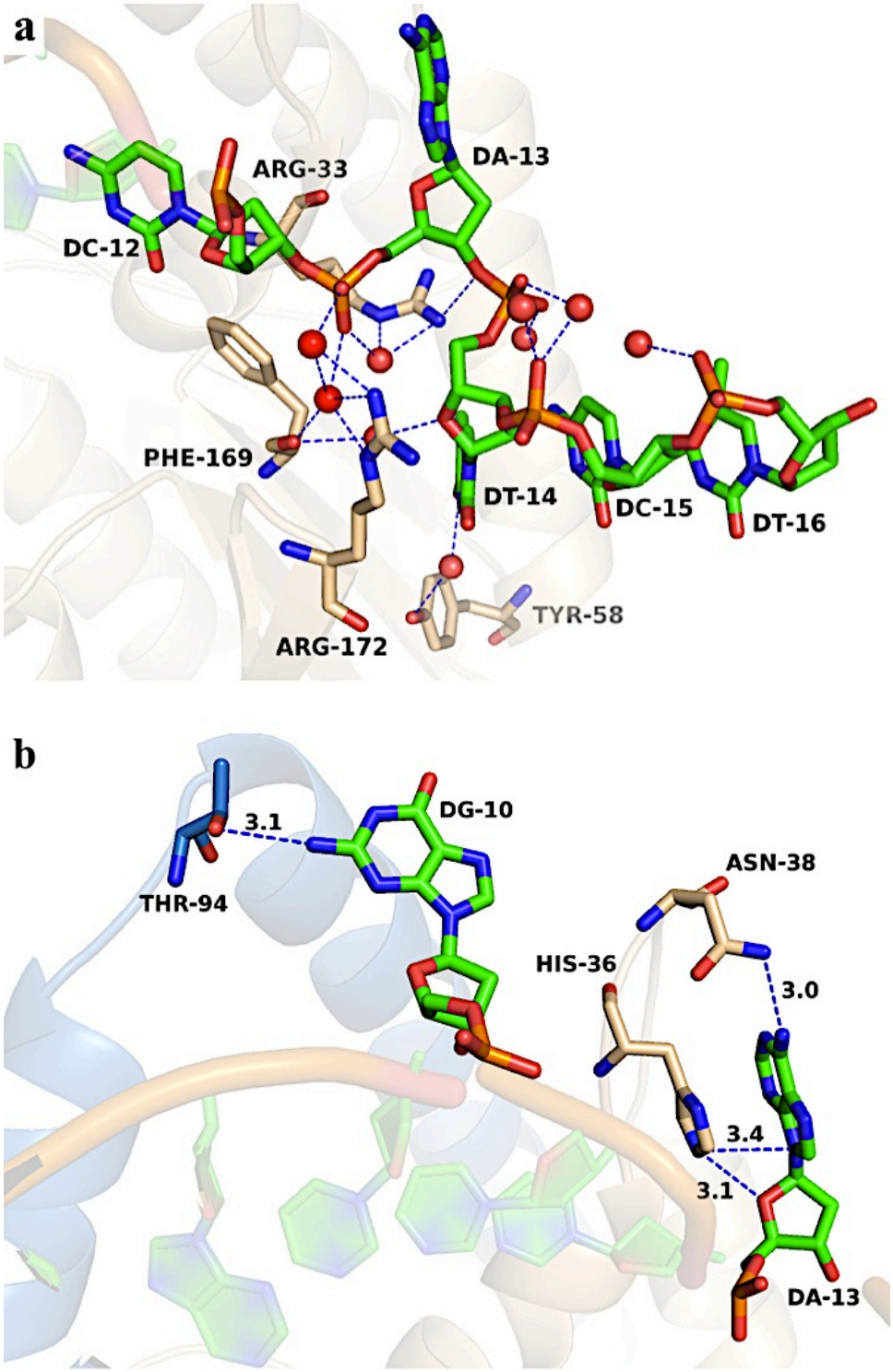



**Figure 6.90:** The interactions between T5FEN-D155K and the DNA 3' overhand. **(a)** Shows indirect interactions of Arg-33, Arg-172, Phe-169 and Tyr-58 with dA-13 and dT-14 *via* water molecules and network interactions between the DNA backbone and water molecules. **(b)** Illustrates the direct interactions of residues: His-36, Asn-38 and Thr-94 with dG-10 and dA-13.

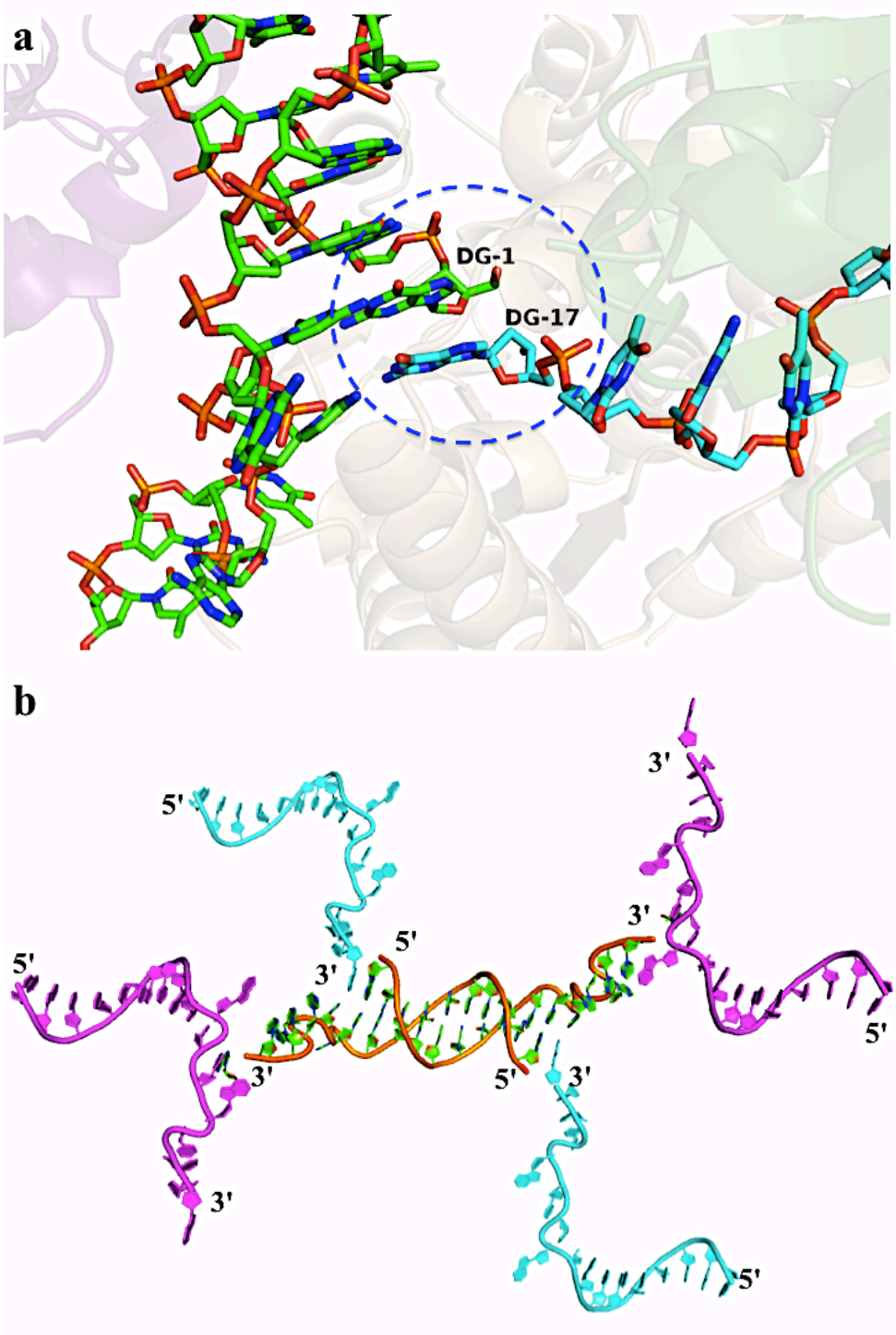



**Figure 6.91:** Passing through the neighboring molecule arch and the DNA structure in the crystal package.
**(a)** Shows the 3' overhang (cyan) threads through the arch hole of the neighboring molecule (wheat) which faces the 5' (green) of the neighboring symmetry related duplex (blue dashed circle). **(b)** Shows the arrangement when crystal symmetry is applied for the 3ov6 DNA with six T5FEN-D155K molecules (protein molecule not shown).

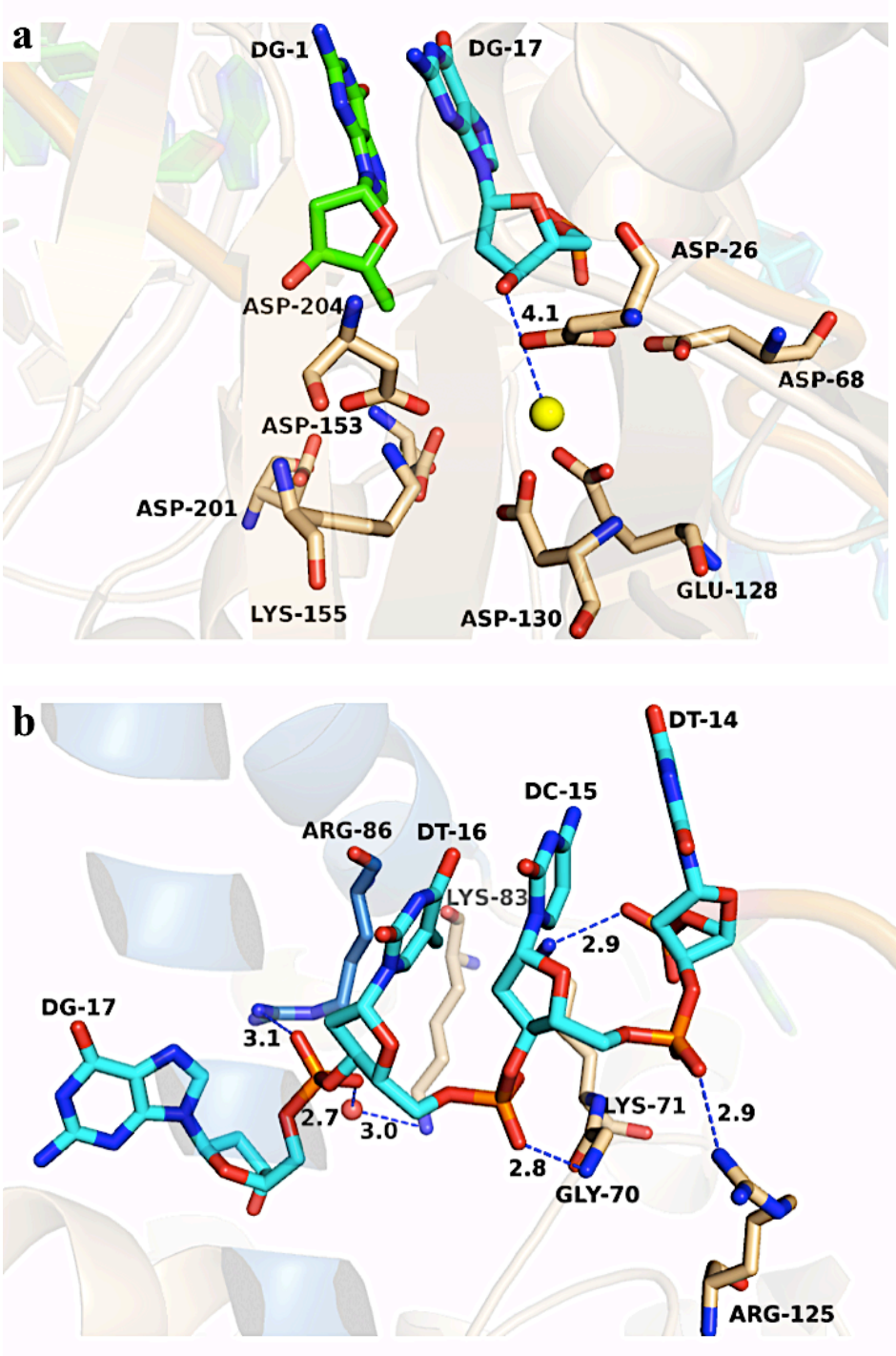



**Figure 6.92:** The DNA interactions underneath the neighboring molecule arch. **(a)** The symmetric dG-1 of 5' (green) and dG-17 of 3' overhang (cyan) pass above the active site and stacked against each other. The calcium ion (yellow sphere) present in M1 of Cat1 site of C3 structure. **(b)** Shows the 3' overhang (cyan) interactions with the basic residues (wheat) under the neighboring molecule arch (blue). Lys-83, Arg-86 and Arg-125 side chains interact with the DNA backbone. The new DNA binding site, Gly-70 and Lys-71, presents and interact with the 3' overhang phosphate groups.

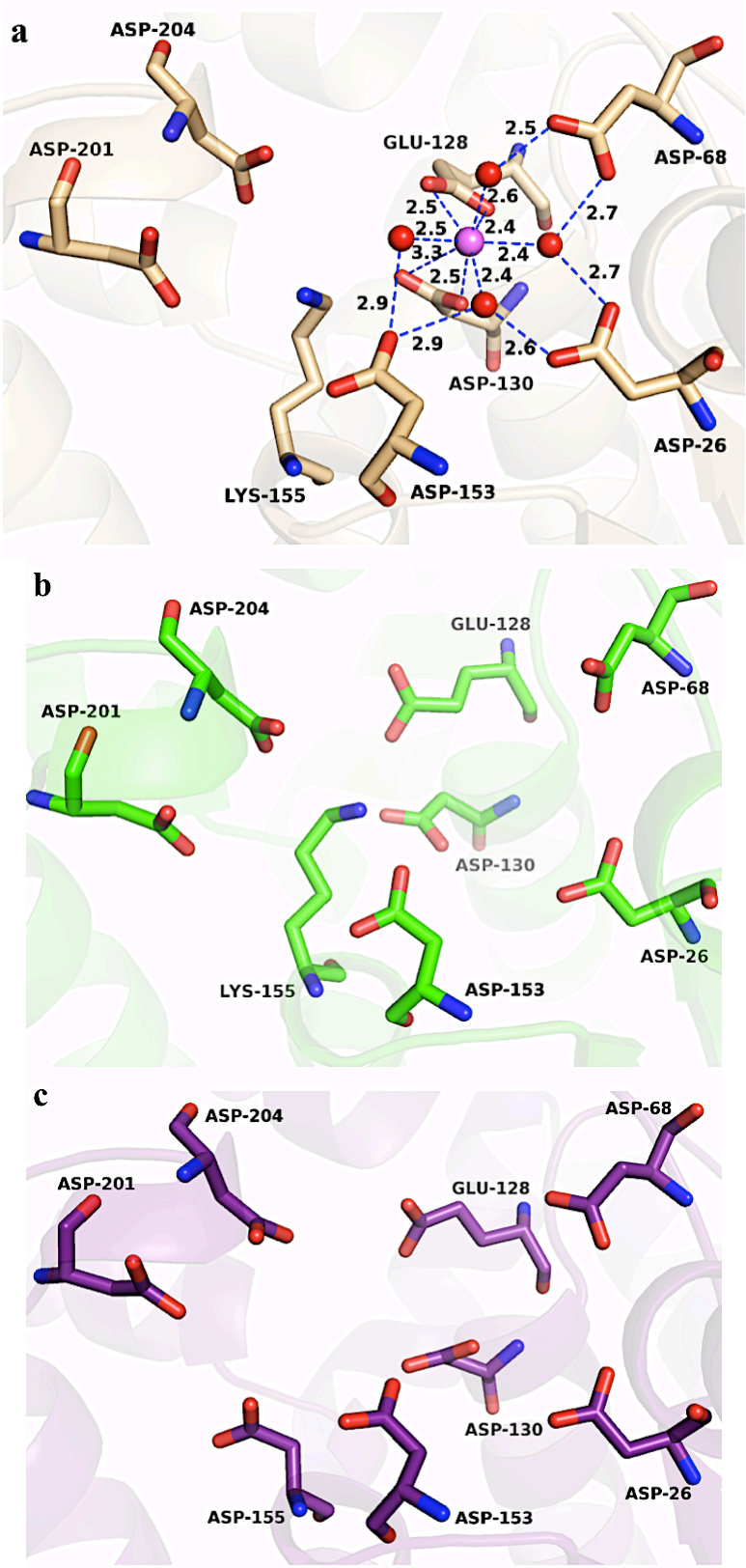



**Figure 6.93:** The active site residues for C3, C4 and CW structures are shown in **(a), (b)** and **(c)** respectively.
One metal ion is present in C3 active site which is $Ca^{2+}$ ion (magenta sphere) at M1 site while the other complexes, C4 and CW, active sites are metal ion-free.

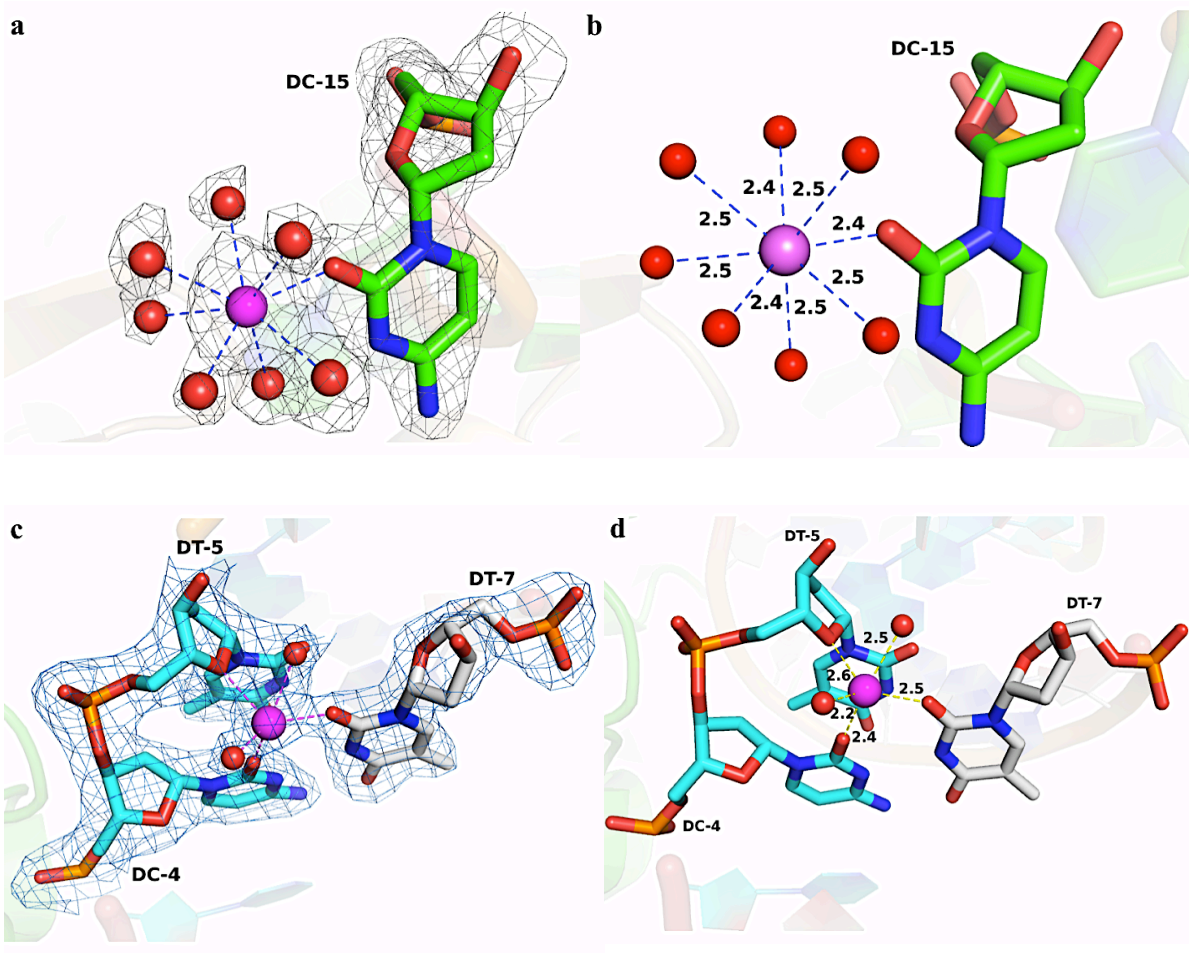



**Figure 6.94:** The metal ions found outside the active site in C3 and C4 structures. **(a)** Shows the 2Fo-Fc electron density map for the $Ca^{2+}$ ion (magenta sphere) presents in C3 and coordinated with seven water molecules and dC-15 while **(b)** presents the ion bond lengths. **(c)** Shows the 2Fo-Fc electron density map for the $Na^{+}$ ion (magenta sphere) presents in C4 and coordinated with two waters, dT-4, dC-5 and dT-7 nucleotides. **(d)** The same $Na^{+}$ ion with the bond lengths.

This metal ion is coordinated by two water molecules, dC-4 base oxygen-2, dT-5 deoxyribose oxygen-4 and interacts also symmetrically with dT-7 base oxygen-2 (Figure 6.10c & d). It appears to be a $Na^+$ ion considering to the characteristics, bond lengths (Harding, 2002; Kuppuraj et al, 2009) and components which were present in the crystallization buffer.

## 6.2 Discussion

In these crystal structures of T5FEN:3ov6 DNA, one copy of the protein interacts with one copy of the 3ov6 DNA. The DNA substrate interacts with the conserved H3TH:$K^+$ motif in the same way for the previously structures of T5FEN-D153K:5ov4 DNA (see sections 4.1.1.3.1 and 4.1.2.2.1) and as it is in other FEN members complexed with DNA substrates (Anstey-Gilbert et al, 2013; Devos et al, 2007; Orans et al, 2011; Tsutakawa et al, 2011). The active sites here in these structures are metal ions-free except for C3 which has calcium ion in M1 of Cat1 site. Of interest, Lys-155 side-chain in C3 active site is also rotated in the same way of C2/A Lys-153 (see section 4.2.4) and occupied a place could be used as metal ion binding site. Additionally, a potassium binding site was observed in two structures again while they are different variant of T5FEN complexed to a different DNA substrate but the $K^+$ ion is found in the same position that has been determined for T5FEN-D153K in complexes with 5ov4 DNA (see sections 4.1.1.3.1 and 4.1.2.2.1) and coordinated with the same residues.

Structurally, the 3' overhang in these complexes shows a similar path to the T4RNase H 3' arm (Devos et al, 2007) and the hFEN:DNA 3' flap (Tsutakawa et al, 2011). However, the last two bases, dT-16 and dA-17, in the 3' end of the symmetrically palindromic duplex penetrate under the helical arch of a neighboring T5FEN molecule (Figure 6.7a). The process that resulted in this the 3' end of this overhang resting under the arch, with the conserved basic side chains of Lys-83 and Arg-86 forming charge interactions with the 5' phosphate of dG-17 is similar to that observed in the 5' overhang of 5ov4 DNA substrate described previously (see section 4.2) but physically it must have approached the neighboring molecule in the

opposite direction compared with the 5ov4 substrate (Figure 6.7a) but results in similar interactions (Figure 6.11). The characters of looping-up and ratchet/barb-like structure that was made by rotation of the second nucleotide in the 5' overhang of 5ov4 DNA in a complex with T5FEN-D153K variant (see sections 4.1.1.3.3 & 4.1.2.2.4) were not observed in this serendipitously obtained structures with 3ov6. The presence of the 3' DNA end under the arch and its passes above the active site can shed a light about the back tracking model and the importance of the protein-protein interactions in these enzymes activity (Williams & Kunkel, 2011). Another interesting point is the presence of the new DNA binding site in the distal side of the arch which interacts again with the dG-17 phosphate oxygen but in different situation from that previously observed with the 5ov4 structure (see section 4.1.2.2.4). The observation of this interaction may suggest that residues on the distal side of the T5FEN arch play a role in substrate binding and may be similar in other FEN members. This could be investigated by mutating these two residues as single and double mutations to Ala or Asn residues and determine the Kd through some assay techniques such as EMSA or anisotropy. It is clear that more investigations are needed with different DNA substrates having a 3' flap and/or 5' flap to study this issue and provide further insight into this physical process

### 6.2.1 Superposition of C4-T5FEN-D155K:3ov6 and CW on C3

These structures, C4 and CW, are superimposed on C3 using secondary structure matching in the LSQKAB program (CCP4, 1994) in order to distinguish the differences that could be found between them. The superposition shows a very well overlay between them (Figure 6.12) with average of RMSD 0.303 Å and 0.44 Å respectively in C$\alpha$ positions over a length of 272 amino acids. These three structures are very similar to each other and no any significant differences can be noticed between them. Additionally, the 3ov6 DNA follows the same trajectory (Figure 6.12c) and interacts with the same residues.

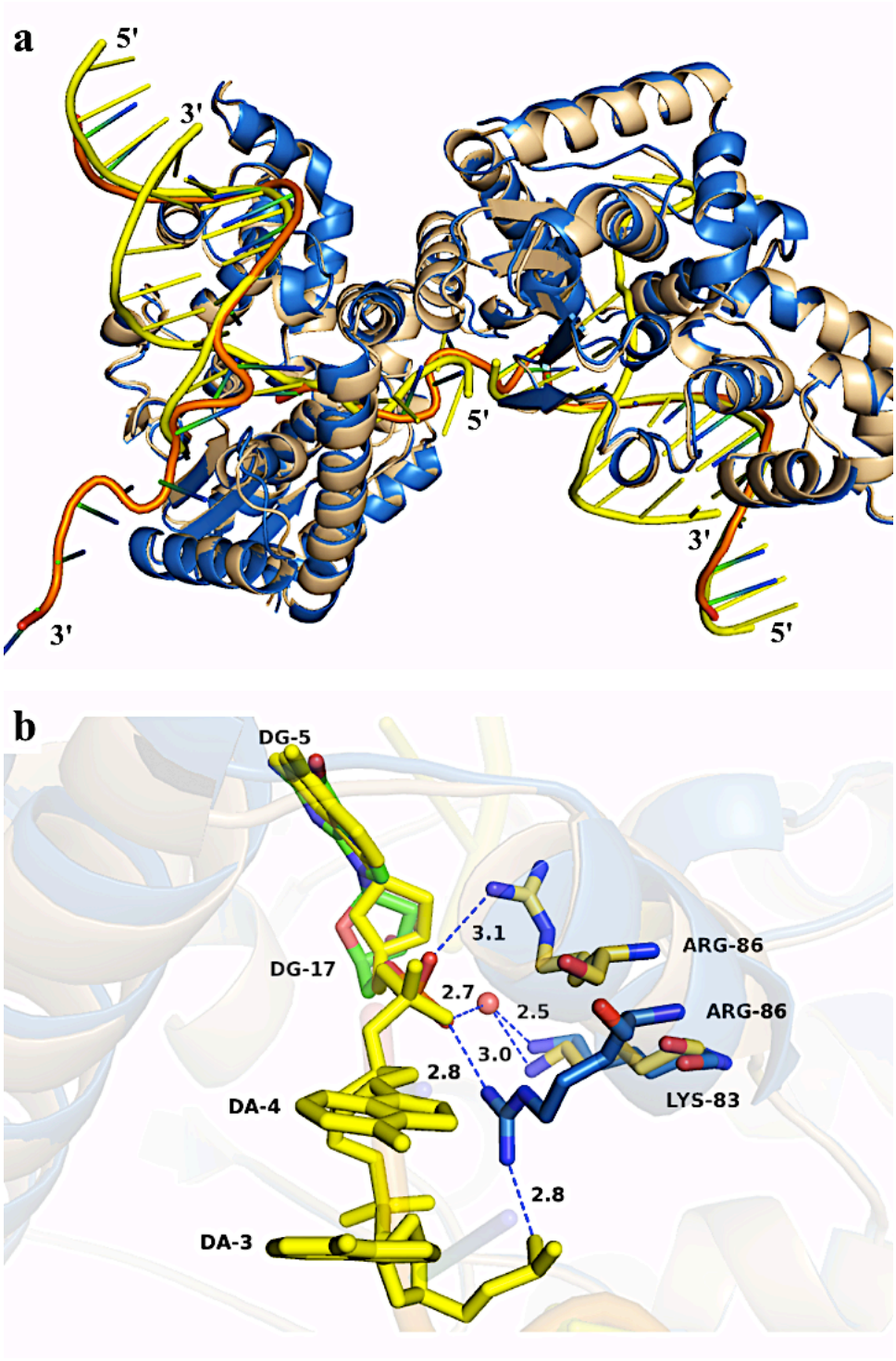



**Figure 6.95:** Superposition of C2/B on C3.
**(a)** Shows the 3' overhang of 3ov6 DNA ( orange backbone and green bases) resting under the arch of the neighboring superposed molecules. **(b)** Illustrates the interaction between the 5' overhang of 5ov4 (yellow) and the 3' overhang of 3ov6 with the arch residues, Arg-86 and Lys-83.

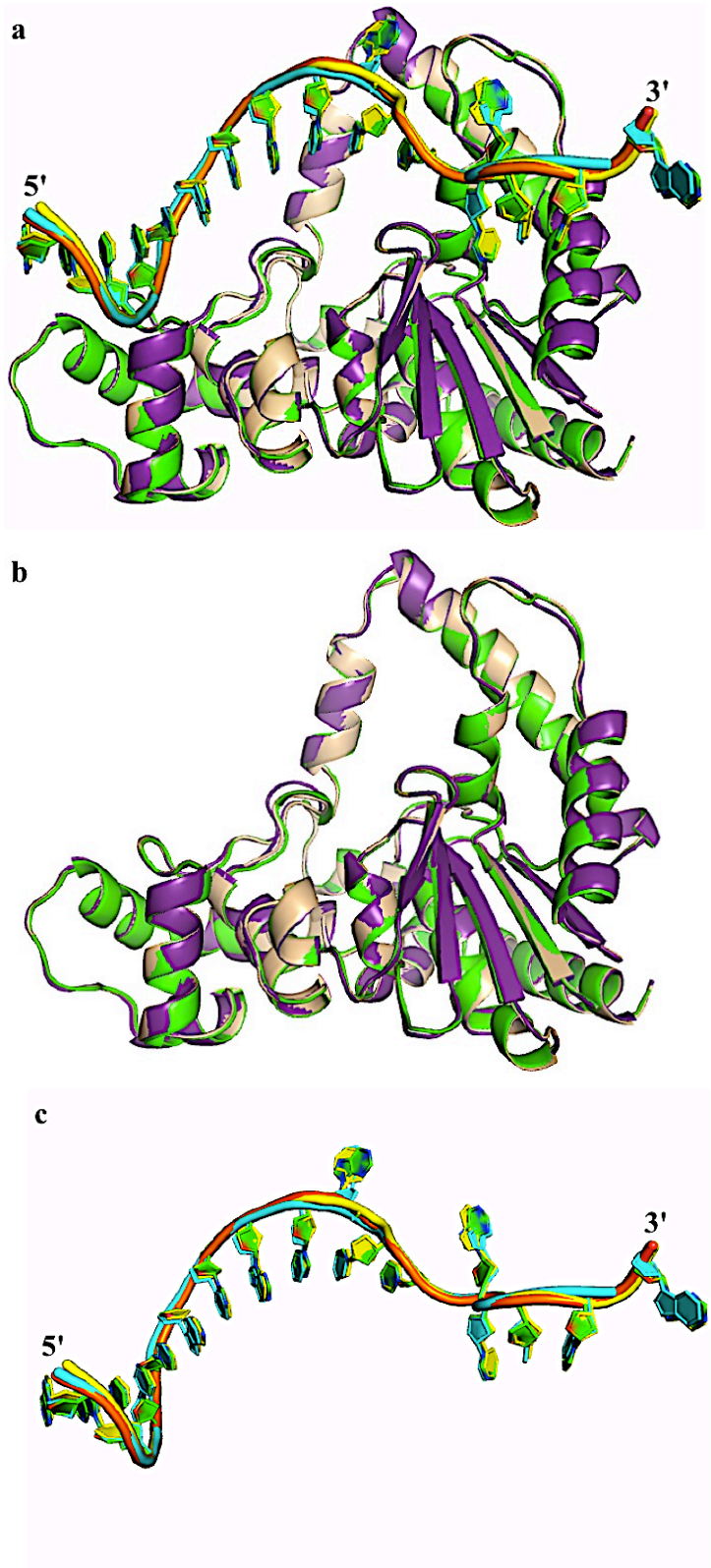



**Figure 6.96:** Superposition of C4 (green) and CW (purple) on C3 (wheat) structures.
**(a)** Shows the three complexes with their DNA substrates while **(b)** illustrates the protein chains without DNA. **(c)** Shows the superposed of the three DNA substrates together. Green DNA for C3, cyan for C4 and yellow for the wild type complex.

#### *6.2.1.1 Superposition of the Active Site Residues*

Superposition of the active sites of C3, C4 and CW illustrates the high level of similarity between these amino acids in C$\alpha$s (Figure 6.13a). The side chain hydroxyl groups of Glu-128 and Asp-130 are rotated in the C3 compared to C4 and CW in order to coordinate the $Ca^{2+}$ ion that occupies M1 in its active site (Figure 6.13a). The side chain of Lys-155 does not occupy the same position in each structure, their side-chain NZ are about 1.7 Å from each other and in C4 it is ~ 3.7 Å apart from the $Ca^{2+}$ ion presents in the C3 active site (Figure 6.13a).

#### *6.2.1.2 The Potassium Binding Site*

A potassium binding site presents in C3 and C4 structures was not observed in CW. Superposition of C4 on C3 shows the two $K^+$ ions in the same position and they are about 0.5 Å away from each other (Figure 6.13b). The absence of the $K^+$ ion from T5FEN-WT:DNA structure could be due to the poor resolution as mentioned before.

### 6.2.2 Superposition of C3 on T5FEN:$Mg^{2+}$ and C2-T5FEN-D153K:5ov4 chains A and B.

This section will focus on comparisons of T5FEN-D155K:DNA with other FEN structures which will be reported for C3 structure only because the high degree of internal similarity between the three complexes of T5FEN with the 3ov6 DNA. The T5FEN native and the two chains C2/A and C2/B, overlay well on C3 (Figure 6.14) with average of RMSD 1.14, 0.75, and 0.94 respectively in C$\alpha$ positions. The similarity between C3 and C2/B is higher than with the other two chains (native and C2/A). According to this analysis the largest location movement can be seen is between T5FEN native and C3 structure at the top of the arch area (a.a. 90-103) (Figure 6.14a). Additionally, His-36 main and side chains are rotated by approximately 180º and change the residue position (Figure 6.15a) in C2/B and C3 comparing to (1EXN) structure. However, in C2/A and in T5FEN native this residue is disordered and cannot be seen.

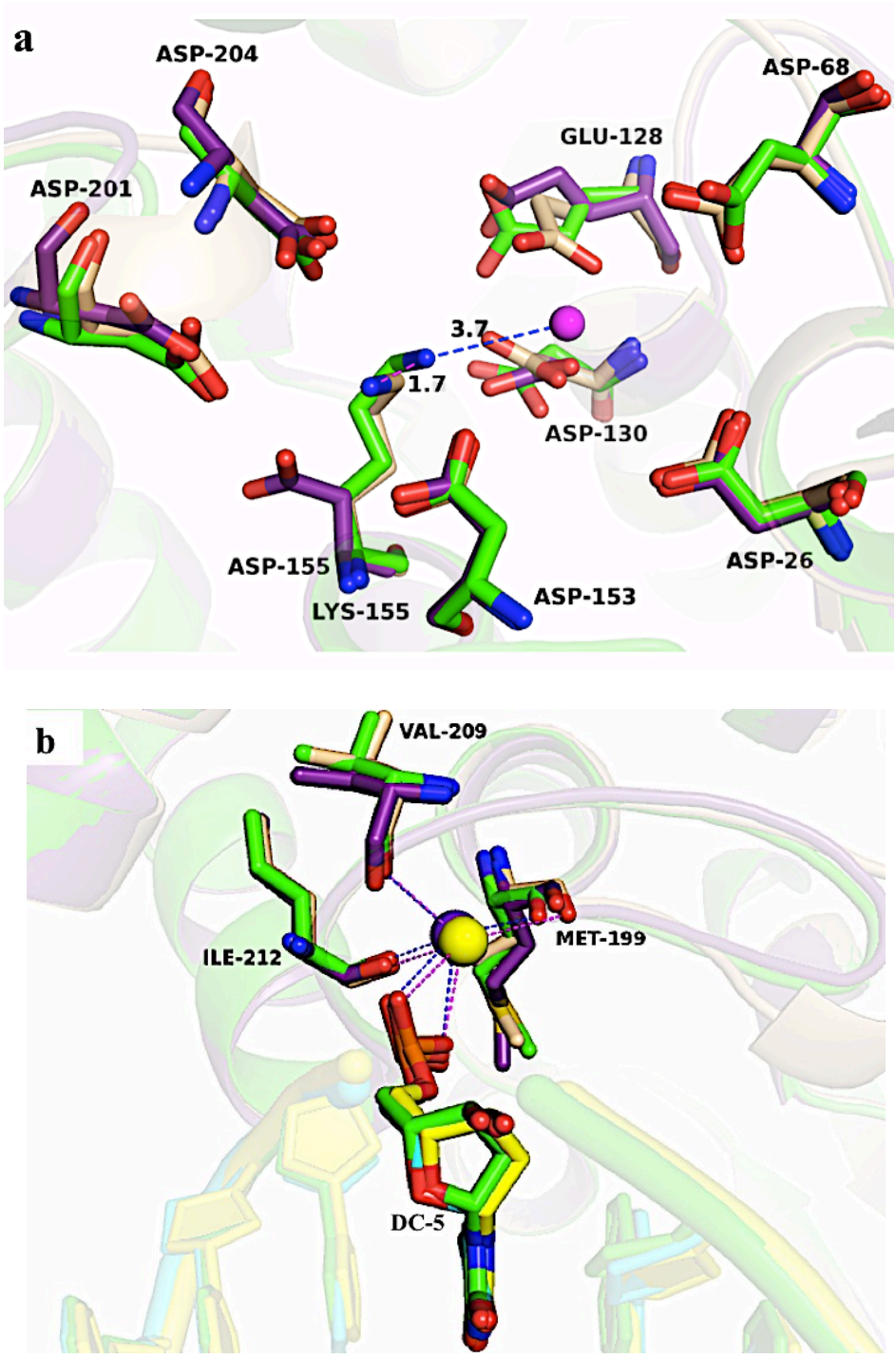



**Figure 6.97:** Superposition of the active site residues and the $K^+$ ions in C3, C4 and CW structures.
**(a)** Shows the active site residues of C4 (green sticks) and WC (purple sticks) superposed on C3 active site (wheat sticks). A $Ca^{2+}$ ion (magenta sphere) presents in M1 of C3-Cat1 site. **(b)** Shows superposition of the $K^+$ ion-binding site in C3, C4 and CW. Potassium ions (yellow and purple spheres) are present in C3 and C4 structures respectively.

Also, the side chain of Phe-169 is rotated by 90º in the three structures containing T5FEN: the native, C2/B and C3 compared to C2/A (Figure 6.15a).

### *6.2.2.1 Superposition of the DNA Substrates*

Superposition of the two DNA substrates, 5ov4 and 3ov6, shows general alignment for strand X from 5ov4 with 3ov6 with looping-up of guanosine-10 from the later one (Figure 6.15b). By deleting or ignoring chain X of the 5ov4 DNA, these two DNA molecules appear together to be like a pseudo-Y or fork DNA similar to the one that was determined in a complex with T4RNase H (Devos et al, 2007) which has both 5' overhang and 3' arms (Figure 6.15b). The interactions between the 5' overhang and T5FEN-D153K:DNA are described in Chapter 4 while the 3' arm interactions explained in the beginning of this chapter in section 6.1.2.4.

### *6.2.2.2 Comparison of the Active Site Residues*

When the carboxylic residues of the active site of T5FEN native enzyme and C2/A, C2/B superposed on C3 active site amino acids they show similarity between them and overlay well one on each other of C$\alpha$ positions (Figure 6.16). The maximum shift is obtained for C2/A active site comparing to other chains (Figure 6.16b) while the lowest level of dissimilarity is notable for the C2/B active site residues (Figure 6.16c). In general, Asp-201 and Asp-204 in all three chains (the native, C2/A and C2/B) present the highest degree of localized movement compared to C3 active site residues.

Six metal ions are present in the superimposed active sites of these proteins (Figure 6.16). Three of these metal ions are $Mg^{2+}$ ions for T5FEN native, two $Mg^{2+}$ for C2/B while the sixth metal ion is $Ca^{2+}$ ion for C3. The magnesium ions occupy M1, M2 and M3 of Cat1 and Cat2 sites as mentioned previously whilst a $Ca^{2+}$ ion lies in M1 of the Cat1 site (Figure 6.16) in the same position as the first $Mg^{2+}$ ions in the native and C2/B structures and they are about 0.7 Å from each other. The mutated Lys-155 of C3 active site was occupied a new place close to M2 in Cat1 site and its NZ

is about 1.9 Å and 5.2 Å from the second and the third $Mg^{2+}$ ions that occupy M2 and M3 respectively in T5FEN native (Figure 6.16a). In the superposed active site of C2/A and C3 the two mutated lysine, Lys-153 and Lys-155, are visible and their NZ are 6.3 Å apart from each other (Figure 6.16b) and both of them expected to be monitors for new metal binding sites in Cat1. The calcium ion which lies in M1 is 4.7 Å and 5.2 Å from Lys-153 and Lys-155 respectively (Figure 6.16b). In addition, in superimposed active site of C2/B and C3the the distance between Lys-153 and Lys-155 side chains, NZ, is 2.5 Å and Lys-153 is 3.0 Å away from $Ca^{2+}$ in M1 whilst Lys-155 is 5.7 Å and 6.2 Å from the C2/B $Mg^{2+}$ ions in M1 and M3 sites respectively (Figure 6.16c).

#### *6.2.2.3 Superposition of the Potassium Binding Sites*

The potassium-binding site is found in C2/A and C2/B and also it was present in C3 structure while absent from T5FEN native. In all of these three chains it overlaid very well and was coordinated by Met-199, Val-209 and Ile-212 residues (Figure 6.17). The $K^+$ ion in C3 is less than an Angstrom from the one in C2/A (Figure 6.17a) and about 1.5 Å from $K^+$ ion in C2/B first position (Figure 6.17b). Deoxycitidine-10, dC-6 and dT-5 phosphate oxygens interact with the $K^+$ ions in C2/A, C2/B and C3 respectively (Figure 6.17).

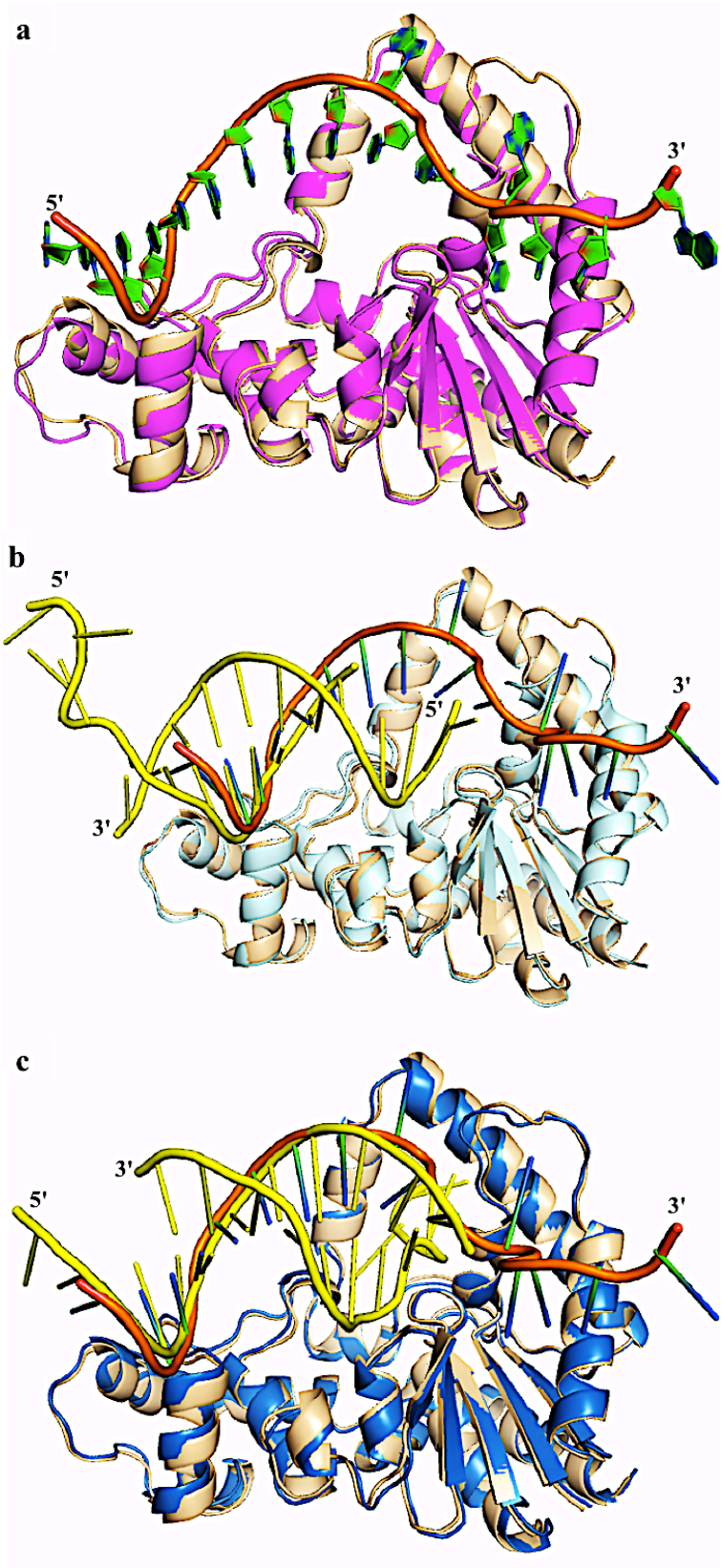



**Figure 6.98:** Superposition of C3 on other T5FEN complexes.
Superposition of T5FEN native (magenta), C2/A of T5FEN-D153K:5ov4 (paly cyan protein and yellow DNA) and C2/B (blue protein and yellow DNA) on C3 (wheat protein and orange DNA) as shown in **(a), (b)** and **(c)** respectively.

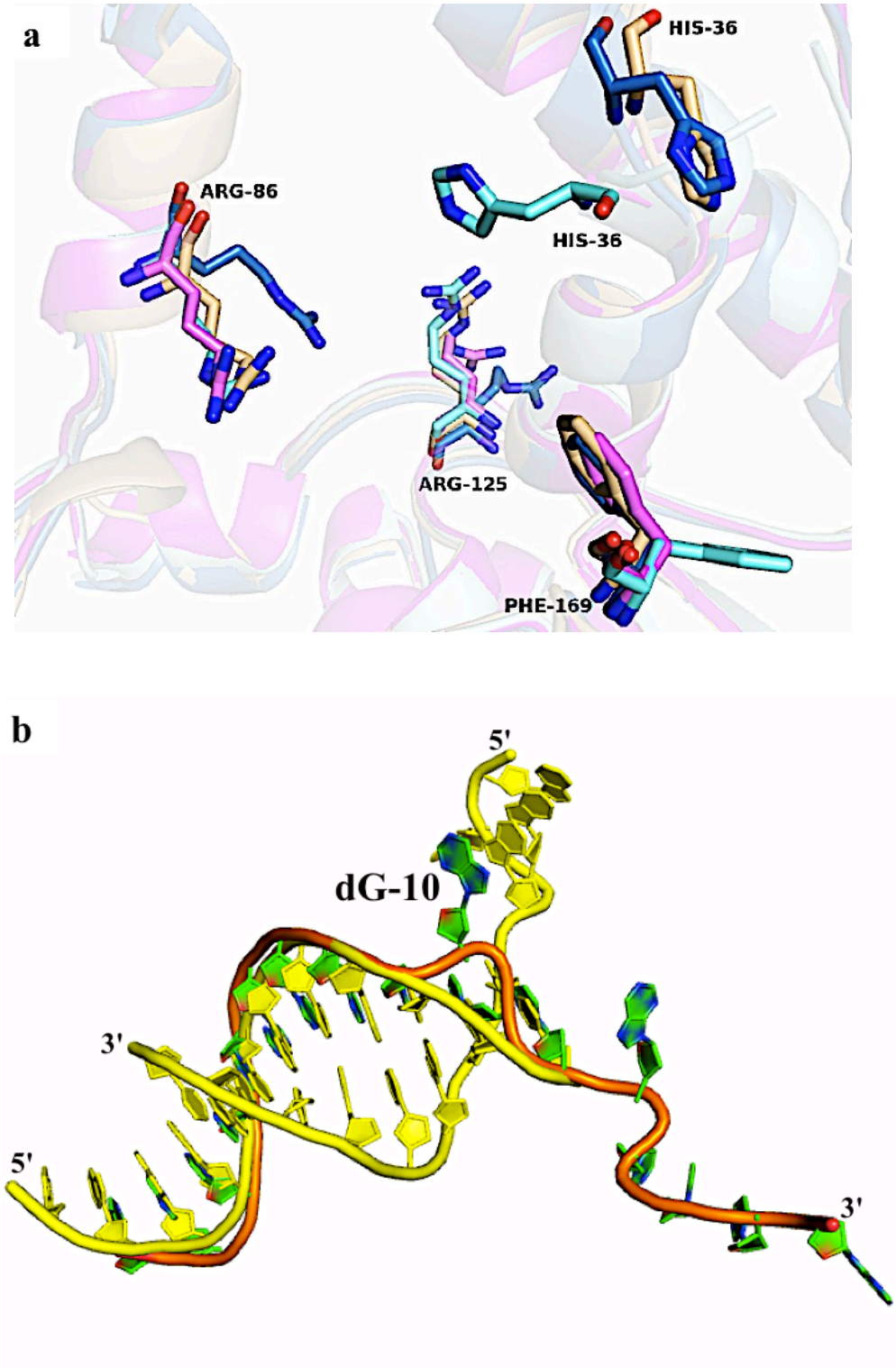



**Figure 6.99:** Conformational changes in some superposed residues.
**(a)** Shows the conformational changes in His-36, Arg-86, Arg-125 and Phe-169 in C2 (paly cyan for chain A and blue for chain B) and C3 (wheat) structures comparing to T5FEN native (magenta). **(b)** Illustrates the superposition of 5ov4 DNA (yellow) on 3ov6 (orange backbone and green bases). The two DNA substrates composed a fork or pseudo-Y DNA.

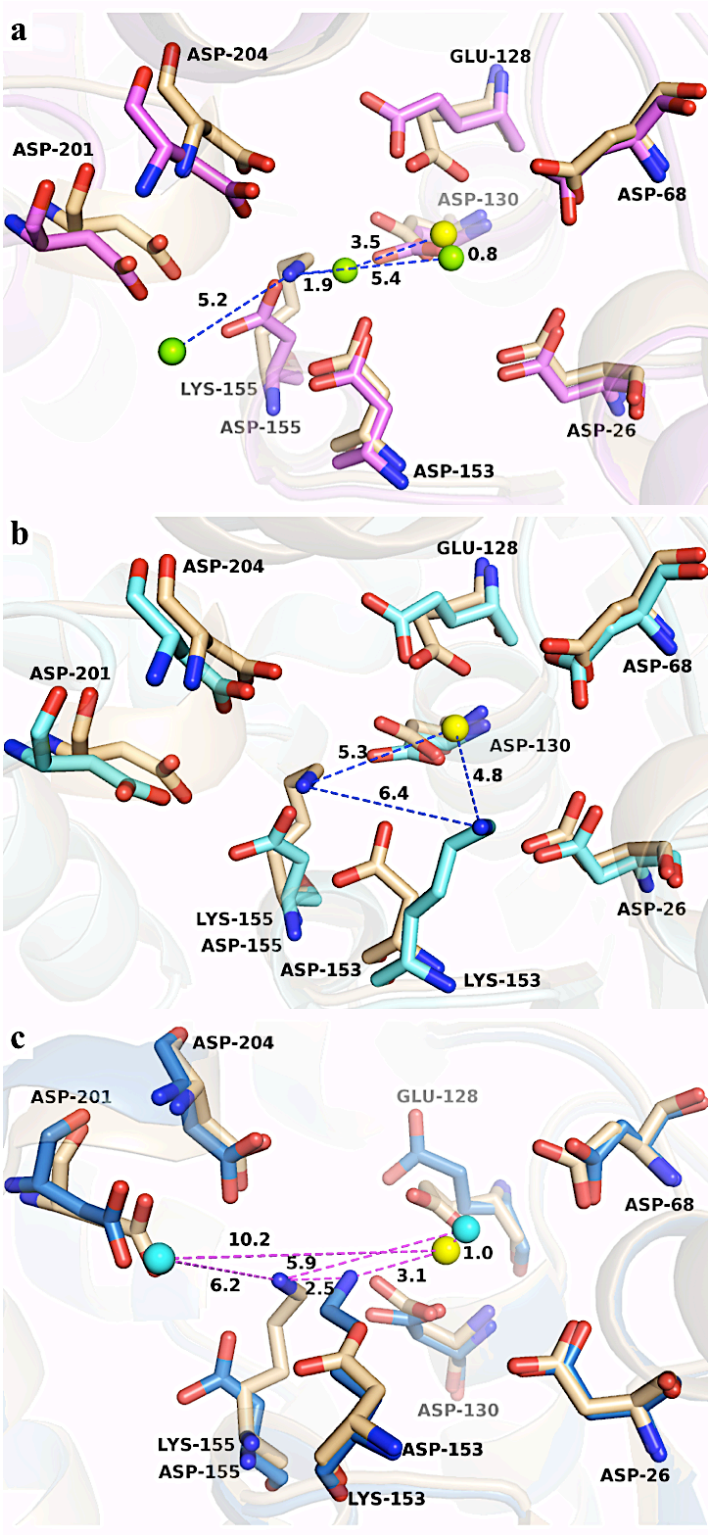



**Figure 6.100:** Superposition of C3 active site residues on other T5FEN complexes. Superposition of the active site residues of T5FEN native (magenta sticks), C2/A (paly cyan sticks), C2/B (blue sticks) on C3 active site (wheat sticks) as shown in **(a), (b)** and **(c)** respectively. The metal ions present in the superimposed active sites are $Ca^{2+}$ (yellow sphere) for C3, $3Mg^{2+}$ (green spheres) for the native structure and $2Mg^{2+}$ (cyan spheres) for C2/B.

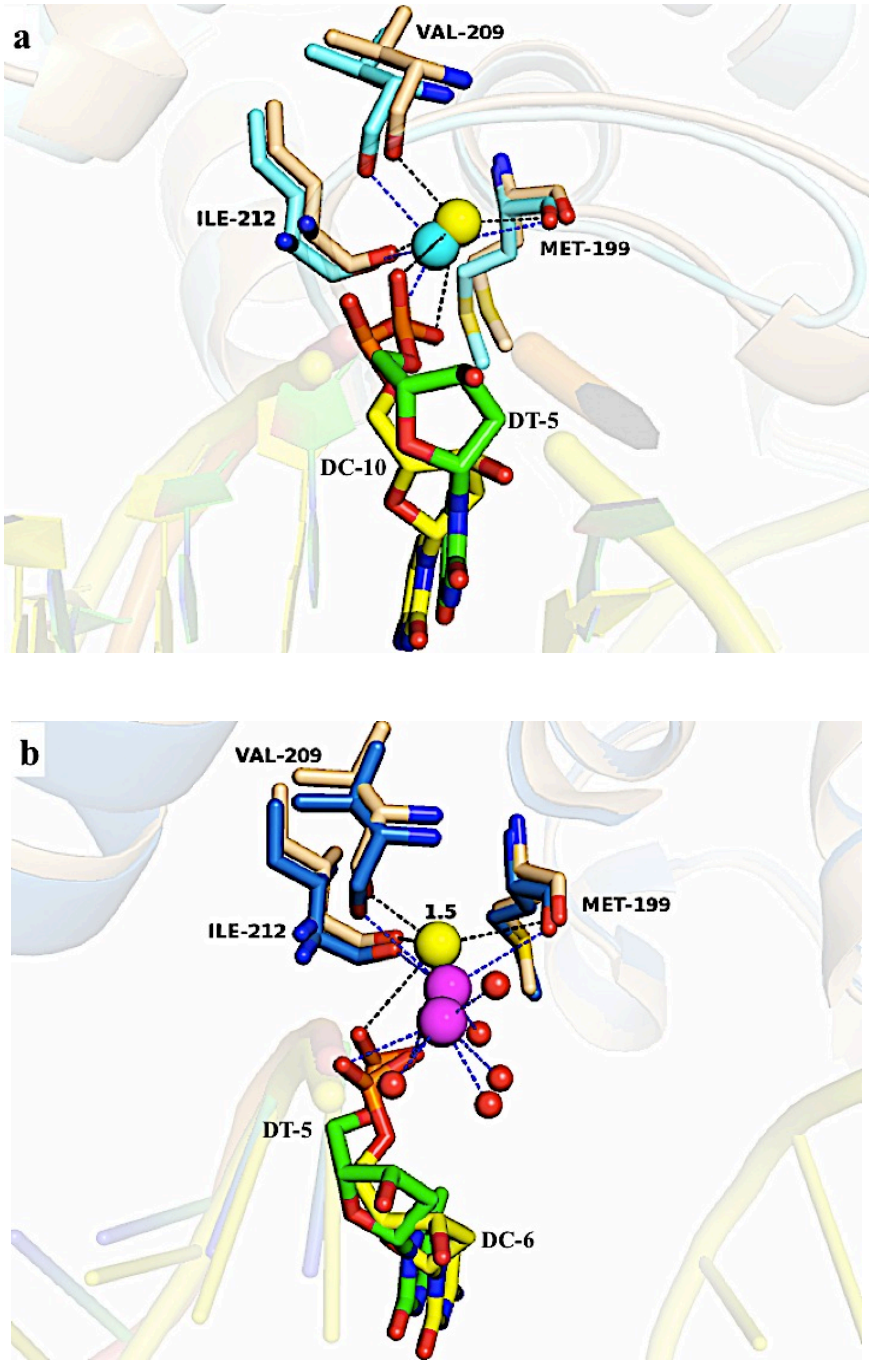



**Figure 6.101:** Superposition of the potassium-binding site.
**(a)** and **(b)** show superimposed of C2/A $K^+$ ion (cyan sphere) and C2/B $K^+$ ion (magenta spheres) on C3 $K^+$ ion (yellow sphere) respectively. In **(b)** the $K^+$ ion presents in two positions (two magenta spheres).

### 6.2.3 Superposition of C3 on T4RNase H:DNA, hFEN:DNA and hEXO1:DNA

Three FEN family members complexed with different DNA substrates can be compared with the C3 structure. In all of these complexes the DNA substrates have a 3' arm, flap or overhang. The three proteins overlay well (Figure 6.18) on C3 in C$\alpha$ positions with average RMSD 2.3 Å, 3.0 Å and 2.7 Å for T4RNase H:DNA, hFEN:DNA and hEXO1:DNA respectively. A high similarity level is observed between C3 and T4RNase H while hFEN has the least similarity. The differences between T5FEN enzyme, C3, and these FEN members are discussed previously in chapters 4 & 5.

DNA substrates do not overlay very well with the 3ov6 DNA as seen in Figure 6.19 due to the looping of some nucleotides that are found in 3ov6 substrate but they follow a similar trajectory around their protein partners in all of these FEN-family enzymes.

#### *6.2.3.2 Comparison of the Active Site Residues*

The conserved active site residues of T4RNase H, hFEN and hEXO1 are superposed on C3 active site amino acids. The C$\alpha$s of these residues overlay fairly well on each other with some shifts in the main chains and/or side chains (Figure 6.20). In the superposed active site with T4RNase H a metal ion is present in M1 of Cat1 site (Figure 6.20a) which is the $Ca^{2+}$ ion for C3 because T4RNase H:DNA was determined without metal ions (Devos et al, 2007) as mentioned previously.

The superimposed active site with hFEN:DNA and hEXO1:DNA show less overlay than T4RNase H:DNA active site and more residues location movements (Figure 6.20b & c). Five metal ions are present in each superimposed active site. Four of them are $Sm^{3+}$ ions belongs to hFEN:DNA substrate and product complexes (Tsutakawa, 2011, Figure 6.19b) while the other four are $2Ba^{2+}$ and $2Mn^{2+}$ for hEXO1:DNA substrate and product respectively (Orans, 2011 #362, Figure 6.20c). The fifth ion in each superposed active site is $Ca^{2+}$ ion for C3. The first $Sm^{3+}$ and $Ca^{2+}$ ions

occupy M1 at Cat1 sit and the $Ca^{2+}$ ion is about 0.9 Å apart from the $Sm^{3+}$ in product and substrate structures while the second $Sm^{3+}$ ion lies in M2 and it is 3.4 Å and 3.8 Å from the $Ca^{2+}$ ion in M1 for product and substrate respectively. The mutated Lys-155 side chain terminal is 1.9 Å and 1.5 Å from the $Sm^{3+}$ ion in the product and substrate M2 active sites respectively (Figure 6.20b). The first $Ba^{2+}$ and $Ca^{2+}$ occupy M1 of Cat1 site and separated by about 2.7 Å whilst the second $Ba^{2+}$ ion locates in M2 sit (Figure 6.20c) and it is 4.2 Å and 1.6 Å from the $Ca^{2+}$ ion and the NZ of the mutated Lys-155 side chain respectively. In hEXO1:DNA product active site, the $Ca^{2+}$ ion is 1.3 Å and 4.5 Å from the $Mn^{2+}$ ions in M1 and M2 respectively (Figure 6.20c) while the Lys-155 side chain terminal is ~ 0.8 Å from the $Mn^{2+}$ ion in M2 site which is support the theory about the new metal ion binding sites can be presents in T5FEN active site.

#### *6.2.3.1 Superposition of the Potassium Binding Site*

The potassium binding site that was recognized firstly in hFEN:DNA and hEXO1:DNA (Orans et al, 2011; Tsutakawa et al, 2011) and recently in T5FEN:DNA was not present in the T4RNase H:DNA structure (Devos et al, 2007). The three residues Met-199, Val-209 and Ile-212 that coordinate the $K^+$ ion in C3 are equivalent to Lys-195, Val-205 and Met-224 in T4RNase H protein which appear far from the superimposed $K^+$ in C3 structure (Figure 6.21a). This fact supports the absence of this site in T4RNase H or it could be present with different conformational changes within the H3TH motif in this enzyme. Superposed of the hFEN:DNA and hEXO1:DNA potassium ions on the C3 $K^+$ ion show a location movement of this ion by about 2.2 Å and 4.4 Å in hFEN:DNA and hEXO1:DNA respectively from its position in C3 (Figure 6.21b & c).

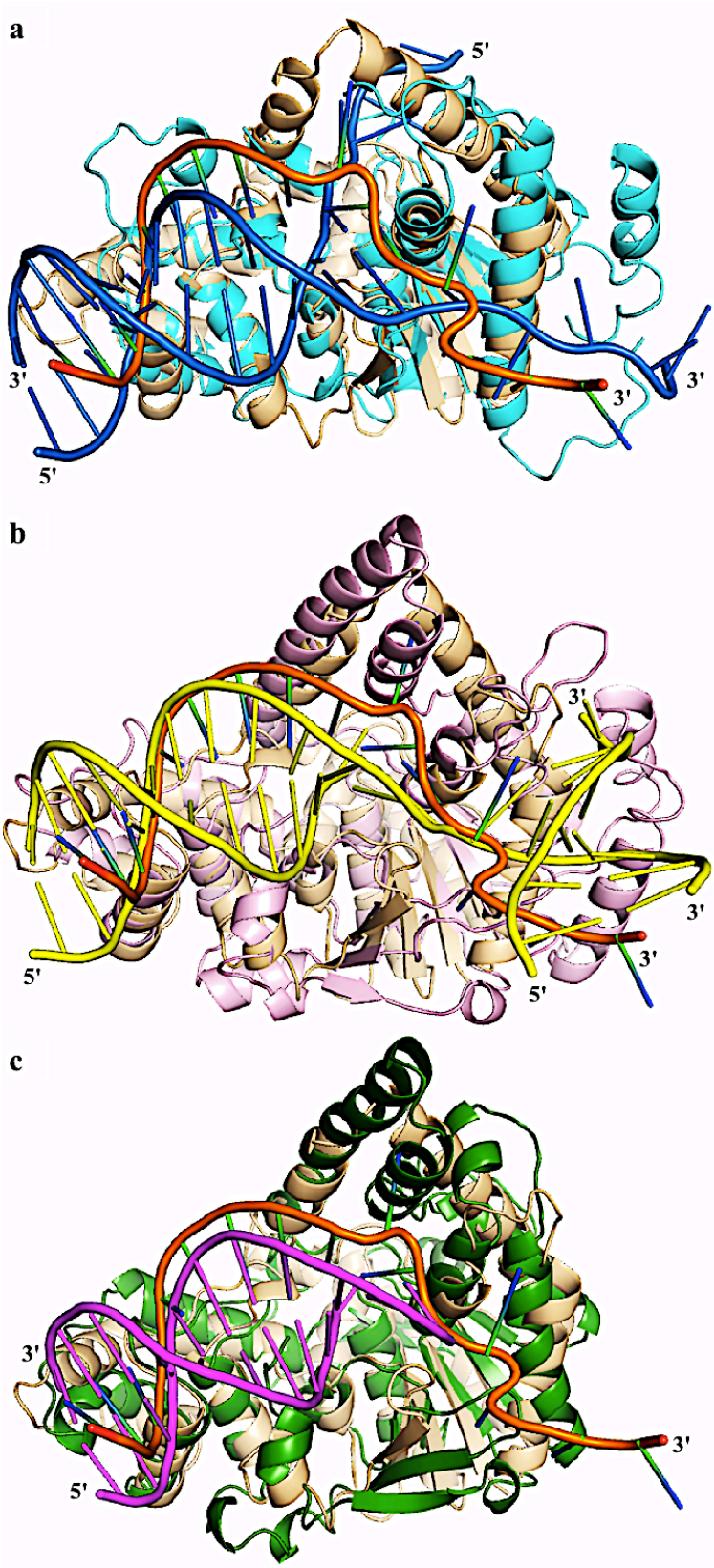



**Figure 6.102:** Superposition of C3 structure on other FEN family members
Superposition of T4RNase H:DNA (cyan protein and blue DNA), hFEN:DNA (pink protein and yellow DNA) and hEXO1:DNA (green protein and magenta DNA) on C3 (wheat protein and orange DNA) as shown in **(a), (b**) and **(c)** respectively**.**

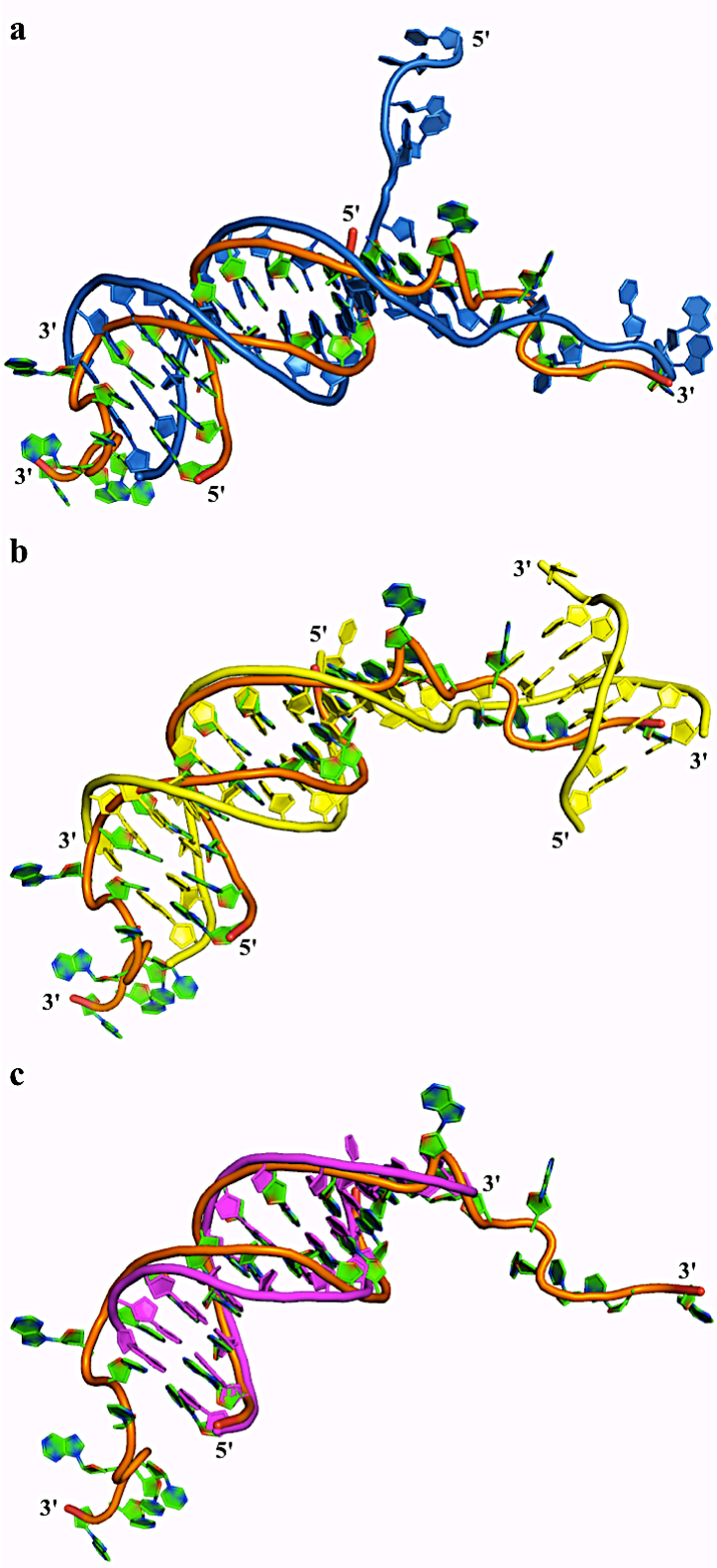



**Figure 6.103:** Superposition of 3ov6 DNA (orange backbone and green bases) as duplex on other FENs DNA substrates.
**(a)** Superposition on pseudo-Y (blue) in T4RNase H:DNA. **(b)** Superposition on duple flap-product (yellow) in hFEN:DNA. **(c)** Superposition on 3' flap (magenta) in hEXO1:DNA.

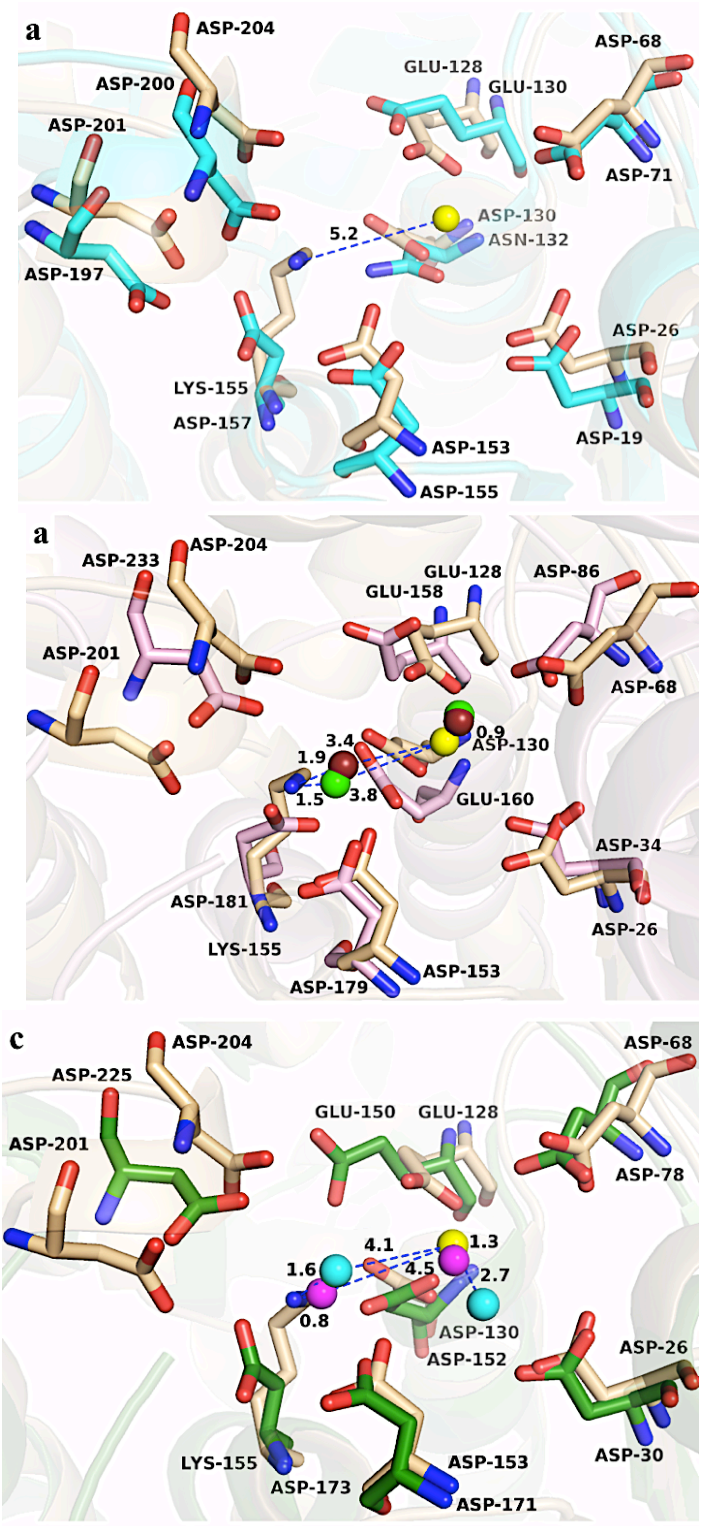



**Figure 6.104:** Superposition of C3 active site residues on some FEN members. **(a), (b)** and **(c)** superposition of the active site in T4RNase H:DNA, hFEN:DNA and hEXO1:DNA on C3 active site respectively. The metal ions present in the superimposed active sites are $Ca^{2+}$ ion (yellow sphere) for C3 structure, $Sm^{3+}$ (green and ruby spheres) for hFEN:DNA substrate and product respectively, $2Ba^{2+}$ (magenta spheres) and $2Mn^{2+}$ (cyan spheres) for hEXO1:DNA substrate and product respectively.

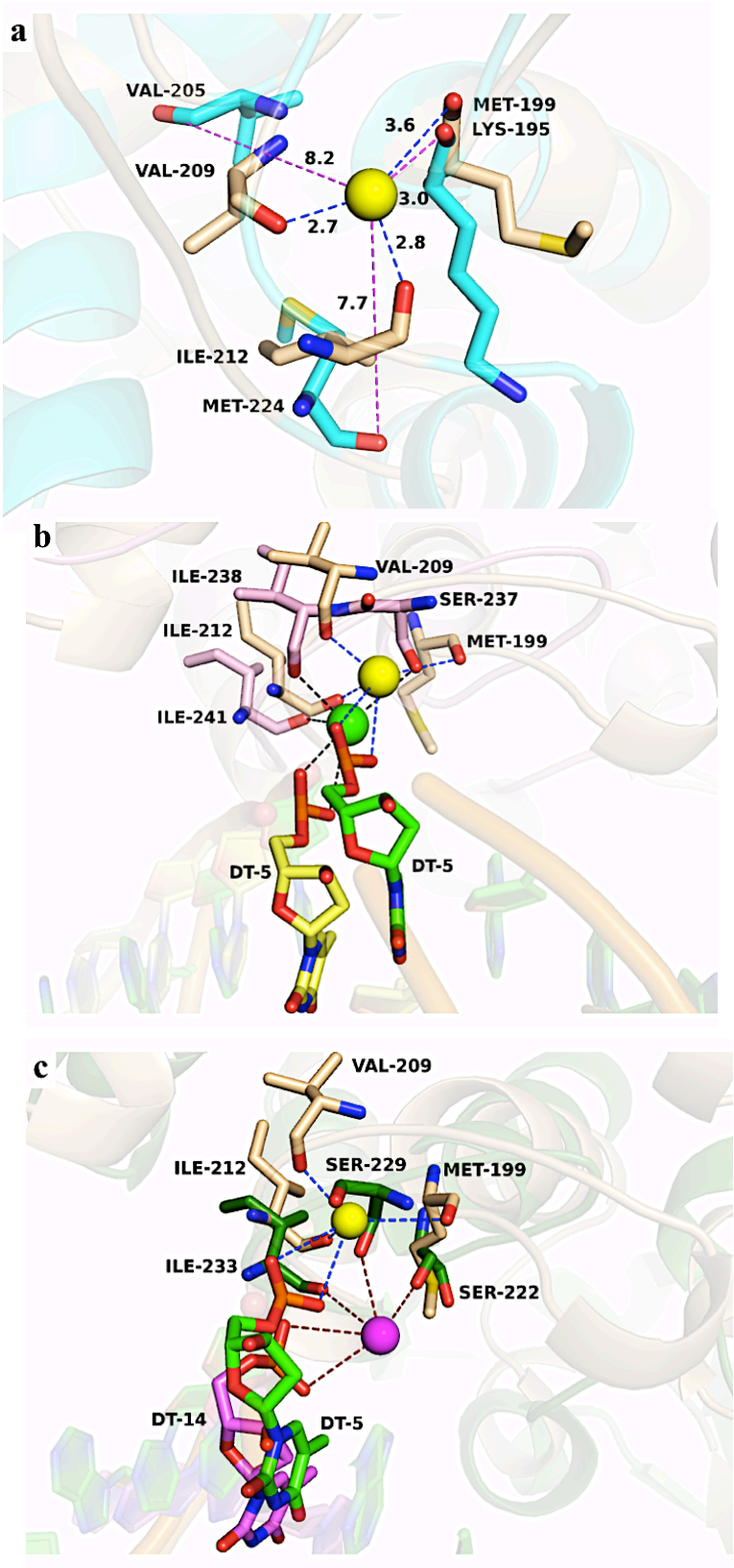



**Figure 6.105:** Superposition of the C3 potassium-binding site on: T4RNase H:DNA, hFEN:DNA and hEXO1:DNA as illustrated in **(a), (b)** and **(c)** respectively. The $K^+$ ion is absent from T4 while presents in hFEN (green sphere), hEXO1 (magenta sphere) and C3 (yellow sphere) structures.

# Chapter 7: Conclusion and Future Work

## 7.1 Conclusion

T5 flap endonuclease enzyme shares a common architecture with FEN family and has conserved catalytic and DNA binding motifs. These proteins can form complexes with a variety of DNA substrates and/or metal ions. A two-metal-ion mechanism was proposed for the FEN enzyme endonucleolytic reaction which takes place one nucleotide into the double-strand. These two metal ions must interact with the scissile phosphate and be 4 Å apart from each other (Beese & Steitz, 1991). Prokaryotic FEN enzymes differ from eukaryotic in some points such as the presence of the N-terminus in the eukaryote active site while in prokaryote this place is a site for metal ions known as Cat2 site except for a subset of FEN-like proteins exemplified by *E. coli* ExoIX which lacks this site (Anstey-Gilbert et al, 2013). In addition, the eukaryote FENs have a specific hydrophobic pocket to accommodate the 3'-single nucleotide that is composed of the C-terminal and the hydrophobic wedge loop while this character absent in prokaryote FENs and in hEXO1.

T5FEN crystal structures in complexes with 5' and 3' overhang that are analyzed and discussed in this thesis provide a clear picture for this FEN enzyme and how it recognize the ss-5' flap. There have been several models suggested to explain how FEN enzymes interact and process the 5' flap or overhang DNA. Tracking, threading, clamping or disorder-to-order mechanism, which do they use? Or can they use more than one mechanism? The T5FEN:DNA structures presented here support the threading model at least for T5 and prokaryotic homologues for processing of 5' flap structures. They have been shown looping-up in the ss-5' overhang before it pushed through the arch hole using the arch positively charged residues to provide electrostatic interactions with the substrate. The threading process made it easy to identify a new DNA binding site which lies in the back far side of the arch and can be present in other FEN family members. The equivalent

amino acids in the T4RNase H (Lys-74 and Ser-75) did interact with the superimposed 5ov4 DNA backbone by their main chains as observed in C2/B but the side chain of Lys-74 can form interaction with the dA-3 phosphate oxygens of the fork DNA which gives an indication that this site could be formed in the T4RNase H enzyme as it is in homolog T5FEN. In hFEN and hEXO1 the equivalent residues are Lys-88 and Ser-89 for hFEN and Ser-80 and Thr-81 for hEXO1. By superposed these two FENs on C2/B it was found that these residues far from 5ov4 DNA backbone which can be refer to the lacking of the real 5' flap or absence of this site in FEN superfamily or may interact through alternative method or have some conformational changes after threading of the 5' flap through the arch in hFEN. In addition, some residues conserved in the arch such as Arg-86 and Tyr-90 hang down from their positions in the native T5FEN and interacted with the DNA backbone to help stabilizing the ss-5' overhang during threading.

Several mechanisms have been observed in order to prevent the ss-DNA from slipping back from the arch after passes through it. These mechanisms included interactions with Arg-86, the formation of the gate-like structure by Tyr-90 and Phe-105 interactions with the distal/trans-arch DNA binding site and formation of the ratchet/barb structure by twisting of the ss-DNA second deoxynucleotide by 180º in the arch far side.

T5FEN-D155K and wild type complexes with the 3ov6 DNA follow a similar path to that was determined for T4RNase H in a complex with short fork DNA and hFEN in complex with DNA product. In C3, C4 and CW structures the insertion dG clamped by the arch residues. Lys-83 and Arg-125 from the neighboring molecule interacted symmetrically with the DNA backbone through hydrogen bonds. Additionally, the 3' overhang is able to pass underneath the neighboring molecule arch which may give a some support to the back tracking model. A binding site for a potassium ion is present in T5FEN:DNA complexes within the H3TH motif flexible loop and engaged in DNA backbone coordination. This site is thought to facilitate processivity in FEN proteins and is present in other FEN members in

complexes with DNA substrates except for T4RNase H:DNA (Devos et al, 2007) and *Af*FEN:DNA (Chapados et al, 2004).

A crystal structure of the parasitic *Trypanosoma brucei* flap endonuclease has been successfully obtained. This structure presents with the FEN family fold but with a disordered arch in the DNA-free sate and it is very similar to hFEN structure. The crystals of *T. brucei* FEN obtained had four molecules in the asymmetric unit in complexes with $Ca^{2+}$ ions. Two $Ca^{2+}$ ions are present in Cat1 site of each molecule active site. Additional $Ca^{2+}$ ion are present in two molecules only in which could be referred to as Cat3 site in FEN superfamily that are lack Cat2 site due to the presence of the N-terminal in its position in prokaryotes FEN family. The $Ca^{2+}$ ions in Cat1 and Cat3 sites are also coordinated by Glu-276 and Glu-277 from the neighboring molecule. Through superposition of *T. brucei* active site with T5FEN native active site that has three $Mg^{2+}$ ions in Cat1 and Cat2 the active site in FEN family can be divided into two sites sharing the Cat1 site. Interestingly, the potassium-binding site could be absent in this FEN member due to the presence of Arg-41 in its equivalent place in other FEN superfamily members and Arg-241 could do the same job of the $K^{+}$ ion or this ion may have another conformations within the H3TH motif.

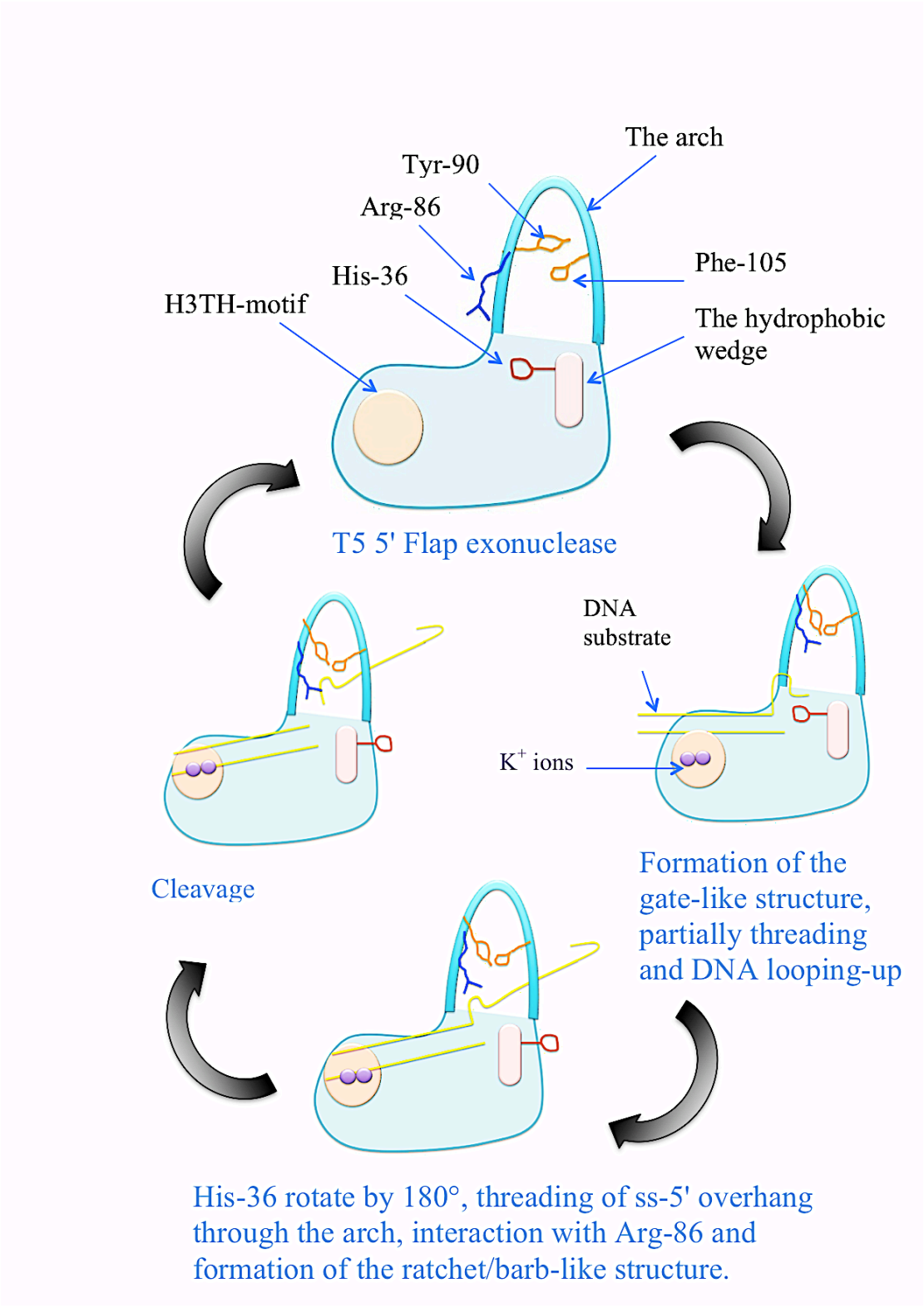



**Figure 7.1:** Diagram shows the overall mechanism of DNA binding and cleavage in T5FEN.

## 7.2 Future Work

In spite of the new findings through working with this project there are still some lines of work with in flap endonucleases proteins required to achieve the full understanding of these enzymes reaction. These investigations must be done in combination between biochemical and physical studies supported with crystallography work. Mutations in some residues that are considered to be important players in DNA binding and processing according to this project results are required. These residues include His-36, Lys-83, Arg-86, Tyr-90, Phe-105 and Arg-125 in T5FEN which can be mutated to alanine or asparagine then measure the DNA dissociation value ($K_d$) to determine the effect of each mutation through some assay techniques such as EMSA or anisotropy. Also double mutations can be investigated in the same way.

In addition, the FEN exonuclease activity is still as a puzzle matter which was posited to have another mechanism from that observed for endonuclease one (Garforth et al, 1999). Mutations in the active site conserved carboxylic residues are required for fully understand of FEN catalysis activity. Replacing some of the carboxylic residues with uncharged side chains amino acid such as alanine, serine or asparagine and then the measure both the catalysis activity and the DNA-binding can do with these mutations. Another point is to determine crystal structure for the wild type and the mutated versions of these enzymes in complexes with different DNA substrates such as double flap, pseudo-Y DNA and variety lengths of 5' or 3' overhangs. It is of interest to produce a complex with a DNA substrate having a 3'-single nucleotide and T5FEN to investigate the reaction and interaction of this type of DNA with a FEN member lacks the 3'-single nucleotide pocket. These complexes can be created with catalytically inactive forms of the enzyme or with the wild type in presence of inhibitor metal ions such as $Ca^{2+}$. Also there is a need for studying the metal ion binding sites in FEN enzymes to determine how many metal ions the active site can occupy in prokaryote and eukaryote FEN members and the role of these metal ions in catalytic reaction.

# Reference


Allawi HT, Kaiser MW, Onufriev AV, Ma WP, Brogaard AE, Case DA, Neri BP, Lyamichev VI (2003) Modeling of flap endonuclease interactions with DNA substrate. *J Mol Biol* 328: 537-554

Amado L, Kuzminov A (2013) Low-Molecular-Weight DNA Replication Intermediates in Escherichia coli: Mechanism of Formation and Strand Specificity. *Journal of Molecular Biology* 425: 4177-4191

Amblar M, de Lacoba MG, Corrales MA, Lopez P (2001) Biochemical analysis of point mutations in the 5'-3' exonuclease of DNA polymerase I of *Streptococcus pneumoniae*. Functional and structural implications. *J Biol Chem* 276: 19172-19181

Anstey-Gilbert C (2006) Structural studies on biological macromolecules: The crystal structure of *E. coli* ErnI. PhD Thesis, Molecular Biology and Biotechnology, University of Sheffield, Sheffield

Anstey-Gilbert CS, Hemsworth GR, Flemming CS, Hodskinson MRG, Zhang J, Sedelnikova SE, Stillman TJ, Sayers JR, Artymiuk PJ (2013) The structure of Escherichia coli ExoIX-implications for DNA binding and catalysis in flap endonucleases. *Nucleic acids research* 41: 8357-8367

Artymiuk PJ, Ceska TA, Suck D, Sayers JR (1997) Prokaryotic 5'-3' exonucleases share a common core structure with gamma-delta resolvase. *Nucleic Acids Res* 25: 4224-4229

Bae SH, Bae KH, Kim JA, Seo YS (2001) RPA governs endonuclease switching during processing of Okazaki fragments in eukaryotes. *Nature* 412: 456-461

Bae SH, Seo YS (2000) Characterization of the enzymatic properties of the yeast Dna2 helicase/endonuclease suggests a new model for Okazaki fragment processing. *J Biol Chem* 275: 38022-38031

Beese LS, Steitz TA (1991) Structural basis for the 3'-5' exonuclease activity of Escherichia coli DNA polymerase I: a two metal ion mechanism. *EMBO Journal* 10: 25-33

Benvenuti M, Mangani S (2007) Crystallization of soluble proteins in vapor diffusion for x-ray crystallography. *Nature Protocols* 2: 1633-1651

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235-242

Bhagwat M, Meara D, Nossal NG (1997) Identification of residues of T4 RNase H required for catalysis and DNA binding. *J Biol Chem* 272: 28531-28538

Blow D (2002) *Outline of Crystallography for Biologists*, New York: Oxford University Press.

Bornarth CJ, Ranalli TA, Henricksen LA, Wahl AF, Bambara RA (1999) Effect of flap modifications on human FEN1 cleavage. *Biochemistry* 38: 13347-13354

Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* 72: 248-254

CCP4 (1994) The CCP4 Suite - programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* 50: 760-763

Ceska TA, Sayers JR, Stier G, Suck D (1996) A helical arch allowing single-stranded DNA to thread through T5 5'-exonuclease. *Nature* 382: 90-93

Chapados BR, Hosfield DJ, Han S, Qiu J, Yelent B, Shen B, Tainer JA (2004) Structural basis for FEN-1 substrate specificity and PCNA-mediated activation in DNA replication and repair. *Cell* 116: 39-50

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D-Biological Crystallography* 66: 12-21

Chikova AK, Schaaper RM (2005) The bacteriophage P1 hot gene product can substitute for the *Escherichia coli* DNA polymerase III theta subunit. *Journal of Bacteriology* 187: 5528-5536

Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, 3rd, Snoeyink J, Richardson JS, Richardson DC (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 35: W375-383

DeLano WL. (2002) The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA.

Delucia P, Cairns J (1969) Isolation of an *E. coli* strain with a mutation affecting DNA polymerase. *Nature* 224: 1164-&

Dervan JJ, Feng M, Patel D, Grasby JA, Artymiuk PJ, Ceska TA, Sayers JR (2002) Interactions of mutant and wild-type flap endonucleases with oligonucleotide substrates suggest an alternative model of DNA binding. *Proc Natl Acad Sci U S A* 99: 8542-8547

Devos JM, Tomanicek SJ, Jones CE, Nossal NG, Mueser TC (2007) Crystal structure of bacteriophage T4 5' nuclease in complex with a branched DNA reveals

how flap endonuclease-1 family nucleases bind their substrates. *J Biol Chem* 282: 31713-31724

Diaz A, Lacks SA, Lopez P (1992) The 5' - 3' exonuclease activity of DNA polymerase I is essential for *Streptococcus pneumoniae*. *Molecular Microbiology* 6: 3009-3019

Doherty AJ, Serpell LC, Ponting CP (1996) The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. *Nucleic Acids Research* 24: 2488-2497

Drenth J (1999) *Principles of Protein X-ray Crystallography*, New York: Springer-Verlag New York Inc.

Dupureur CM (2008) Roles of metal ions in nucleases. *Curr Opin Chem Biol* 12: 250-255

Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60: 2126-2132

Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallographica Section D-Biological Crystallography* 66: 486-501

Feng M (2002) Unravelling the mechanisms of DNA cleavage as catalyzed by 5 nucleases,. PhD Thesis, Medical School, University of Sheffield,

Feng M, Patel D, Dervan JJ, Ceska T, Suck D, Haq I, Sayers JR (2004) Roles of divalent metal ions in flap endonuclease-substrate interactions. *Nature Structural & Molecular Biology* 11: 450-456

Fijalkowska IJ, Schaaper RM, Jonczyk P (2012) DNA replication fidelity in *Escherichia coli*: a multi-DNA polymerase affair. *Fems Microbiology Reviews* 36: 1105-1121

Finger LD, Blanchard MS, Theimer CA, Sengerova B, Singh P, Chavez V, Liu F, Grasby JA, Shen B (2009) The 3 '-Flap Pocket of human Flap Endonuclease 1 is critical for substrate binding and catalysis. *J Biol Chem* 284: 22184-22194

Flemming CS (2011) Structural Studies on DNA Binding Proteins. PhD Thesis, Molecular Biology and Biotechnology, University of Sheffield,

Frenkel GD, Richards.Cc (1971) Deoxyribonuclease induced after infection of *Escherichia-coli* by bacteriophage T5 characterization of enzyme as 5`-exonuclease. *J Biol Chem* 246: 4839-&

Friedrich-Heineken E, Henneke G, Ferrari E, Hübscher U (2003) The acetylatable lysines of human Fen1 are important for endo- and exonuclease activities. *J Mol Biol* 328: 73-84

Fukushima S, Itaya M, Kato H, Ogasawara N, Yoshikawa H (2007) Reassessment of the in vivo functions of DNA Polymerase I and RNase H in bacterial cell growth. *J Bacteriol* 189: 8575-8583

Garcia-Diaz M, Bebenek K (2007) Multiple functions of DNA polymerases. *Critical Reviews in Plant Sciences* 26: 105-122

Garforth SJ, Ceska TA, Suck D, Sayers JR (1999) Mutagenesis of conserved lysine residues in bacteriophage T5 5 '-3 ' exonuclease suggests separate mechanisms of endo- and exonucleolytic cleavage. *Proceedings of the National Academy of Sciences of the United States of America* 96: 38-43

Garforth SJ, Patel D, Feng M, Sayers JR (2001) Unusually wide co-factor tolerance in a metalloenzyme; divalent metal ions modulate endo-exonuclease activity in T5 exonuclease. *Nucleic Acids Res* 29: 2772-2779

Garforth SJ, Sayers JR (1997) Structure-specific DNA binding by bacteriophage T5 5'->3' exonuclease. *Nucleic Acids Research* 25: 3801-3807

Georgescu RE, Kurth I, O'Donnell ME (2012) Single-molecule studies reveal the function of a third polymerase in the replisome. *Nature Structural & Molecular Biology* 19: 113-116

Gloor JW, Balakrishnan L, Bambara RA (2010) Flap Endonuclease 1 Mechanism Analysis Indicates Flap Base Binding Prior to Threading. *J Biol Chem* 285: 34922-34931

Grasby JA, Finger LD, Tsutakawa SE, Atack JM, Tainer JA (2012) Unpairing and gating: sequence-independent substrate recognition by FEN superfamily nucleases. *Trends in Biochemical Sciences* 37: 74-84

Guo Z, Kanjanapangka J, Liu N, Liu S, Liu C, Wu Z, Wang Y, Loh T, Kowolik C, Jamsen J, Zhou M, Khue T, Chen Y, Zheng L, Shen B (2012) Sequential posttranslational modifications program FEN1 degradation during cell-cycle Progression. *Molecular Cell* 47: 444-456

Guo Z, Qian L, Liu R, Dai H, Zhou M, Zheng L, Shen B (2008) Nucleolar localization and dynamic roles of flap endonuclease 1 in ribosomal DNA replication and damage repair. *Molecular and Cellular Biology* 28: 4310-4319

Hanahan D (1983) Studies on transformation of *Escherichia coli* with plasmids *Journal of Molecular Biology* 166: 557-580

Harding MM (2002) Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium. *Acta Crystallogr D Biol Crystallogr* 58: 872-874

Harrington JJ, Lieber MR (1994a) The characterization of a mammalian DNA structure-specific endonuclease. *EMBO J* 13: 1235-1246

Harrington JJ, Lieber MR (1994b) Functional domains within FEN-1 and RAD2 define a family of structure-specific endonucleases: implications for nucleotide excision repair. *Genes Dev* 8: 1344-1355

Hasan S, Stucki M, Hassa PO, Imhof R, Gehrig P, Hunziker P, Hubscher U, Hottiger MO (2001) Regulation of human flap endonuclease-1 activity by acetylation through the transcriptional coactivator p300. *Molecular Cell* 7: 1221-1231

Hastings PJ, Hersh MN, Thornton PC, Fonville NC, Slack A, Frisch RL, Ray MP, Harris RS, Leal SM, Rosenberg SM (2010) Competition of *Escherichia coli* DNA polymerases I, II and III with DNA Pol IV in stressed cells. *PloS one* 5: e10862-e10862

Hemsworth GR (2009) Structural Studies on the Flap Endonuclease, ExoIX. PhD Thesis, Molecular Biology and Biotechnology, University of Sheffield,

Henneke G, Friedrich-Heineken E, Hubscher U (2003) Flap endonuclease 1: a novel tumour suppresser protein. *Trends Biochem Sci* 28: 384-390

Hosfield DJ, Mol CD, Shen BH, Tainer JA (1998) Structure of the DNA repair and replication endonuclease and exonuclease FEN-1: Coupling DNA and PCNA binding to FEN-1 activity. *Cell* 95: 135

Hwang KY, Baek K, Kim HY, Cho Y (1998) The crystal structure of flap endonuclease-1 from *Methanococcus jannaschii*. *Nature Structural Biology* 5: 707-713

Indiani C, O'Donnell M (2013) A proposal: Source of single strand DNA that elicits the SOS response. *Frontiers in bioscience : a journal and virtual library* 18: 312-323

Ishimi Y, Claude A, Bullock P, Hurwitz J (1988) Completeenzymatic-synthesis of DNA containing the SV40 originof replication. *J Biol Chem* 263: 19723-19733

Jeruzalmi D, Yurieva O, Zhao YX, Young M, Stewart J, Hingorani M, O'Donnell M, Kuriyan J (2001) Mechanism of processivity clamp opening by the delta subunit wrench of the clamp loader complex of E-coli DNA polymerase III. *Cell* 106: 417-428

Joyce CM, Grindley ND (1984) Method for determining whether a gene of *Escherichia coli* is essential: application to the *polA* gene. *Journal of Bacteriology* 158: 636-643

Kaiser MW, Lyamicheva N, Ma W, Miller C, Neri B, Fors L, Lyamichev VI (1999) A comparison of eubacterial and archaeal structure-specific 5'-exonucleases. *J Biol Chem* 274: 21387-21394

Kantardjieff KA, Rupp B (2003) Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci* 12: 1865-1871

Kao HI, Campbell JL, Bambara RA (2004) Dna2p helicase/nuclease is a tracking protein, like FEN1, for flap cleavage during Okazaki fragment maturation. *J Biol Chem* 279: 50840-50849

Kim Y, Eom SH, Wang JM, Lee DS, Suh SW, Steitz TA (1995) Crystal structure of *Thermus aquaticus* DNA-polymerase. *Nature* 376: 612-616

Kirby TW, Harvey S, DeRose EF, Chalov S, Chikova AK, Perrino FW, Schaaper RM, London RE, Pedersen LC (2006) Structure of the Escherichia coli DNA polymerase III epsilon-HOT proofreading complex. *J Biol Chem* 281: 38466-38471

Kleywegt GJ, Read RJ (1997) Not your average density. *Structure* 5: 1557-1569

Kolter R, Helinski DR (1979) Regulation of initiation of DNA replication. *Annu Rev Genet* 13: 355-391

Kong XP, Onrust R, O'Donnell M, Kuriyan J (1992) Three-dimensional structure of the beta subunit of *E. coli* DNA polymerase III holoenzyme: a sliding DNA clamp. *Cell* 69: 425-437

Kornberg A, Baker TA (1992) *DNA Replication*, 2nd edn. New York: W.H. Freeman and Company.

Kornberg A, Lehman IR, Bessman MJ, Simms ES (1956) Enzymic synthesis of deoxyribonucleic acid. *Biochim Biophys Acta* 21: 197-198

Kucherlapati M, Yang K, Kuraguchi M, Zhao J, Lia M, Heyer J, Kane MF, Fan K, Russell R, Brown AM, Kneitz B, Edelmann W, Kolodner RD, Lipkin M, Kucherlapati R (2002) Haploinsufficiency of Flap endonuclease (Fen1) leads to rapid tumor progression. *Proc Natl Acad Sci USA* 99: 9924-9929

Kuppuraj G, Dudev M, Lim C (2009) Factors Governing Metal-Ligand Distances and Coordination Geometries of Metal Complexes. *Journal of Physical Chemistry B* 113: 2952-2960

Laemmli UK (1970) Cleavage of structural proteins during assembly of head of bacteriophage-T4. *Nature* 227: 680-&

Laskowski RA, Macarthur MW, Moss DS, Thornton JM (1993) PROCHECK - a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* 26: 283-291

Lee KH, Kim DW, Bae SH, Kim JA, Ryu GH, Kwon YN, Kim KA, Koo HS, Seo YS (2000) The endonuclease activity of the yeast Dna2 enzyme is essential in vivo. *Nucleic acids research* 28: 2873-2881

Leslie AGW (1994) MOSFLM. *Joint CCP4 & ESF-EACBM Newsletter on Protein Crystallography* 26

Lieber MR (1997) The FEN-1 family of structure-specific nucleases in eukaryotic DNA replication, recombination and repair. *Bioessays* 19: 233-240

Liu R, Qiu J, Finger LD, Zheng L, Shen B (2006) The DNA-protein interaction modes of FEN-1 with gap substrates and their implication in preventing duplication mutations. *Nucleic Acids Res* **34:** 1772-1784

Liu P, Qian L, Sung J-S, de Souza-Pinto NC, Zheng L, Bogenhagen DF, Bohr VA, Wilson DM, III, Shen B, Dernple B (2008) Removal of oxidative DNA damage via FEN1-dependent long-patch base excision repair in human cell mitochondria. *Molecular and Cellular Biology* 28: 4975-4987

Lyamichev V, Brow MAD, Dahlberg JE (1993) Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science* 260: 778-783

Matsui E, Musti KV, Abe J, Yamasaki K, Matsui I, Harata K (2002) Molecular structure and novel DNA binding sites located in loops of Flap Endonuclease-1 from *Pyrococcus horikoshii*. *J Biol Chem* 277: 37840-37847

Matthews BW (1968) Solvent content of protein crystals. *J Mol Biol* 33: 491-497

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. *Journal of Applied Crystallography* 40: 658-674

McHenry CS (2011) DNA replicases from a bacterial perspective. In *Annual Review of Biochemistry, Vol 80*, Kornberg RD, Raetz CRH, Rothman JE, Thorner JW (eds), Vol. 80, pp 403-436.

McInerney P, Johnson A, Katz F, O'Donnell M (2007) Characterization of a triple DNA polymerase replisome. *Molecular Cell* 27: 527-538

McPherson A (1999) *Crystallization of Biological Macromolecules*: Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Menezes MR, Sweasy JB (2012) Mouse models of DNA polymerases. *Environmental and molecular mutagenesis* 53: 645-665

Moreau S, Morgan EA, Symington LS (2001) Overlapping functions of the Saccharomyces cerevisiae Mre11, Exo1 and Rad27 nucleases in DNA metabolism. *Genetics* 159: 1423-1433

Moyer RW, Rothe CT (1977) Role of T5 gene-D15 nuclease in generation of nicked bacteriophage-T5-DNA. *Journal of Virology* 24: 177-193

Mueser TC, Nossal NG, Hyde CC (1996) Structure of bacteriophage-T4 RNase-H, a 5' to 3' RNA-DNA and DNA-DNA exonuclease with sequence similarity to the RAD2 family of eukaryotic proteins. *Cell* 85: 1101-1112

Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53: 240-255

O'Donnell M (2006) Replisome architecture and dynamics in *Escherichia coli*. *J Biol Chem* 281: 10653-10656

Ogawa T, Okazaki T (1984) Function of RNase H in DNA replication revealed by RNase H defective mutants of *Escherichia coli*. *Mol Gen Genet* 193: 231-237

Oka A, Sugimoto K, Takanami M, Hirota Y (1980) Replication origin of the Escherichia-coli K-12 chromosome - the size and structure of the minimum DNA segment carrying the information for autonomous replication. *Molecular and General Genetics* 178: 9-20

Okazaki R, Arisawa M, Sugino A (1971) Mechanism of DNA chain growth slow joining of newly replicated DNA chains in DNA polymerase I deficient *Escherichia coli* mutant. *Proceedings of the National Academy of Sciences of the United States of America* 68: 2954-&

Ollis DL, Kline C, Steitz TA (1985) Domain of *Escherichia coli* DNA polymerase-I showing sequence homology to T7 DNA polymerase. *Nature* 313: 818-819

Ooi L-l (2010) *Principles of X-ray Crystallography*, New York: Oxford University Press.

Orans J, McSweeney EA, Iyer RR, Hast MA, Hellinga HW, Modrich P, Beese LS (2011) Structures of human Exonuclease 1 DNA complexes suggest a unified mechanism for nuclease family. *Cell* 145: 212-223

Otwinowski Z, Minor W (1997) Processing of X-ray Diffraction Data Collected in Oscillation Mode. *Methods in Enzymology* Volume 276: Macromolecular Crystallography, part A: 307-326

Patel N, Atack JM, Finger LD, Exell JC, Thompson P, Tsutakawa S, Tainer JA, Williams DM, Grasby JA (2012) Flap endonucleases pass 5 '-flaps through a

flexible arch using a disorder-thread-order mechanism to confer specificity for free 5 '-ends. *Nucleic Acids Research* 40: 4507-4519

Paul AV, Lehman IR (1966) Deoxyribonucleases of *Escherichia coli* a deoxyribonuclease induced by infection with phage T5. *J Biol Chem* 241: 3441-&

Pelletier H, Sawaya MR (1996) Characterization of the metal ion binding helix-hairpin-helix motifs in human DNA polymerase beta by X-ray structural analysis. *Biochemistry* 35: 12778-12787

Pickering TJ, Garforth SJ, Thorpe SJ, Sayers JR, Grasby JA (1999) A single cleavage assay for T5 5'-->3' exonuclease: determination of the catalytic parameters for wild-type and mutant proteins. *Nucleic Acids Res* 27: 730-735

Qiu JZ, Li XW, Frank G, Shen BH (2001) Cell cycle-dependent and DNA damage-inducible nuclear localization of FEN-1 nuclease is consistent with its dual functions in DNA replication and repair. *J Biol Chem* 276: 4901-4908

Reagan MS, Pittenger C, Siede W, Friedberg EC (1995) Characterization of a mutant strain of *Saccharomyces cerevisiae* with a deletion of the RAD27 gene, a structural homolog of the RAD2 nucleotide excision repair gene. *Journal of Bacteriology* 177: 364-371

Reyes-Lamothe R, Sherratt DJ, Leake MC (2010) Stoichiometry and Architecture of Active DNA Replication Machinery in *Escherichia coli*. *Science* 328: 498-501

Rhodes G (2006) *Crystallography Made Crystal Clear*, London: Academic Press.

Sakurai S, Kitano K, Yamaguchi H, Hamada K, Okada K, Fukuda K, Uchida M, Ohtsuka E, Morioka H, Hakoshima T (2005) Structural basis for recruitment of human flap endonuclease 1 to PCNA. *EMBO J* 24: 683-693

Sayers JR, Artymiuk PJ (1998) Flexible loops and helical arches. 5: 668-670

Sayers JR, Eckstein F (1990) Properties of overexpressed Phage-T5 D15 exonuclease - similarities with *Escherichia coli* DNA polymerase I 5'-3' exonuclease. *J Biol Chem* 265: 18311-18317

Sayers JR, Eckstein F (1991) A single-strand specific endonuclease activity copurifies with overexpressed T5 D15 exonuclease. *Nucleic Acids Research* 19: 4127-4132

Sengerova B, Tomlinson C, Atack JM, Williams R, Sayers JR, Williams NH, Grasby JA (2010) Bronsted Analysis and Rate-Limiting Steps for the T5 Flap Endonuclease Catalyzed Hydrolysis of Exonucleolytic Substrates. *Biochemistry* 49: 8085-8093

Setlow P, Brutlag D, Kornberg A (1972) Deoxyribonucleic acid polymerase: two distinct enzymes in one polypeptide. I. A proteolytic fragment containing the polymerase and 3' leads to 5' exonuclease functions. *J Biol Chem* 247: 224-231

Setlow P, Kornberg A (1972) Deoxyribonucleic acid polymerase: two distinct enzymes in one polypeptide. II. A proteolytic fragment containing the 5' leads to 3' exonuclease function. Restoration of intact enzyme functions from the two proteolytic fragments. *J Biol Chem* 247: 232-240

Shen B, Qiu J, Hosfield D, Tainer JA (1998) Flap endonuclease homologs in archaebacteria exist as independent proteins. *Trends Biochem Sci* 23: 171-173

Sobhy MA, Joudeh LI, Huang XJ, Takahashi M, Hamdan SM (2013) Sequential and Multistep Substrate Interrogation Provides the Scaffold for Specificity in Human Flap Endonuclease 1. *Cell Reports* 3: 1785-1794

Sommers CH, Miller EJ, Dujon B, Prakash S, Prakash L (1995) Conditional lethality of null mutations in RTH1 that encodes the yeast counterpart of a mammalian 5`-exonuclease to 3`-exonuclease required for lagging strand DNA synthesis in reconstituted systems *J Biol Chem* 270: 4193-4196

Stewart JA, Campbell JL, Bambara RA (2009) Significance of the dissociation of Dna2 by Flap Endonuclease 1 to Okazaki fragment processing in *Saccharomyces cerevisiae*. *J Biol Chem* 284: 8283-8291

Sun X, Wu P, Zheng L, Thrower D, Partikian A, Qiu J, Shen B (2002) Suppression of *Saccharomyces cerevisiae rad27* null mutant phenotypes by the 5' nuclease domain of *Escherichia coli* DNA polymerase I. *Curr Genet* 41: 379-388

Syson K, Tomlinson C, Chapados BR, Sayers JR, Tainer JA, Williams NH, Grasby JA (2008) Three metal ions participate in the reaction catalyzed by T5 flap endonuclease. *J Biol Chem* 283: 28741-28746

Taft-Benz SA, Schaaper RM (2004) The theta subunit of *Escherichia coli* DNA polymerase III: a role in stabilizing the epsilon proofreading subunit. *J Bacteriol* 186: 2774-2780

Tanner NA, Hamdan SM, Jergic S, Schaeffer PM, Dixon NE, van Oijen AM (2008) Single-molecule studies of fork dynamics in Escherichia coli DNA replication. *Nature Structural & Molecular Biology* 15: 170-176

Thayer MM, Ahern H, Xing D, Cunningham RP, Tainer JA (1995) Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *EMBO J* 14: 4108-4120

Tock MR, Frary E, Sayers JR, Grasby JA (2003) Dynamic evidence for metal ion catalysis in the reaction mediated by a flap endonuclease. *EMBO J* 22: 995-1004

Tomlinson CG, Syson K, Sengerova B, Atack JM, Sayers JR, Swanson L, Tainer JA, Williams NH, Grasby JA (2011) Neutralizing mutations of carboxylates that bind metal 2 in t5 flap endonuclease result in an enzyme that still requires two metal ions. *The Journal of Boil Chem* 286: 30878-30887

Tsutakawa SE, Classen S, Chapados BR, Arvai AS, Finger LD, Guenther G, Tomlinson CG, Thompson P, Sarker AH, Shen BH, Cooper PK, Grasby JA, Tainer JA (2011) Human Flap Endonuclease structures, DNA double-base flipping, and a unified understanding of the FEN1 superfamily. *Cell* 145: 198-211

Tsutakawa SE, Tainer JA (2012) Double strand binding-single strand incision mechanism for human flap endonuclease: Implications for the superfamily. *Mechanisms of Ageing and Development* 133: 195-202

Wang BC (1985) Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol* 115: 90-112

Weiss MS, Hilgenfeld R (1997) On the use of the merging R factor as a quality indicator for X-ray data. *Journal of Applied Crystallography* 30: 203-205

Williams RS, Kunkel TA (2011) FEN Nucleases: Bind, Bend, Fray, Cut. *Cell* 145: 171-172

Winter G (2010) xia2: an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography* 43: 186-190

Wu Z, Lin Y, Xu H, Dai H, Zhou M, Tsao S, Zheng L, Shen B (2012) High risk of benzo alpha pyrene-induced lung cancer in E160D FEN1 mutant mice. *Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis* 731: 85-91

Xu Y, Potapova O, Leschziner AE, Grindley NDF, Joyce CM (2001) Contacts between the 5 ' nuclease of DNA polymerase I and its DNA substrate. *J Biol Chem* 276: 30167-30177

Yao NY, Georgescu RE, Finkelstein J, O'Donnell ME (2009) Single-molecule analysis reveals that the lagging strand increases replisome processivity but slows replication fork progression. *Proceedings of the National Academy of Sciences of the United States of America* 106: 13236-13241

Yeeles JTP (2014) Discontinuous leading-strand synthesis: a stop-start story. *Biochem Soc Trans* 42: 25-34

Zhang J (2012) Biochemical Studies on T5 Exonuclease. PhD Thesis, Medical School, University of Sheffield, Sheffield

Zheng L, Dai H, Hegde ML, Zhou M, Guo Z, Wu X, Wu J, Su L, Zhong X, Mitra S, Huang Q, Kernstine KH, Pfeifer GP, Shen B (2011) Fen1 mutations that

specifically disrupt its interaction with PCNA cause aneuploidy-associated cancer. *Cell Research* 21: 1052-1067

Zheng L, Dai H, Zhou M, Li M, Singh P, Qiu J, Tsark W, Huang Q, Kernstine K, Zhang X, Lin D, Shen B (2007) Fen1 mutations result in autoimmunity, chronic inflammation and cancers. *Nat Med* 13: 812-819

Zheng L, Li M, Shan JX, Krishnamoorthi R, Shen BH (2002) Distinct roles of two Mg2+ binding sites in regulation of murine flap endonuclease-1 activities. *Biochemistry* 41: 10323-10331

Zheng L, Zhou M, Chai Q, Parrish J, Xue D, Patrick SM, Turchi JJ, Yannone SM, Chen D, Shen B (2005) Novel function of the flap endonuclease 1 complex in processing stalled DNA replication forks. *EMBO Rep* **6:** 83-89

Zubay G (1993) *Biochemistry*, Third edn.: Wm. C. Brown Publishers.

# Appendix1: hFEN1:DNA

Human flap endonuclease is a member of FEN superfamily and parts of FEN enzymes that has the same structure form. The structure of this protein was solved for the first time in a complex with PCNA (Sakurai et al, 2005). In addition, structures for hFEN in complexes with DNA substrate and product were obtained for the wild type and mutant versions (Tsutakawa et al, 2011; Tsutakawa & Tainer, 2012). The structure of hFEN-D179K mutant in a complex with 5ov4 DNA substrate was obtained at ~3 Å resolution. Data collection and processing statistics are written in Table 1. Human FEN-WT enzyme was modified by Jing Zhan, Medical school, University of Sheffield. The modification process included the removing of 45 amino acids from the C-terminal to prevent forming a flexible loop which known from previse structures as affect the crystallization process and mutate the aspartic acid (Asp-179) to lysine (Lys-179) to avoid metal ion binding in M2 site.

**Table Ap.1.1:** Data Collection and processing for hFEN:5ov4 DNA complex.

| **Data collection and processing** | | | |
|---|---|---|---|
| X-ray source | | Senchrotron | |
| Wavelength | | 0.9795Å | |
| Space group | | P $2_1$ $2_1$ $2_1$ | |
| Crystal system | | Orthorhombic | |
| Cell dimensions | | | |
| a (Å) | $\alpha$ (°) | 77.19 | 90 |
| b (Å) | $\beta$ (°) | 123.3 | 90 |
| c (Å) | $\gamma$ (°) | 148.6 | 90 |
| Resolution range overall (Å) | | 59.88 -3.16 | |
| Resolution range highest shell (Å) | | 3.08-3.08 | |
| Number of observations | | 178087 (13722) | |
| Number of unique reflections | | 26938 (1961) | |
| $R_{merge}$ | | 0.124 (0.997) | |
| $R_{pim}$ | | 0.048 (0.373) | |
| I/$\sigma$I | | 15.3 (2.4) | |
| Competence | | 99.9 | |
| Multiplicity | | 6.6 (7.0) | |

*Data in brackets indicates highest resolution shell.

$$R_{merge} = \sum_{hkl}\sum_{i}\left|Ii(hkl) - \langle I(hkl)\rangle\right| / \sum_{hkl}\sum_{i} Ii(hkl)$$

$$R_{pim} = \sum_{hkl}\left[\frac{1}{N-1}\right]^{1/2}\sum_{i}\left|Ii - \langle I\rangle\right| / \sum_{hkl}\sum_{i} Ii$$

Where <*I(hkl)*> is the mean intensity of the reflection.

# Appendix 2: Theory of X-ray Crystallography

## 2.1 X-ray and Macromolecular (protein) Crystallization

Since 1895 when X-rays were discovered by Conrad Roentgen they have become a powerful tool in chemistry and biology to study the molecules. X-ray crystallography is an experimental technique used to elucidate the three-dimensional structures of single crystals of macromolecules as well as of small molecules. All of the known proteins in the living cells are share an identical structural principle in that they are composed of approximately 20 different amino acids arranged in a different lengths of linear chains called polypeptides. Proteins' activities and functions depend on their three-dimensional structures that allow formation of active sites or specific binding sites. In order to understand these functions it is of great importance to investigate the protein structures at the molecular level. Biological crystallography is used to investigate the three-dimensional structures of macromolecules that consist of proteins alone as native or in complexes. Proteins, DNA, RNA and different complexes of them can form single crystals and be studied by X-ray crystallography. The protein group also can be divided into two types of proteins: globular and fibrous proteins, and of these the first type can be crystallized. The most important step in structure determination is to form single crystals for the desired protein that are large and of good quality. Protein crystallization is a complex process and not clearly understood but it can be described through two phases or stages; nucleation and growth as shown in (Figure 2.1).

The target protein is mixed with different solutions in a drop form. Each drop contains protein, buffer, and precipitant. The mixture drops are surrounded with the appropriate volume of crystallisation solution and then are incubated at low temperature. Nucleation is the first step in crystallization and here protein molecules transfer from a freely diffusing state to an ordered crystalline one. The concentration of precipitant in the surrounding crystallization solution is higher than in the mixture drop which causes

solvent molecule migration from drop to solution. Consequently, the concentration of the protein in the drop increased and it becomes supersaturated in an unstable stage. Small aggregates of protein or crystal nuclei will appear which may precipitate or grow to large single crystals in the next step of crystallization. The second stage is crystal growth which allows crystal nuclei to grow in all three dimensions by adding ordered layers of molecules over the nucleation centres at lower levels of supersaturation. Thus protein crystals move through phase space from undersaturation to supersaturation to induce nucleation and then encourage growth by reducing supersaturation (Figure 2.1).

Living systems are based almost entirely on aqueous chemistry within narrow ranges of temperature and pH. On the basis of this fact, biological macromolecular crystallization is affected by a variety of environmental factors that can influence the ultimate sizes and quality of crystals produced or may affect the probability of crystals occurring at all. The temperature, pH and the type as well as the concentration of precipitants and of the target protein are the most important and influential elements in the crystallization process and the production of good crystals consists in optimizing or balancing these factors (McPherson, 1999); (Benvenuti & Mangani, 2007); (Ooi, 2010).

### 2.1.1 Sample Preparation

The sample of the target macromolecule that will be used in the crystallisation process should be well prepared. An appropriate and good method of purification is needed to obtain a pure and homogeneous sample of target protein. The sample must usually be more than 90% pure and should be at a high concentration (often >10mg/ml).

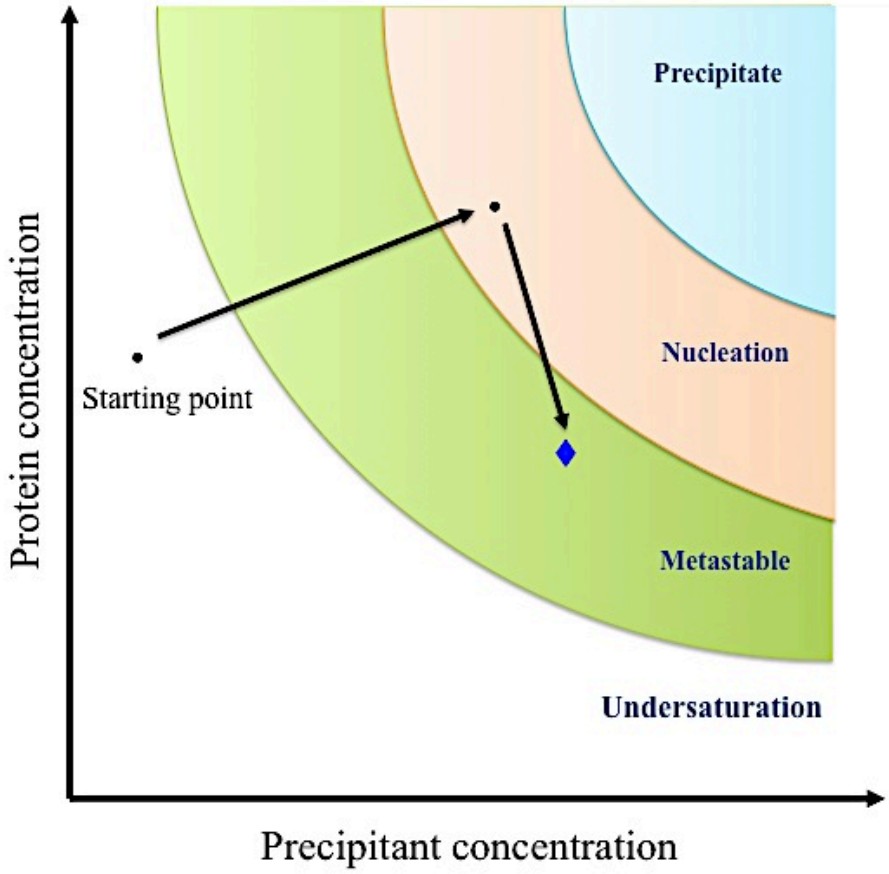



**Figure 2.106:** A theoretical phase diagram for protein crystallization growth.
The schema shows the protein and precipitant concentrations during a crystallization experiment. The arrows indicate the protein movement from an undersaturated to a supersaturated state forming the crystals.
Adapted from (Rhodes, 2006).

The desired protein has to be soluble, folded and stable, and its activity can be checked using assays to confirm this before start growing crystals.

### 2.1.2 Crystallization Methods

Several techniques have been developed and are used to grow protein crystals. A detailed description of some common methods will be illustrated below.

#### *2.1.2.1 Batch Method*

In this method the concentrated protein solution is simply mixed with the precipitant solution in a plate well. To prevent water evaporation the plate wells of solution mixtures should be covered with oil. In the batch technique protein can pass to a highly supersaturated state quickly which may form single crystals. However, this method is the oldest one that used in crystallography labs and was not used during this project.

#### *2.1.2.2 Vapour Diffusion Method*

Vapour diffusion method which is also known as isothermal distillation, is the most common method and is widely used in macromolecular crystallization. In this method a mixture drop is formed by mixing the target protein with crystallization buffer in typically a 1:1 ratio in a closed container containing a reservoir of the precipitant. Initially, the concentration of the crystallization buffer is equal to that in the reservoir but is reduced to a half after being mixed with protein. Consequently, water/solvent vapour diffuses out of the mixture drop into the reservoir to equalize the precipitant concentration between the drop and the reservoir. The precipitant and protein concentrations increase in the drop moving the protein to reach supersaturation and start forming crystals. Vapour diffusion includes two methods sharing the same principle of action, but differing in experimental application. These are called the hanging drop and the sitting drop.

#### 2.1.2.2.1 The Hanging Drop Method

Hanging drop vapour diffusion is the most popular technique in macromolecular crystallisation. In this method coverslips are treated with a siliconising solution to conserve the surface tension of the mixture drop that will hang from these glass covers. Between 1 µl to 10 µl volumes of target protein can be mixed with crystallisation solution in a 1:1 ratio and then inverted over a well containing 500 µl to 1000 µl of crystallisation buffer. The coverslip are then stuck on the well rim using a layer of oil or grease to create an airtight seal area between mixture drop and reservoir, and allowing the vapour diffusion to occur (Figure 2.2a).

#### 2.1.2.2.2 The Sitting Drop Method

Sitting drop is the second type of vapour diffusion method and it has the same principle. Here the drop is set up on a microbridge while the well around the drop is filled with the precipitant solution and then the system is sealed by a layer of oil or grease and a coverslip placed on the top of the well (Figure 2.2b). One of the benefits of this method is that the volume of the drop is unlimited while this limitation appears in the hanging drop technique. In addition, some precipitant solutions have low surface tension that make them spread out on the coverslip with the hanging drop method which makes the sitting drop technique preferable for these solutions. Furthermore, this method is easily used to test a large numbers of crystallization conditions automatically by robots to determine appropriate crystallization conditions.

#### *2.1.2.3 Dialysis Method*

The principle of this method depends on using a semi-permeable membrane which allows small molecules to pass through it but prevent the macromolecules from doing so. The protein sample is placed in a microdialysis button chamber, which is usually used in this technique, and then covered with an appropriate semi-permeable membrane before being submerged in the crystallization solution (Figure 2.2c). The membrane should be selected according to the size of its pores to be suitable for the protein

molecular weight to remain in the chamber while allowing diffusion of the precipitant through the membrane into the sample. When the concentration of the diffusible molecule is equal on both membrane sides equilibration is achieved. The protein sample transfers from an undersaturated state to supersaturated one and crystals will hopefully grow. The dialysis button can be moved between different crystallization solutions which is the major advantage of this technique.

## 2.2 X-ray Generation

Proteins or macromolecules are composed of arrangements of atoms joined by bonds. Bonding distances in protein typically 1.5 Å to 3.5 Å and therefore to position the atoms we need to be able to measure these short of distances and hence use X-rays of a comparable wavelength. X-rays are a kind of electromagnetic radiation which can range in wavelength from 0.1 Å to 1000 Å and thus can be used to determine the precise atomic positions as well as the bound lengths and angles of molecules within a single crystal. There are two types of X-rays one of them known as ‘soft’ X-rays have wavelengths of more than 2 Å and are significantly absorbed by air and strongly by water. However, the second kind of X-rays are called ‘hard’ X-rays which have shorter wavelengths between 0.6 Å and 1.6 Å and this kind is most suitable for crystallography studies, and not significantly absorbed by air or water. For crystallography experiments, X-rays are produced using a rotating anode generator or generated at synchrotrons.

### 2.2.1 Rotating Anode Generators

This system of X-ray generator has superseded the sealed X-ray tube which is composed of a vacuum with two ends: one of them has a heated cathode while an anode presents in the other end within a sealed tube. The metal cathode emits electrons that move towards and hit a rotating metal anode at high speed. These electrons are accelerated between the cathode and the anode by the large potential difference (usually between 30-50 kV). The energy of the electrons that hit the anode is transferred to the electrons in the

orbitals of the metal atoms in the anode which cause electron transitions. In this case, the lower energy orbital electrons can be released from the atom (e.g. the K-shell) and then replaced by electrons from the higher energy level (e.g. the L- shell). As a result, energy is released by the atoms as heat and photons of electromagnetic radiation in the X-ray region. The wavelength of this radiation is affected by the metal type of the anode and in crystallography labs, copper is the most commonly used due to its high heat conductivity which allows rapid dissipation of the heat generated during this process.

A copper anode can release two X-ray wavelengths (Figure 2.3): one of them produces X-ray of wavelength 1.542 Å that is the result of L to K shell electron transitions and is known as Cu Kα radiation, while the other one from M to K shell produces a shorter wavelength X-ray called Cu Kβ radiation. For crystallographic experiments, X-rays of a single wavelength are used. Consequentially, the primary X-ray radiation is monochromated using a crystal monochromator, e.g. a graphite crystal, to filter out unwanted X-rays, e.g. Cukβ, which gives X-rays of a single wavelength shorter than 1.542 Å. However, a molybdenum anode produces X-rays with a wavelength of 0.71 Å for shorter wavelength purpose.

### 2.2.2 Synchrotron Sources

Researchers were aware of the need to find other sources that can produce much more intense X-rays and since the 1960s synchrotrons were established to use in crystallography experiments. Synchrotrons are large particle accelerators in which electrons are accelerated in a circular vacuum to near the speed of light using electric fields and are forced to alter direction using a magnetic field to produce electromagnetic radiation which is known as synchrotron radiation. Synchrotrons are polygonal in shape (Figure 2.4) and made of straight sections and the bending magnets are lie in the junction areas. The electrons are stored in an auxiliary magnetic storage ring to circulate them numerous times to emit radiation at a variety of frequencies.

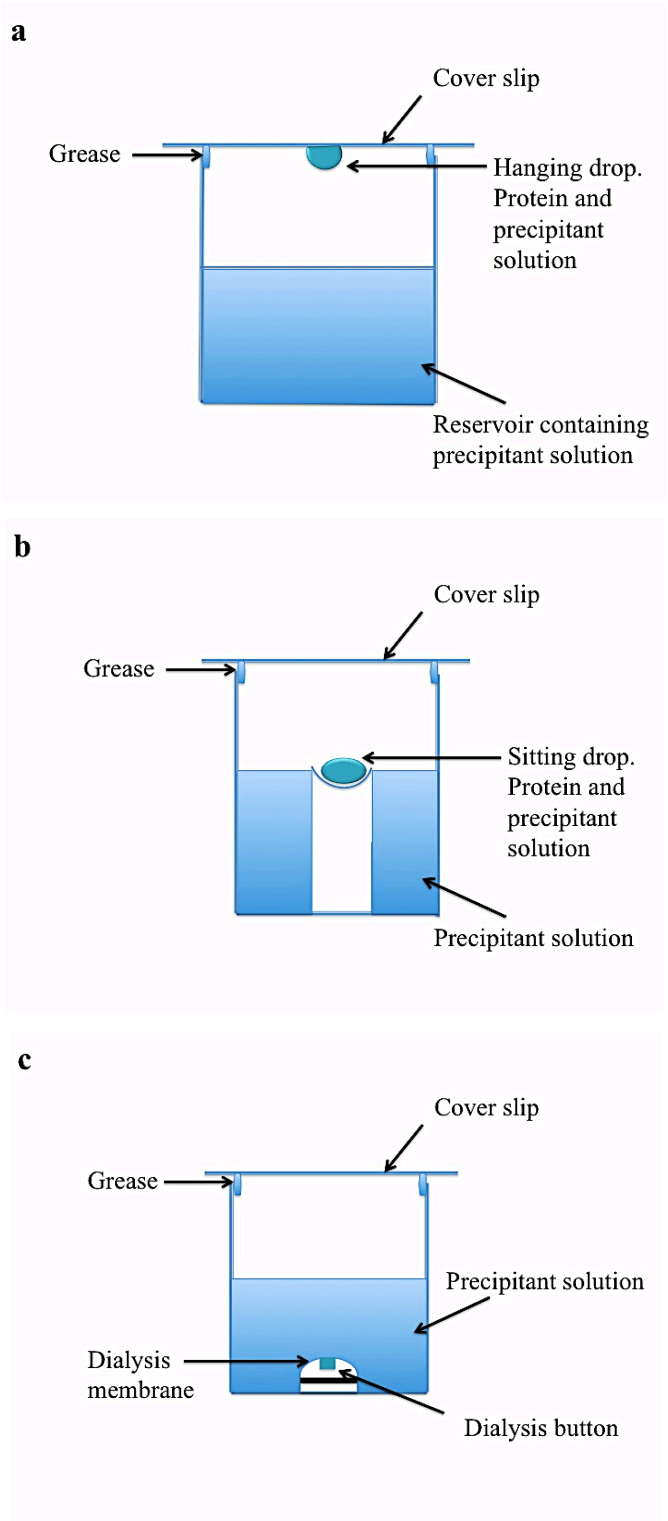
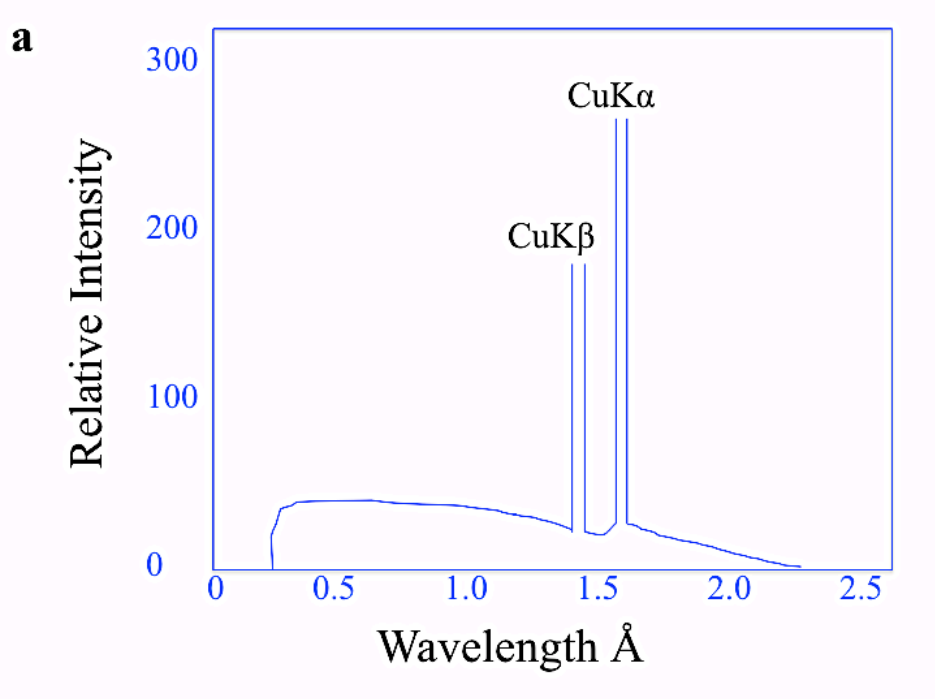



**Figure 2.107:** Diagram illustrating the most common methods used for growing protein crystals.
**(a)** Hanging drop while **(b)** is sitting drop vapour diffusion and **(c)** dialysis method.



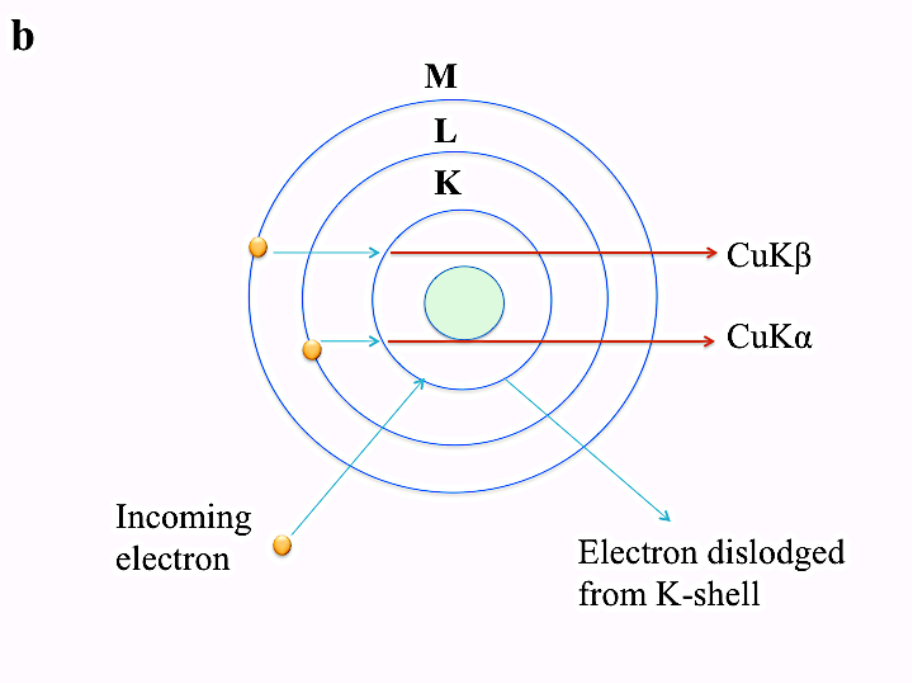




**Figure 2.108:** Diagram for X-ray rotating anode generator.
**(a)** Shows an X-ray spectrum generated by a copper anode bombarded with 50keV electrons. **(b)** Shows the three orbitals, M, K and L in an atom and how X-ray produce. Adapted from (Blow, 2002).

The bending magnets deflect electrons which changes their direction and hence acceleration and so electromagnetic radiation is released in the form of X-rays. The straight sections in the storage ring contain other magnets called "insertion devices" which are used to produce more intense X-rays (Figure 2.4). Wigglers and undulators are widely utilised insertion devices that cause the electrons to fluctuate their direction multiple times emitting X-rays beams of significantly higher intensity. Undulators produce more intense X-rays because each one has the magnetic poles closer together which cause more transitions of the electron along its path. On the other hand, wigglers provide a wide fan of X-rays and can supply several end stations. A Synchrotron produces multiple wavelength X-rays and then the desired wavelength is obtained by the use of the appropriate monochromators which filters out unwanted wavelengths. The important benefit from using synchrotron radiation in the field of crystallography is in phase determination to collect data and determine the phase angles for proteins that contain inserted heavy atoms in their structure. The wavelength of the X-rays can be selected such that the beam conforms to the absorption edge of scattering heavy atom such as selenium. On the other hand, synchrotrons are very expensive facilities and need a large space which limits their numbers around the world.

## 2.3 X-ray Diffraction

X-rays interact with the electron clouds around atoms. They can be thus diffracted from ordered arrays of molecules in crystals where the crystals act to amplify the signal. X-ray is scattered by the protein atoms when it hits the crystal causing the diffraction patterns, by measuring these scattering, it is possible to calculate an image of the crystal contents. X-rays are electromagnetic radiation and therefore consist of photons of energy which can be regarded as either particles or electromagnetic waves with specific wavelength. X-rays can interact with the atoms of the protein crystal through two ways: coherent and incoherent. When X-rays in wave form hit the crystal, the electrons of the crystal atoms oscillate with an identical frequency of this wave.

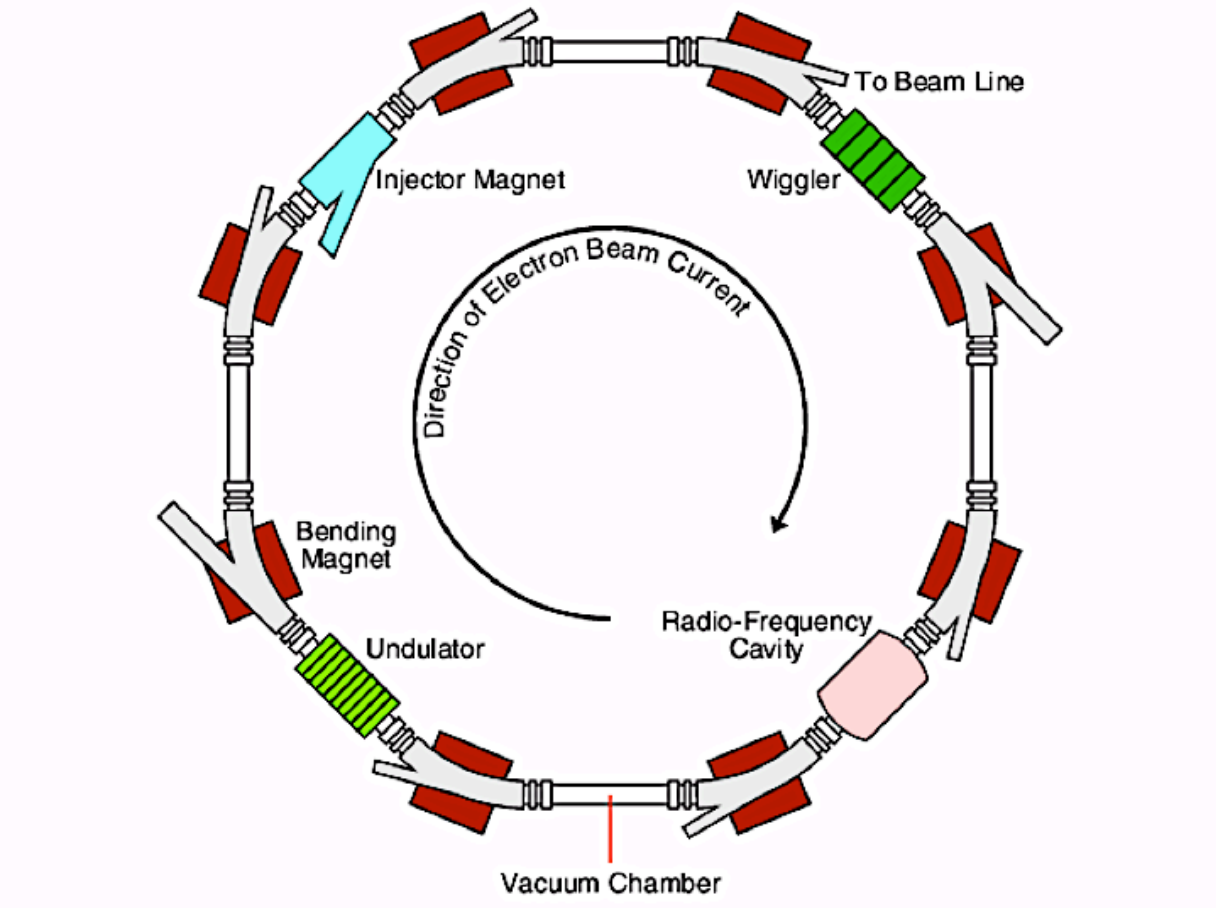



**Figure 2.109:** A synchrotron radiation source and its components.
Red parts represent the bending magnets and in some straight sections in-between them. Two devices are shown in green and light green colors for wigglers and undulators respectively. The light blue color illustrates an injector magnet which injects electrons in discrete pulses to exist inside the storage ring in gray.
Figure is from (http://www.pd.chem.ucl.ac.uk/pdnn/inst2/ring.gif).

When these electrons return to their original a secondary wave of a new X-ray is emission. The new wave has the same wavelength as the original X-ray beam but is 180° out of phase which is known as coherent scattering. On the other hand, incoherent scattering (which is also known as Compton scattering) is the other way that X-rays can interact with the electrons of the crystal atoms. In Compton scattering, the X-ray encounters with loosely bound electrons which diverts the X-ray from its original path causing loss of energy and change of the wavelength. Macromolecular structure determination depends upon coherent scattering while incoherent scattering does occur but it is small and is ignored.

Additionally, there are two types of interference for the scattered wave: constructive interference which is in phase or destructive interference that is out of phase (Figure 2.5).

### 2.3.1 Bragg`s Law

In 1913, William and Lawrence Bragg derived a simple equation explaining X-ray diffraction by crystals and since that time it has been known as Bragg`s law. Crystals are composed of numerous repeating units that are known as unit cells and can reflect the X-ray beam. Consider a single crystal with aligned planes of lattice points which are separated by a distance d. X-ray beam hits the crystal at an angle θ on atoms Z and B (Figure 2.5). The beam strikes atom Z and also passing on to the next layer and being scattered by atom B. For the two parts of the beams to be in phase and parallel and must remain in phase which means the extra distance must be an integral number of wavelengths (nλ) and can be written as:

$$AB + BC = n\lambda$$

From (Figure 2.6) *d* is the hypotenuse of the triangle ABZ and by using simple trigonometry *d* and θ can be related to the distance AB and presented as:

$$AB = d\sin\theta$$

And since $AB = BC$

Then $n\lambda = 2AB$

Which gives $n\lambda = 2d\sin\theta$

Thus, Bragg`s law has been derived. A diffraction pattern can be formed when a wavelength (λ), interplanar distance (*d*) and glancing angle (θ) have an appropriate relationship. The architecture of protein crystal is composed of many unit cells that are can be broken down into imaginary parallel planes known as Miller planes or Miller indices (*h, k, l*). These can be identified by a set of three integer values.

### 2.3.2 Atomic Scattering

The measure of X-ray waves scattered from an atom is known as atomic scattering factors and can be described by atomic form factors (*f*). As shown in (Figure 2.7) it varies with two *sinθ/λ* and it depends on the number of the electron in the atom and their distribution in the electron density cloud. For low scattering angle which is equal to zero, the atomic scattering factor is directly proportional to the atomic number, while it decreases at higher angles. The overall scattered wave from an atom, *f,* is thus given by the sum of the individual contributions over the volume of the atom.

$$f = \int_r p(r)\exp(2\pi r \bullet S)dr$$

Where $\rho$ is the electron density, **r** is the distance from the centre of the atom and **S** is the scattering vector.

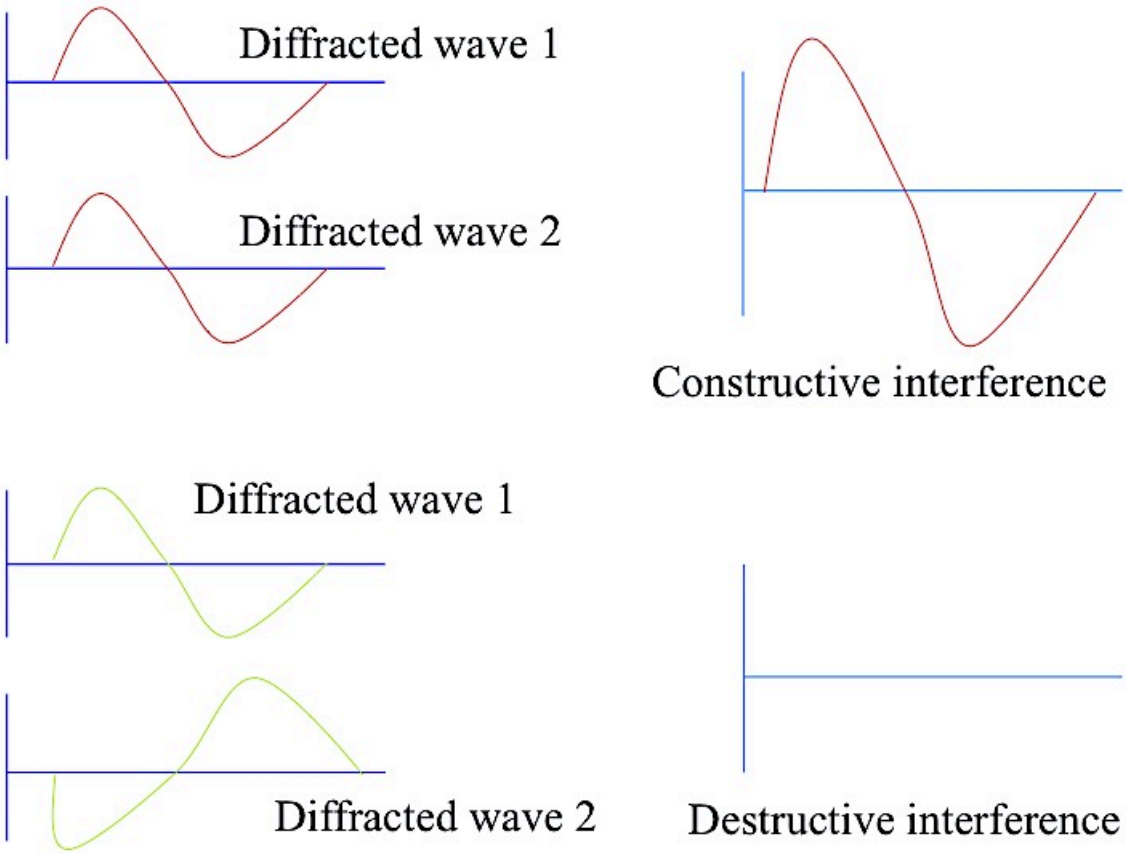



**Figure 2.110:** Constructive and destructive interference.
Figure was adapted from www.asdlib.org/onlineArticles/ecourseware/Bullen_XRD/

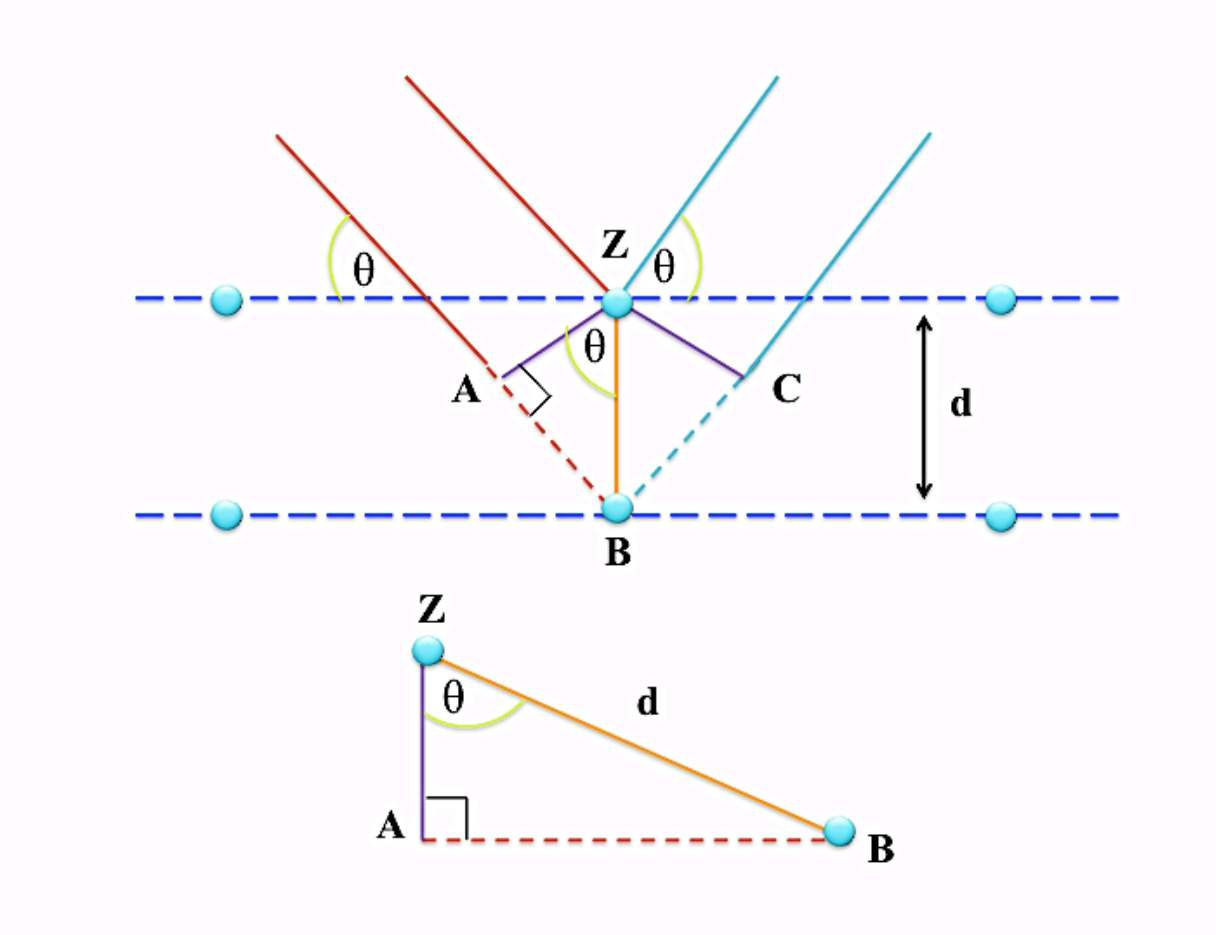



**Figure 2.111:** Bragg`s law derivation using the reflection geometry and trigonometry.
The two paths of the X-ray beams scattered from Z and B points produce a diffracted beams. The lower one travels an extra distance (AB + BC). d is the distance between Z and B and is the hypotenuse of the right-angled triangle ABZ.

In an atom, the electron density decreases with distance from the nucleus and therefore the factor ($f$) drops. Additionally, fluctuations in the atom position will decrease scattering at high angles, which is characterized by the temperature or B-factor. Also heavy elements contain more electrons which affect the scattering factor value.

### 2.3.3 Structure Factors

Formation of an image for the crystal structure in three-dimensional using the scattered ray depends on three factors: the direction, the amplitude and the phase. As it explains above, the atomic scattering factor only describes the amount of ray diffracted by an atom with no phase information. Structure factors of any X-ray reflections are a mathematical function which describes both the amplitude and the phase of the scattered X-ray. To consider this, the wave can be represented as a vector with an angle and magnitude relative to a fixed origin or equivalently, as a complex number.

The summation of waves is dependent on their amplitude and phase, so indicating the waves as vectors simplifies the calculation of the resultant wave.

The structure factor of a scattered wave from a single atom is given by:

$$F = f \exp(2\pi i r \bullet S)$$

Where $f$ is the atomic scattering factor, **r** is its position relative to a fixed origin and **S** is the scattering vector. The last part of equation ($2\pi i\mathbf{r}\bullet\mathbf{S}$) gives the phase difference between the wave scattered from position **r** and the origin.

The intensity of the reflection given by a crystal can be represented as a quantity by the structure factor calculation which plays a critical role in the solution and refinement of the protein structure.

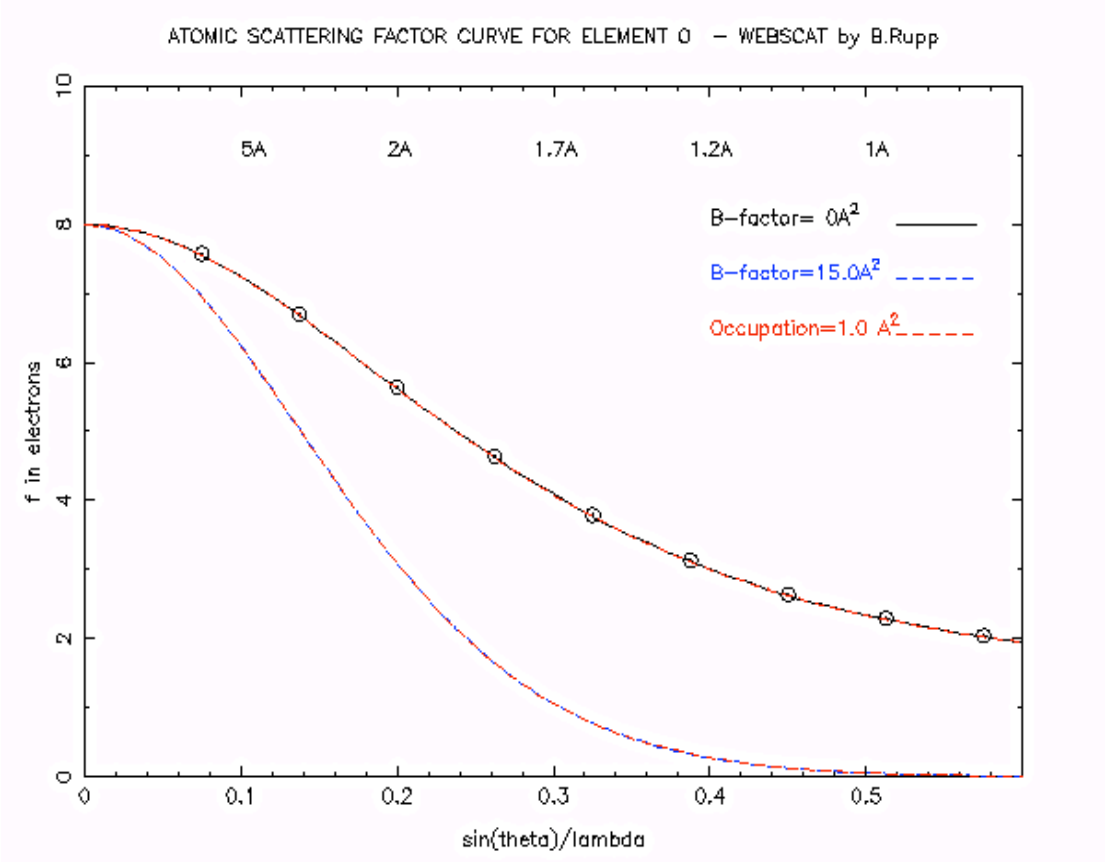



**Figure 2.112:** Atomic scattering factor curve for an oxygen atom with increasing scattering angle.
It can be seen that the scattering factor at zero scattering angle is equal to the number of electrons in the atom, which is 8 for an oxygen atom. The amount of scattering decreases with increasing scattering angle (blue dashed line), and it is further dependent on the B-factor of the atom (black line).
Generated from (http://www.ruppweb.org/xray/comp/scatfac.htm).

### 2.3.4 Molecule Scattering

Any molecule is composed of many atoms and each atom will scatter the X-ray beam individually. The structure factor for a molecule **G(S)** can be calculated by summing the structure factors from individual atom`s scattering.

$$G(S) = \sum_{j=1}^{atoms} f_j \exp(2\pi i r_j \bullet S)$$

Where $2\pi r_j$ **•S** equals the phase difference between a wave scattered from a molecule at position $\mathbf{r}_j$ and the origin.

### 2.3.5 Crystal Scattering

Crystals are built of unit cells which are the basic building block for each crystal. One molecule or more can be found within every unit cell. Within a unit cell, the asymmetric unit is the smallest portion that can be rotated and translated to generate one unit cell using the symmetry operators.

The amplification of the scattered X-rays by a crystal can be obtained if the unit cells in the crystal scatter X-rays in phase. In one dimensional scattering, the phase difference between unit cells is given by $2\pi$**a•S.** If this difference is equal to $2\pi$**h** where ***h*** is an integer, the adjacent unit cells will scatter in phase or, in other words, if **a•S** = ***h***. Each crystal has translation vectors **a**, **b** and **c** that describes the repeating distances between its unit cells. The values of translation vectors are given as **a•S** = ***h***, **b•S** = ***k*** and **c•S** = ***l*** considered to each reflection (spot) in a diffraction.

Where ***h***, ***k*** and ***l*** are integers. These are called the Laue equations. The position vector $r_j$ that is used in the molecular structure factor equation can be given as fractional coordinates of the unit cell $\boldsymbol{ax_j} + \boldsymbol{by_j} + \boldsymbol{cz_j}$. By combining this and the Laue equations the structure factor for the unit cell can be calculated as:

$$F(hkl) = \sum_{j=1}^{N} f_j \exp[2\pi i(hx_j + ky_j + lz_j)]$$

Where $x_j$, $y_j$ and $z_j$ are the fractional coordinates of the $j^{th}$ atom in the unit cell. The *hkl* terms represent the index of the reflection and refer to a given set of crystal lattice planes. If the positions of all the atoms in the unit cell are known the diffraction pattern of any crystal can be calculated using this equation. The above equation is a summation of waves, so in order to represent F(*hkl*) as a single resultant wave, it can be calculated as:

$$F(hkl) = |F(hkl)| \exp[i\alpha(hkl)]$$

Where $|F(hkl)|$ represents the structure factor amplitude while $\alpha(hkl)$ represents its phase.

## 2.4 Diffraction Detection

To date, there is no a known scientific method that can be used to measure the phase of diffracted X-rays but it is possible to measure their position and intensity. Image plate detectors and charged coupled devices (CCD camera) are two types of detectors most commonly used for collecting X-ray diffraction in protein crystallography.

### 2.4.1 Image plate Detectors

On an in-house X-ray source the type of detector that used is an image plate detector. They are composed of plastic sheets coated with a thin layer of small barium halide phosphor crystals (BaFBr) covered by bivalent europium ($Eu^{2+}$). These materials can be stimulated when struck by X-ray radiation and $Eu^{2+}$ can become excited into the metastable state $Eu^{3+}$. In this situation, red laser light of 633 nm wavelength is then used in a detector scanning process that causes the metastable $Eu^{3+}$ to relax back to the $Eu^{2+}$ state. As a result, a 390 nm wavelength radiation in proportion to the intensity of the absorbed X-

ray is emitted. Image plate detectors use a fine laser coupled to a photomultiplier to read the diffraction pattern and then create a diffraction image. After that the whole plate can be erased by illumination with visible light returning the europium back to its $Eu^{2+}$ state and reused again (Figure 2.8a).

Image plate detectors have the advantages that they are reasonable in cost, have a high spatial resolution and a high recording potency and also they are available in large sizes. The major disadvantage for this type of detector is that they require about 2 minutes to read the plate and then form the diffraction image after each exposure. The importance of this issue clearly appears in collecting data at room temperature. The crystal can be damaged and its lifetime reduced by the generating of free radicals as a result of X-ray exposure which can prevent collecting enough data and solving the structure.

### 2.4.2 Charged Coupled Devices (CCD)

The first charged coupled device detectors (CCDs) were introduced in 1994 and since then they have become widely available. This type of detector is commonly found in digital cameras and camcorders but in crystallography laboratories it is still quite rare as an in-house detector. However, while it is almost the only detection device used at synchrotron stations. When an X-ray proton impinges on the camera it is converted into visible light by the phosphor coated face of the CCD detectors (Figure 2.8b). The visible light is then processed to form an electric charge image that can be digitized. CCD detectors have a lot of advantages compared with image plate detectors but the major one is that the read-out time is significantly reduced to few seconds or part of second. Some of the disadvantages that can be reported against CCD detector are that they are expensive and their maintenance cost is large; also there are high computer requirements in processing power and data storage.

### 2.4.3 Cryoprotectant

Macromolecule crystals are composed of protein (possibly also DNA or RNA) and solvent, which includes water molecule, so that the proteins stay hydrated all the time. Once a protein crystal has grown and the composition of the components is known, the crystal should be prepared for collecting the data. Usually, data are collected at low temperature (typically 100 K) under a stream of nitrogen gas to prevent the free-radicals damage that is caused by X-ray beam. Prior to flash-freezing, crystal it is necessary to treat protein crystals with cryoprotectant that forms an amorphous glass around the crystal. Cryoprotectant must be characterized with two things: does not form ice ring and should not damage the crystal. Consequently, cryoprotectant is often made up of crystal's precipitant buffer mixed with a suitable percentage of anti-freezing solutions such as glucose, sucrose, ethylene glycol, MPD, glycerol and PEG 400-20000 at an appropriate percentage concentration. Furthermore, cryoprotectant solution should be tested alone without the crystal against the ice ring formation. This test is made by doused the loop in the cryoprotectant and then subjected to X-ray. A diffracted image can be collected and checked by eyes for any ice ring appearance. When the good cryoprotectant solution is obtained it can be used directly and start collecting data.

### 2.4.4 Crystal Mounting

For crystal mounting tiny loops about 0.05-1.0 mm in diameter are used. These loops joined to metal pins on magnetic bases which can be easily mounted onto the X-ray apparatus. The chosen loop should be appropriate to the crystal size and then immerse in the crystal growth buffer close to the protein crystal. The crystal is then picked up by the loop and held through surface tension mechanism. The cryoprotectant drop is ready and the crystal is placed into this drop by touching the loop to it and usually only left in the cryoprotectant drop for few seconds then picked out again. After that the loop carrying the crystal should immediately plunge into liquid nitrogen or the nitrogen gas stream of the data collection apparatus.

### 2.4.5 Mosaicity

The mosaicity of a crystal is a measure of the misalignment between crystal blocks containing several or many unit cells. If the mosaic is ranged from 0.1-0.2 degrees the crystal can be described as a good one while crystals with more than 2.0 degrees mosaic are bad crystals. Additionally, crystal with zero degree mosaicity is considered as a perfect one. Crystal mosaicity can be affected by some factors such as poor handling and flash-freezing of crystal. For X-ray crystallography, data still can be collected from crystals with high mosaicity but it is not good as that collected from low mosaicity crystal. This is due to the spread out of the intensity of the spot which make the spots appears smeared on the detector surface.

## 2.5 Data Processing

Data processing in protein crystallography is the way by which sets of raw diffraction data is converted into a list of Bragg reflections and are used to calculate an electron density map. Throughout this work computational suites of programs were used such as MOSFLM (Leslie, 1994), Xia2 (Winter, 2010) and the CCP4 (CCP4, 1994). A brief description for each step will be explained in the following sections.

### 2.5.1 Auto-Indexing (iMOSFLM)

The auto-indexing process for images involves the determination of the crystal orientation, lattice type, and possible point group and cell parameters. For this purpose the MOSFLM (Leslie, 1994) or Xia2(Otwinowski & Minor, 1997) (Winter, 2010) programs were used in this project. A spot search is run on the image and each spot is assigned an index consisting of three integers *h, k* and *l.* This is done by calculating a prediction of the diffraction image based on the cell dimensions and orientation and then attempting to fit the real image with the predicted one. A list of unit cell dimensions and possible lattice types for the observed diffraction pattern is estimated automatically in this stage.

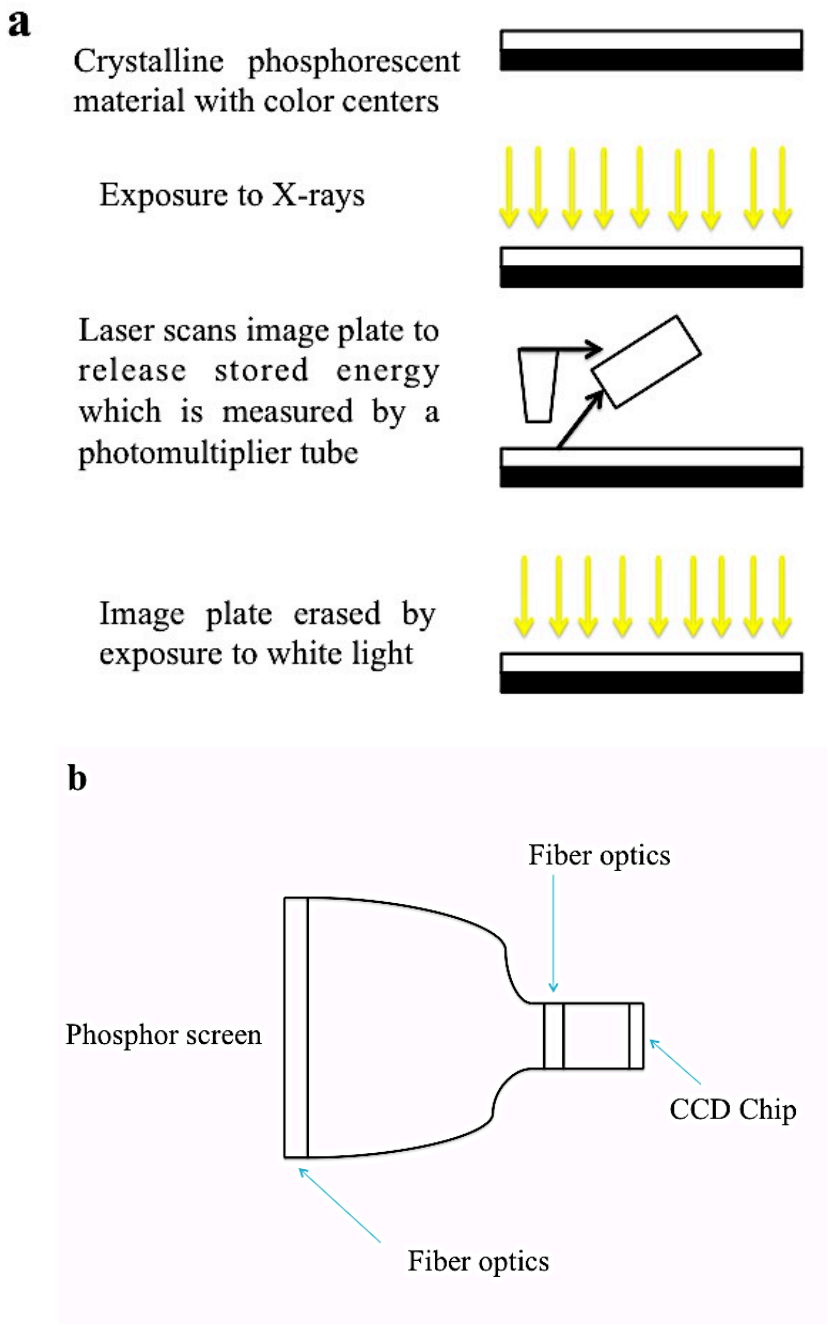



Figure 2.113: Diagram showing two types of X-ray detectors.
(a) Image plate detectors. (b) CCD detector. Adapted from (Drenth, 1999).

A single diffraction image is sometimes enough to identify the lattice type and cell dimensions correctly while using a series of images, or images that are 90° apart, will give better results. At this time, a decision must be taken for which is the most likely indexing matrix. Following this an estimation of the mosaicity of the crystal is made and the program then refines the orientation and cell parameters and displays the matrix on the diffraction image. Also the program suggests a strategy for data collection that the crystallographers can follow for collecting data.

### 2.5.2 Data Integration (iMOSFLM)

Integration of data is the process of determining the intensity of the reflections on each collected image with subtracting of the detector background from the observed diffraction spot. Two techniques help to obtain this: 'mask and count' and 'profile fitting'. The first method 'mask and count' can be used whereby a spot is located and masked, the pixels constituting the spot are summed and the background that is assigned from the adjacent pixels is subtracted to obtain the final intensity. This procedure works efficiently with strong reflections but it appears less successful with weak reflections which opens the way for the second process 'profile fitting'. In this method a standard profile for spots in each of several areas on the detector is generated using strong spots. These profiles are then fitted to each reflection in each image and then integrated to produce a list of the intensity of each reflection. The final result of the integration procedure is a single *.mtz file that is used in subsequent data processing.

### 2.5.3 Scaling Data and Merging

SCALA (CCP4, 1994) is used to combine reflections from different images and calculate a scaling factor between these images. The importance of these scale factors is that they account for the quality against some influences such as crystal and air absorbance, X-ray radiation damage to the crystal, beam

intensity fluctuations and detector sensitivity. The first image of data collection is assumed to be not affected by the above physical factors and as so it is used as a standard image to derive the scale factors. SCALA produces an output log file containing the final data statistics that can be used to evaluate the quality of the dataset and suitable to use for phasing such as $R_{merge}$, completeness, a list of indices (*h, k, l*), and signal to noise ratio (I/sig(I)). The main indicators of data quality from merging and scaling is $R_{merge}$ which measures the agreement between the independent measurements of symmetry related reflections in a data set as given by the equation:

$$R_{merge} = \frac{\sum_N \left[ \sum_N \left( I_i(hkl) - \bar{I}(hkl) \right) / n \right]}{\sum_N \bar{I}(hkl)}$$

Where $I_i(hkl)$ is the $i^{th}$ measurement of reflection *hkl* and $\bar{I}(hkl)$ is the mean value of the *n* equivalent reflections.

$R_{merge}$ depends on the redundancy of the data, which means that low redundancy records low $R_{merge}$ but the data are less accurate. This issue with $R_{merge}$ can be overcome by a precision-indicating R factor ($R_{pim}$) which gives the precision of the averaged measurement of a full data set (Weiss & Hilgenfeld, 1997) and can be calculated as:

$$R_{pim} = \sum_{hkl}[(1/N-1)]^{1/2} \sum_i |I_i(hkl) - I(hkl))| / \sum_{hkl} \sum_i I_i(hkl)$$

Completeness, which ideally should be at least 90%, clarifies the amount of available data that have been collected for a given resolution. Another crucial indicator for data quality is I/sigI which indicates the signal to noise ratio of the data and should be > 1.0.

## 2.6 Phase Determination

The raw collected data is used to calculate an electron density map to solve the protein structure which depends on the phase determination. The determination of the phase of the structure factors cannot be measured

directly but it can be obtained using experimental methods including multi-wavelength anomalous dispersion (MAD), multiple isomorphous replacement (MIR), or molecular replacement (MR). MR method can be used if a similar molecule`s structure has been experimentally determined.

In this project all structures were determined using molecular replacement procedure which will be described in the following sections.

### 2.6.1 The Patterson Function

Structural information can be obtained without any phase information using the Patterson function $P(u\ v\ w)$. It can be calculated by setting the phase angle to zero and using the intensities as coefficients:

$$P(u\ v\ w) = \frac{1}{V} \sum_{hkl} \left| F(hkl) \right|^2 \cos[2\pi(hu + kv + lw)]$$

Where $u$, $v$ and $w$ are Patterson unit cell coordinates while the real unit cell coordinates are $x$, $y$ and $z$. In the Patterson unit cell a peak at ($uvw$) will correspond to a vector between atoms at ($x, y, z$) and ($x+u, y+v, z+w$) in real space. As the measured intensities are equal to the structure factor squared, the Patterson function can be calculated directly from the diffraction data. It is the convolution of the structure with itself that results in the peaks formation. These peaks are the interatomic distance vectors sized by the product of the number of electrons in the atoms. They also represent vectors between atoms both within the same molecule and between different molecules. For example, if N is the atoms number of a molecule, $N^2$ is the number of peaks for intra-molecule vectors that will be observed in the Patterson function. Additionally, the repeating arrangement of molecules in the crystal lattice produces vectors between molecules. In small molecules the structure can be directly defined from the Patterson map because N is small but in macromolecules such as proteins it is not the same case due to the large number of C, N and O atoms within these macromolecules. In spite of that, the Patterson function is useful for phasing in macromolecules crystallography as they can be used to locate the heavy atoms positions using

MAD and MIR methods. Additionally, it is the basis of the rotation and translation functions in MR technique.

### 2.6.2 Molecular Replacement

This method is based on using a known homologous structure as a phase information source to derive the initial phases for an unknown structure. The MR technique is very popular for determining macromolecular 3D structures and it is useful for binding studies or studies of complexes formed from previously solved proteins. The protein Data Bank (PDB) provides a unique resource for a large number of structural folds and protein structures due to the increasing number of protein models that are deposited in it recently (Berman et al, 2000). The concept is to manoeuvre a structural model for the protein of interest around the unit cell until it lies in an orientation that would account for the observed diffraction. 25% sequence identity is often considered as a minimum level of homology required for molecular replacement. In this method the model must be rotated and translated around the unit cell which involves searching in six-dimensions, and so it is typically divided into two separate steps with three-dimensions for each step: first the rotation and then the translation functions. The crucial step in MR is the selection and preparation of the search model that should be well prepared which means removing non-conserved parts between the search model and the studied protein such as loop and insertions regions. In MR the CHAINSAW program (CCP4, 1994) is used for pruning non-conserved residues back to the gamma atom using the alignment between investigated protein and model sequences in the bla form file (Blast output files). The output file produced in pdb form and labeled according to the investigated protein sequence is used in the MR program PHASER.

#### *2.6.2.1 The Rotation Function*

The rotation function searches are to produce the best orientation which fits the Paterson function of the search model to the structure under investigation. Two types of vectors can be noticed in the Patterson function, inter-molecular

vectors that present within a single molecule and intra-molecular vectors that exist between different molecules in the crystal. However, for the successful use of the rotation function it is necessary that the intra-molecular vectors which are concerned with the orientation of the molecule in the unit cell, dominate the Patterson function. Therefore, the inter-molecular vectors which are concerned with the position of the molecule in the unit cell, should be excluded. This is achieved by comparing the two data maps within a fixed radius of the origin, commonly 75% of the maximum dimension of the search model. The data resolution can affect the success of the rotation function because increasing the resolution can result in emphasing details between the model and the investigated structure, which makes the differences between them more obvious. Based on this fact, the rotation function is usually performed for high resolution limits between 3-6 Å that provide more similarity between the model and the observed structure. It is also important to have good quality low-resolution data to provide information about the overall shape of the molecule. In this stage a number of possible solutions may be found and all are tested in the following translation function to obtain the correct one.

#### *2.6.2.2 The Translation Function*

At the end of the rotation function the model should be correctly orientated rotationally with respect to the investigated structure. The next step is to place the model in the correct position in the unit cell, which can be obtained through the translation function. If the crystal contains only lattice symmetry the origin of the crystal can be set as the origin of the model but if the rotational symmetry is present in the crystal it requires a translation search for the correct position relative to the crystallographic rotation axes. The translation function identifies relationships between the molecules in the crystal using the inter-molecular vectors excluded from the rotation function. Several solutions of rotation function may have been obtained during the first step and therefore it is normal to perform a translation function search on all of them to choose the one which produces the best final result. However, some solutions of rotation and translation cause clashes between molecules in

the cell and cannot be applied. Additionally, different candidate spacegroups can be tested in this stage and selection made for the best correlation between model and observed data with no or very few clashes.

#### *2.6.2.3 Maximum Likelihood Method In PHASER*

Maximum likelihood is a statistical method used to compare the consistency between the search model and the observations. It is used in the PHASER program (McCoy et al, 2007) and REFMAC5 program that were used for phase determination and structure refinement in molecular replacement during this project. The molecular replacement programs use Patterson function and correlation coefficients between the observed and calculated Patterson maps to evaluate rotation and translation functions. The PHASER program uses a maximum likelihood target to evaluate the quality of a particular model rotation or translation and it is more sensitive to the correct orientation. For each trial orientation within the rotation or translation function, PHASER estimates probability functions for the calculated structure factor amplitudes of the model and compares them with $|F_{obs}|$. The orientations with highest probability for producing the observed structure factors are chosen as solutions for the molecular replacement search.

## 2.7 Electron Density Calculation

When the X-ray beam hits a protein crystal it will be scattered and the scatterers are the clouds of electrons around the atoms in the molecules within the crystal. The diffraction pattern that is observed is due to the distribution of the electrons, known as the molecule`s electron density which reflects what the molecule look like. Determination of the phases of the scattered X-rays permits the calculation of an electron density map that can be used to build an atomic model of the protein. The structure factor (F) from reflection *hkl* is related to the electron density ($\rho$) at position (r) by the following equation:

$$F(hkl) = \int_{cell} p(r)\exp(2\pi r \bullet S)\,dv$$

Where **S** is the diffraction vector representing the reflection of index *hkl*. The real space electron density can also be represented as $\rho(xyz)$ and the dot product **r • S** as $(hx + ky + zl)$ and so the equation can be rewritten:

$$F(hkl) = V\int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1} p(xyz)\exp[2\pi i(hx+ky+lz)]\,dx\,dy\,dz$$

This equation allows the diffraction pattern from a known crystal structure to be calculated and so the Fourier transform of **F**(*hkl*) allows the calculation of electron density at coordinates *xyz* by the equation:

$$P(xyz) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l} F(hkl)\exp[-2\pi i(hx+ky+lz)]$$

These values allow ρ(xyz) to be calculated at grid points across the asymmetric unit to create an electron density map that will indicate the atom`s positions in the crystal and allow a model of the protein structure to be built.

## 2.8 Density Modification

Primary phases determined experimentally or *via* molecular replacement technique commonly contain errors. This affects the initial electron density map and leads to insufficient quality and difficulty in its interpretation. Therefore, phase information must be improved through a variety of density modification methods such as solvent flattening, non-crystallographic symmetry (NCS) averaging and histogram matching.

### 2.8.1 Solvent Flattening

Macromolecular crystals contain protein molecules surrounded by solvent molecules that are both ordered and disordered. Solvent molecules occupy a

large portion of the crystal (normally >50%). Hydrogen bonds can be formed between the ordered solvent molecule and protein which can be observed as spheroids in high resolution maps. On the other hand, the disordered solvent molecules have few features and lower average of electron density. Flattening the density in the solvent parts of the map results in reduction of the background noise throughout the map and improves the phases (Wang, 1985). The basis of solvent flattening is simply that a mask is defined which assigns the unit cell into protein parts and solvent parts. The electron density in the solvent area is given a mean value and so is flattened while the protein regions are not changed. Consequently, new structure factors with improved phases can be calculated using the new electron density distribution and then used with the observed structure factors to generate a new improved set of phases and this improves the electron density map for the protein part as well.

### 2.8.2 Non- Crystallographic Symmetry (NCS) Averaging

Non-crystallographic symmetry (NCS) is found in some protein crystals that contain internal symmetry which is not part of the overall crystal symmetry. NCS can be useful in phase improvement as new averaged density values can be calculated. Firstly, the position of the NCS symmetry axis that relates the NCS related regions is found by a self-rotation function analysis. Secondly, these regions are masked off and grid points within these masked regions are then averaged (Kleywegt & Read, 1997).The NCS averaging procedure is best used when 3 or more NCS related molecules are presented in the asymmetric unit.

## 2.9 Structure Refinement

When an initial model for the protein structure has been derived from an electron density map, it is important to improve this interpretation to give the best fit using a computational process known as structure refinement. The procedure of refinement performed by the REFMAC program (Murshudov et al, 1997) used in this project attempts to refine the initial model in order to better account for the observed data. Refinement programs improve the

agreement between | $F_{obs}$| from the collected data and | $F_{calc}$| from the built model by making small changes to the model. It can be done by minimizing |$F_{obs}$| - | $F_{calc}$| through making small changes to the *xyz* coordinates of the atoms in the model and their B-factors.

### 2.9.1 Refinement Parameters

#### *2.9.1.1 The B-factor*

B-factor or "temperature factor" measures the atom oscillation around the position specified in the model. It ranges from 1 Å to 100 Å and depends on the data set resolution. Atoms are in continuous movement and it is known that the side chain atoms in protein molecules have a freedom of movement more than those in the main chain. This movement can affect the data diffraction which makes the attribution of the B-factor for each atom in the structure one of the refinement parameters. In macromolecular structures, the B-factor can be used to judge the type of the ligand or metal ion atoms that could present in the structure by comparing the B-factor value of these atoms with the B-factor values of the surrounding atoms which must be reasonable agreement with each other.

#### *2.9.1.2 The Occupancy*

Another refinement parameter is the occupancy that describes the specific extent of the presence of each atom in the model. In protein structure the atoms have a full occupancy set at 1.0, unless the atom has two or more possible positions such as when the side chain has two conformations with equal frequency. In these cases, the occupancy can normally be set to 0.5 or any other value depending on the B-factor of this atom and the surrounding atoms.

### 2.9.2 Monitoring of Refinement

During the refinement process it is possible for the inherent errors in the phase information to be present into the model and as a result the quality of the refinement should be measured to inhibit the introduction of any bias into the structure. For this purpose the R factor monitor has been introduced as a measurement tool. It is the sum of the differences between observed $|F_{obs}|$ and calculated $|F_{calc}|$ structure factor amplitudes divided by the sum of observed structure factor amplitudes and presented as:

$$R = \frac{\sum_{hkl} || F_{obs} | - k | F_{calc} ||}{\sum_{hkl} | F_{calc} |}$$

The R-factor is a useful indicator of the refinement progress and the model quality. When the differences between the observed structure and the model are low, the R factor is recorded as a low value which reflects the good agreement between them. The introduction of model bias cannot be monitored using R factor measurement alone. Consequently, the free R-flag dataset is used in the form of a second R-factor called the free R-factor ($R_{free}$). This time the calculated structure factors are compared to some of the experimental data never used in the refinement process. These reflections are chosen randomly and often from 5-10% of the total data.

$$R_{free} = \frac{\sum_{hkl \subset T} || F_{obs} | - k | F_{calc} ||}{\sum_{hkl \subset T} | F_{obs} |}$$

Where $hkl \subset T$ represents all reflections form the test set T of unique reflections. Therefore, a fall in R-factor paralleled by a similar fall in the free R-factor during refinement indicates a genuine improvement of the model. If the case is not this and the free R-factor does not record any reduction in its value while the R factor goes down, this may mean that the refinement is over fitting and introducing model bias. The target of each cycle of

refinement is to obtain better phases for the observed data that will be used to generate a new map and re-build the model until the best possible structure is obtained.

## 2.10 Model Validation

Before depositing of a protein structure in a public dataset it is necessary to validate the final model. Model validation depends on experimental proof such as electron density maps in addition to prior chemical knowledge of proteins. It should be done during and after the process of model building.

### 2.10.1 Ramachandran Diagram

The Ramachadran diagram or plot is based on torsion angles of the polypeptide backbone which are not restricted in refinement. It is used to monitor the bond angles to be close to their theoretical values during the refinement process and any change in these values are often considered as an indicator of accuracy. The main chain $\phi$ and $\psi$ angles are plotted for each amino acid along the polypeptide chain. These points are clustered in preferred (allowed) regions and more limited sub-regions for secondary structure elements while a small number of conformational angles may locate in unfavoured (disallowed) regions.

### 2.10.2 PROCHECK and MolProbity

In recent years the development of programs and web services that provide broad-spectrum based evaluation of protein structure quality has increasesd. In this project the PROCHECK (Laskowski et al, 1993) and MolProbity (Chen et al, 2010) suites of programs have been used for model validation. The PROCHECK program gives a graphic output comprising a number of plots and a comprehensive residue-by-residue listing. Additionally, it allows the comparison against prospective values expected for the resolution and also highlights regions that may need further investigation. The deviation of bond lengths and angles as well as $\phi$ and $\psi$ angles and side chain torsions

from expected values can be analyzed by this program which gives a clear view of the model's overall stereochemical quality (Laskowski et al, 1993).

MolProbity is the second program that has been used in order to analyse the geometry of the models in this work. Hydrogen atoms can be added to all residues for an all-atom contact analysis. This can help in avoiding clashes between the residues and can recommend flipping of some amino acids such as asparagine, glutamine and histidine. Analysing Cß deviation values can indicate dubious side-chain torsions. MolProbity writes a script to run in Coot highlighting poorly fitting areas by taking the researcher to the electron density for possible corrections. The MolProbity program can perform some local corrections automatically and other errors can be presented in chart and graphical forms that help users in rebuilding the protein model. MolProbity can also be downloaded for free from http://molprobity.biochem.duke.edu for use as a local server (Chen et al, 2010).