THE UNIVERSITY
of EDINBURGH

Structural Studies of Two Proteins Involved in the Maintenance of Genomic Stability

FEN 1 and DNA-PKcs

THE UNIVERSITY *of* EDINBURGH

James M. Parker B.Sc. (Hons)., M.Sc.

A Thesis submitted for the degree of

Doctor of Philosophy

Institute of Structural and Molecular Biology

The University of Edinburgh

2nd November 2015

Declaration

I hereby declare that this thesis was composed by me, and the research presented is my own, except where otherwise stated. This work has not been submitted for any other degree or professional qualification.

James Parker

Monday 2nd November 2015

Acknowledgements

I would like to thank Dr Laura Spagnolo for affording me the opportunity to perform the research for my PhD studies in her lab and under her supervision. Her expertise, guidance and leadership were invaluable to me throughout my studies. I also would like to warmly thank all past and present members of the group, with special mention to Giuseppe Cannone for being an excellent colleague and friend, whose help really has been immeasurable.

Thanks should also go to Dr Judit Debreczeni for her support and supervision whilst performing the industrial part to this CASE studentship at AstraZeneca. She helped to shine a light on the exciting and sometimes unforgiving world of drug discovery in an industrial setting, only serving to strengthen my resolve to be a part of this ever-changing landscape. I want to thank the Biotechnology and Biological Sciences Research Council and AstraZeneca for funding this research.

I am eternally grateful to my wife Laura, who stood by me through placements, long hours and boring practice seminars, feigning interest in all the little intricacies that I took immense pride in. The cliché of saying I could not have done it without her is I'm afraid to say, very apt in this instance. She also brought my precious children Harriet and Rory, into the world, for whom all of this is for.

Last, but by no means least, I want to thank Malcolm, Dianne and Richard Parker. You taught me how important hard work and learning were, and gave me the tenacity needed to achieve my goals, thank you!

Table of Contents

List of Abbreviations

AI	*Arabinose Inducible*
Amp	*Ampicillin*
AP	*Alkaline Phosphatase*
APS	*Ammonium Persulphate*
ATM	*Ataxia Telangiectasia Mutated*
ATP	*Adenosine Tri-Phosphate*
ATR	*ATM and RAD3 Related*
bp	*Base Pair*
Cam	*Chloramphenicol*
CAPS	*N-Cyclohexyl-3-Aminopropanesulfonic acid*
Conc^{n}	*Concentration*
CSS	*Complex formation Significance Score*
CV	*Column Volumes*
DAC	*Diamond Anvil Cell*
Dbf	*Dumb-bell former*
DDR	*DNA Damage Response*
DNA	*Deoxyribonucleic Acid*
DNA-PK	*DNA – Protein Kinase*
DNA-PKcs	*DNA-PK catalytic sub unit*
DSB	*Double Strand Break*
E.coli	*Escherichia coli*
ECL	*Electrochemiluminescence*
EDTA	*Ethylene Diamine Tetra-acetic Acid*
EM	*Electron Microscopy*
EtBr	*Ethidium Bromide*
FAT Domain	*FRAP, ATM, TRRAP Domain*
FATC Domain	*C Terminus of the FAT domain*
FEN 1	*Flap Endonuclease 1*
FKBP	*FK506 Binding Protein*
FOM	*Figure of Merit*
FRAP	*FKBP12-Rapamycin Associated Protein*

GST	*Glutathione-S-Transferase*
HEPES	*4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid*
HIC	*Hydrophobic Interaction Chromatography*
His	*Histidine*
HPLC	*High Performance Liquid Chromatography*
HPMX	*High Pressure Macromolecular Crystallography*
HR	*Homologous Recombination*
HRP	*Horseradish Peroxidase*
IgG	*Immunoglobulin G*
IPTG	*Isopropyl-β-D-1-thiogalactopyranoside*
IR	*Ionising Radiation*
γH2AX	*Phosphorylated histone H2AX*
Kan	*Kanamycin*
LB	*Luria Bertani*
Lig4	*DNA Ligase IV*
LV	*Loop Volumes*
M	*Molar*
MALDI-TOF	*Matrix Assisted Laser Desorption/Ionisation-Time Of Flight*
MCS	*Multiple Cloning Site*
MES	*2-(n-morpholino) ethanesulfonic acid*
MOPS	*3-(n-morpholino) propanesulfonic acid*
MRN	*Mre11/Rad50/Nbs1*
mRNA	*messenger RNA*
MS	*Mass Spectrometry*
mTOR	*mammalian Target of Rapamycin*
NBT-BCIP	*Nitro-Blue Tetrazolium-5-bromo-4-chloro-3'-indolyphosphate*
NF	*Nuclease Free*
NHEJ	*Non Homologous End Joining*
OD_{600}	*Optical Density at 600nm*
Pab	*Pyrococcus abyssi*
Pfu	*Pyrococcus furiosus*
PBST	*Phosphate Buffered Saline Tween 20*
PCR	*Polymerase Chain Reaction*
PDB	*Protein Data Bank*
PEG	*Polyethylene Glycol*
PEI	*Polyethyleneimine*

PI(3)K	*Phosphatidylinositol-3-OH kinase*
PIPES	*1,4-piperazinediethanesulfonic acid*
PCNA	*Proliferating Cell Nuclear Antigen*
PVDF	*Polyvinylidene Fluoride*
r.m.s.d.	*Root mean squared deviation*
RNA	*Ribonucleic Acid*
ROS	*Reactive Oxygen Species*
RPA	*Replication Protein A*
RSCC	*Real Space Correlation Coefficient*
SDS	*Sodium Dodecyl Sulphate*
SDS-PAGE	*SDS-Polyacrylamide Gel Electrophoresis*
SOC	*Super Optimal broth with Catabolite repression*
SSB	*Single Strand Break*
Sso	*Sulfolobus Solfactaricus*
ssDNA	*single stranded DNA*
SUMO	*Small Ubiquitin-like Modifier*
TAE	*Tris-Acetate EDTA*
TB	*Terrific Broth*
TDA	*Thermal Denaturation Assay*
TEMED	*Tetramethyldiethylamine*
Tet	*Tetracycline*
TRIS	*Tris(hydroxymethyl)aminomethane*
TRRAP	*Transformation/transcription domain Associated Protein*
TSG	*Tumour Suppressor Gene*
UV light	*Ultra-Violet light*
V(D)J	*Variable, Diverse, Joining gene segment*
v/v	*volume by unit volume*
w/v	*weight by unit volume*
XLF	*XRCC4 Like Factor*
XRCC4	*X-ray Cross Complementing gene 4*

List of Figures

List of Tables

Lay Summary

A persons DNA is constantly under threat from damage caused by environmental factors, as well as the normal replication processes it performs whilst generating new cells. If the damage is serious enough it can kill the cell or lead to the formation of cancer, in order to protect the cells from death or cancer the cells have a series of repair pathways it employs in a variety of situations.

DNA-PK is an enzyme that works in the Non-Homologous End Joining Pathway repairing DNA double strand breaks that can be caused by factors such as Ionising Radiation, UV light and chemotherapy drugs. If a high-resolution structure was produced it could be used to help with the development of drugs to be used in combination with existing cancer therapy to make them more effective. This study will show the work performed to attempt to generate soluble correctly folded protein to serve these structural purposes.

DNA undergoes replication in order to produce new cells. This involves a large raft of proteins, all of which have specific functions. One particular protein is involves in the removal of a small RNA primer, whose role is the initiation of the replication process but also has to be removed from the final DNA sequence. Flap Endonuclease 1 (FEN 1) is the protein charged with this task. This study will use the Archeon *Pyrococcus abyssi* as a model organism. It is an organism that thrives under high pressure and high temperature, as well as having a DNA replication system analogous to ours, however much simpler.

Abstract

Genomic stability refers to an organism's ability to maintain and pass forward its genetic information. There are a raft of proteins and pathways whose sole purpose is maintaining this stability through swiftly replicating DNA as well as accurately repairing damage caused through contact with endogenous and exogenous DNA damaging elements. This study will focus on the structural aspects of two proteins that play a part in different areas of genome maintenance.

Flap Endonuclease 1 (FEN 1) works in DNA replication, where it is tasked with removing a small RNA flap that is created during Okazaki fragment formation. This flap removal is essential to mature these fragments into one continuous strand of nascent DNA. Using the archeon *Pyrococcus abyssi (Pab)* as a model system has the advantage of possessing simple replicative machinery, whilst bearing striking similarities with the human system. *Pab* is a hyperthermophilic, piezophile meaning it thrives in conditions of high temperature and pressure.

DNA-dependent protein kinase (DNA-PK) is a holoenzyme that plays a role in the Non Homologous End Joining (NHEJ) pathway by repairing DNA double strand breaks (DSB's). In cancer therapy, a patient is exposed to DNA damaging elements, leading to an ever-increasing population of DSBs. If an inhibitor of DNA-PKcs were introduced along with this therapy it could potentiate its effect, as the cancerous cells will be less able to repair the damage. The aim of this part of the study is to determine a protocol to generate pure, soluble, correctly folded protein for the purposes of biophysical characterisation and X-ray crystallographic structural studies.

Chapter 1

Introduction

1.1 Genome Instability & Cancer

1.1.1 – DNA damage

The genetic information contained within a cell is under the constant threat, both internally and externally, of damage to its integrity. Endogenous damage can come from maintenance processes such as transcription and replication that can occasionally produce faulty products, whereas exogenous damage can come from ultraviolet (UV) light and ionising radiation (IR), as well as certain cytotoxic chemicals. If the damage that is caused is not repaired rapidly and efficiently, this can lead to mutations, compromising the integrity of the genome as well as the survival of the cell.

The levels of DNA damage have been estimated in humans to be in the region of 10,000 to 1,000,000 events *per* day *per* cell (1). DNA damage such as this falls primarily into three categories. Firstly, there can be chemically induced alteration of the nucleotides in the DNA sequence and secondly, there can be mutations that arise due to errors in the replicative process. Finally the chemical alteration of the DNA backbone can occur.

Nucleotide alteration can itself occur as a result of a multitude of factors. These can include oxidation, where the DNA is damaged by reactive oxygen species (ROS) produced

during normal metabolic processes within the cell. One common product of DNA oxidation in mammals is the guanine analogue 8-hydroxyguanine (OH^8Gua)[1] (2).

Another nucleotide alteration is alkylation, whereby the nucleotide is modified, usually through the addition of a methyl group. One prevalent methylation event is the alteration of guanine, forming 6-*O*-methylguanine. This modification stops the guanine binding with its traditional pyrimidine partner cytidine, but with the alternative pyrimidine thymine. This creates a DNA mismatch that can lead to a permanent mutation if not repaired correctly (3).

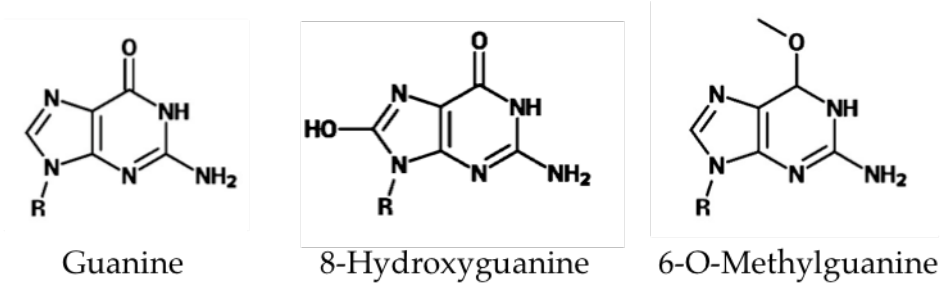


Figure 1.1 - Chemical structure of guanine as well as two products of its modification. Oxidation can lead to addition of a hydroxyl group to position 8 of the guanine, forming 8-hydroxyguanine; Alkylation can cause modification of the carboxyl group at position 6, forming 6-O-methylguanine.

Hydrolysis of the bases can also occur, leading to deamination, depurination and depyrimidation. This then leaves an unpaired base, which can be repaired through base excision repair (BER) (4). Another means by which the nucleotides can be chemically altered is through the formation of either inter- or intra-strand crosslinks. This is the formation of a covalent bond between two nucleotides affecting the ability of this portion of the sequence to take part in transcription and replication (5). Several pathways working in collaboration with one another usually repair this type of cytotoxic damage, BER and homologous recombination (HR) (6).

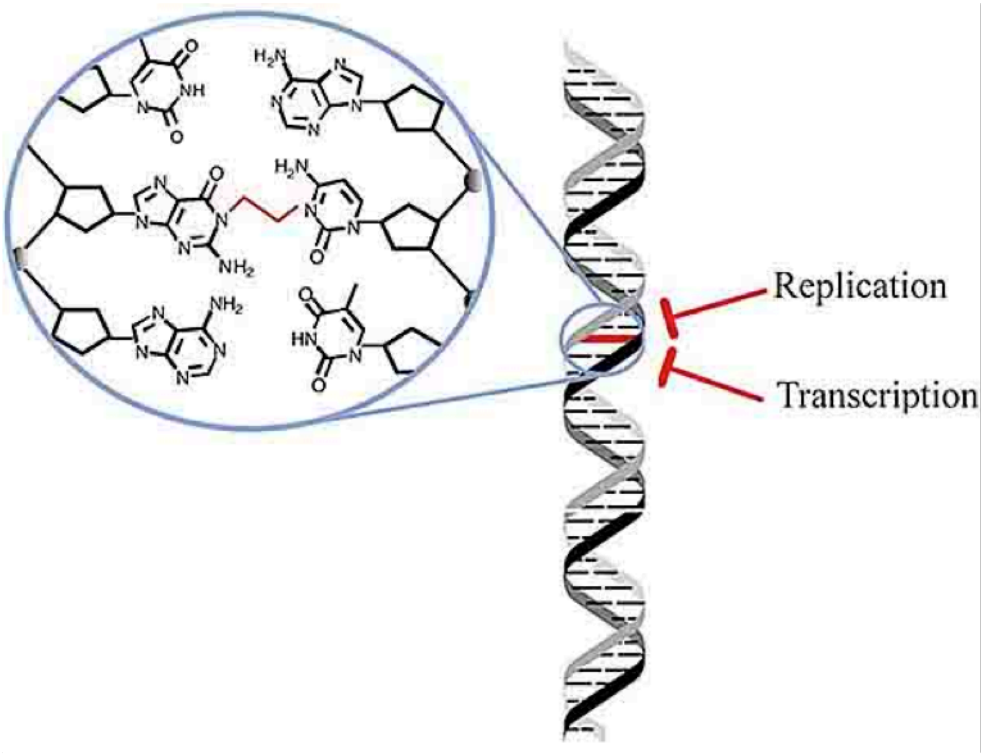


Figure 1.2 – Illustration of an interstrand crosslink between guanine and cytosine. This covalent link between nucleotides abrogates further downstream processing such as replication and transcription (7).

Both polycyclic aromatic hydrocarbons (PAHs) such as benzo[a]pyrene, commonly found in coal tar, and heterocyclic aromatic compounds (HACs) such as dibenzacridine, commonly found in cigarette smoke, are generated in the environment and are attributed to an increased incidence of cancer (8). Their reactive metabolites can covalently bind to nucleotides leading to the formation of depurinic adducts, an example of which can be seen in figure 1.3(B), with similar effects to nucleotide hydrolysis (8).

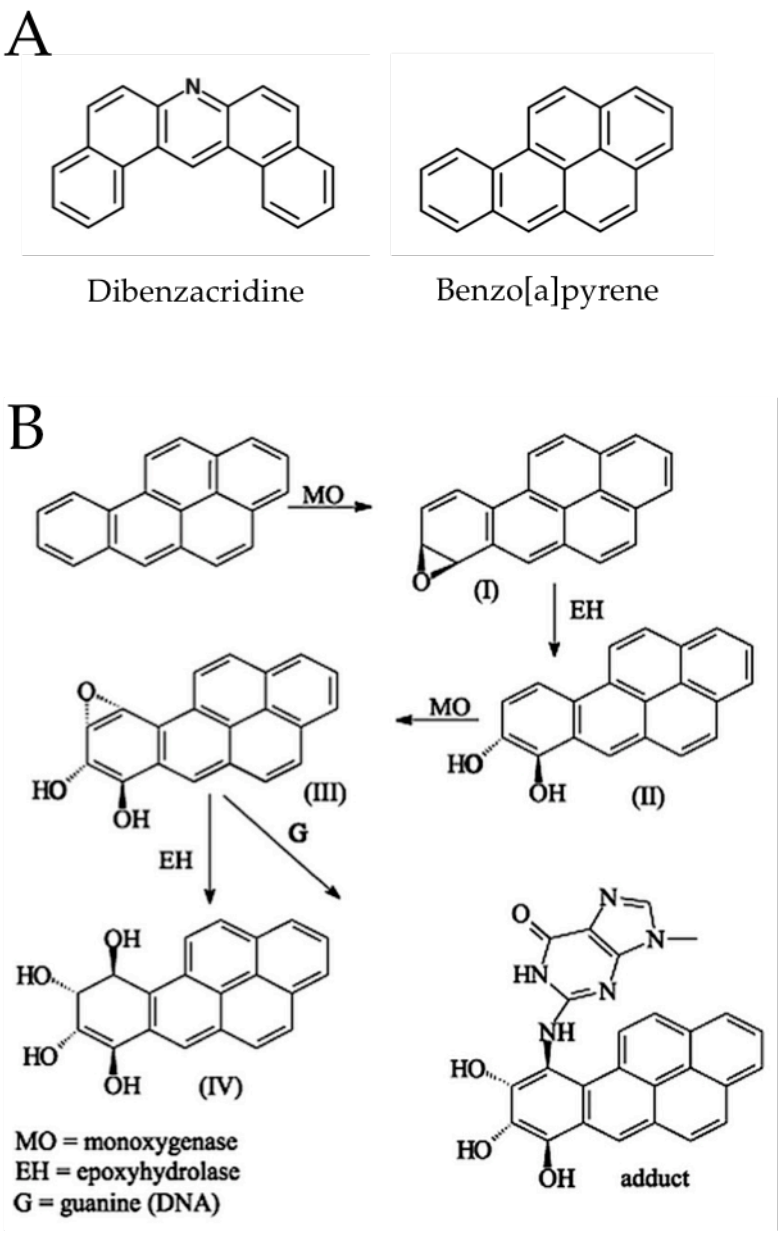


Figure 1.3 - (A) Dibenzacridine, an example of a heterocyclic aromatic compound; Benzo[a]pyrene, an example of a polycyclic aromatic hydrocarbon. (B) Figure published by Formenton-Catai *et al* describing the oxidative metabolic pathway of benzo[a]pyrene that leads to the formation of a DNA adduct with guanine (9).

DNA damage, as a result of replication, can lead to nucleotides being either added, deleted or mutated spontaneously (10). Depending on the nature of the nucleotide alteration, it could be relatively innocuous, but it could be quite damaging to the cell. If a nucleotide

were added early in a protein coding exon sequence of DNA and not subsequently repaired, it would lead to a frame shift, thus permanently changing all amino acids synthesised downstream of the alteration.

Chemical alteration of the DNA backbone is a significant type of DNA damage, and can potentially be the most destructive to the cell. The covalent bonds in the backbone of the DNA can be broken down when they come into contact with exogenous factors such as IR and certain chemotherapeutic drugs (11). This can lead to the formation of single strand breaks (SSBs) or, if the intensity of the radiation or concentration of drug is high enough, double strand breaks (DSBs). DSBs can be extremely cytotoxic and it has been shown that even one unrepaired DSB can be sufficient to trigger cell death (11, 12). If they go unrepaired they can lead to permanent cell cycle arrest and death, either through apoptosis or loss of genomic material (13). If they go on without sufficient repair, gene deletion can occur, along with chromosomal abnormalities, translocation and loss, all of which have been linked with cancer (11, 13) .

DSBs are also an integral part of endogenous cellular processes. A key example of this occurring is V(D)J recombination (variable, diverse, joining gene segments). This is the process of generating antigen-binding diversity within developing B- and T- lymphocytes. These gene segments are flanked by recombination signal sequences, made up of conserved regions that ensure accurate segment joining. Further randomisation is generated through combinations of small deletions as well as insertions of either random or complementary nucleotides. The occurrence of these small deletions is believed to be due to the formation of DSBs, the repair of which occurs by the same pathways used when the cause of the DSB is exogenous (12, 14).

With a vast array of repair processes to both detect and repair damage to the DNA, a considerable majority of the threats to genome integrity are removed or repaired. Some

damage eludes repair and can have a variety of phenotypes. If a cell were to pass on any DNA damage, this damage would be fully incorporated into the nascent strand and therefore would be undetectable with the canonical repair processes and therefore not be able to be repaired, remaining a permanent part of the genome from that point. This can potentially be extremely toxic to the cell and so safeguards are in place to stop the cell cycle from continuing at the point DNA damage has been detected.

1.1.2 – Carcinogenesis

If a mutation is caused either to an oncogene, tumour suppressor gene, or a gene involved in the DNA repair pathway, it can lead to the formation and progression of cancer. An oncogene is dominant in its nature and is formed when a normal gene, termed a 'proto-oncogene', is mutated resulting in the cell becoming cancerous (15, 16). The proto-oncogene is often a gene that forms proteins affecting cell division and differentiation. The mutated proto-oncogene then encodes for increased expression of the resultant series of proteins, which can lead to the cell being unable to initiate or proceed with apoptosis, a trait commonly seen in cancer (16). Tumour suppressor genes (TSGs), such as p53, are genes that code for proteins that function in a variety of ways. Firstly they repress certain cell cycle proteins, leading to inhibition of cell division. Some TSGs code for cell adhesion proteins that work to prevent cellular dispersion, and therefore metastasis (17). A mutation in any of these genes can therefore inhibit their ability to perform their primary function.

The final result of a cellular system that contains depleted numbers of functioning TSG products, and over-expressed oncogene products, is a heterogeneous tumour cell population that has no growth regulation and undergoes replication without restraint, all the while failing to respond to archetypal apoptosis signals (16).

1.2 DNA Damage Repair

1.2.1– Overview of DNA repair pathways

1.2.1.1 - Mismatch repair

As mentioned previously, a nucleotide mismatch can occur in a variety of manners, but the pathway for their repair is thought to be conserved across both prokaryotes and eukaryotes. The mismatch repair (MMR) pathway has been well characterized in *E. coli*, but the mammalian pathway is less understood. Homologues for the key components have been determined, such as MutS, which is responsible for mismatch recognition and binding as well as binding to insertion/deletion loops, which evade repair by proofreading nucleases. These roles are performed in Eukarya by MutSα (Msh2/Msh6 heterodimer) and MutSβ (Msh2/Msh3 heterodimer), where MutSα binds preferentially to the mismatches and single insertion/deletion loops, and MutSβ binds to 2-4 base pair insertion/deletion loops (18). At this point the parent and daughter strands are discriminated from one another and the protein ExoI excises the mismatched portion of the daughter strand. The single strand binding polymerase Polδ and DNA ligase resynthesise and re-ligate the repaired DNA (19).

1.2.1.2 - Base excision repair

As previously mentioned, BER plays a key role in the repair of nucleotide hydrolysis and inter-strand crosslinks, but it is in fact most abundantly used to repair adduct formation as a result of both PAHs and HACs (20, 21). The pathway is initiated when the damage is detected by one of 11 distinct glycosylases, the choice of which relates to the specific lesion type. Once bound the glycosylase excises the affected base, leading to the formation of an abasic site (AP-site), which undergoes incision by an endonuclease APE1 (Human AP Endonuclease). The DNA polymerase, Polβ, then processes the subsequently formed DNA

ends, which are then ligated together with DNA Ligase I. This is termed the single nucleotide-BER pathway. When a longer portion of nucleotides, between 2 and 10, need to be removed, the long-patch BER pathway is employed. This uses the same factors as mentioned previously, with the notable addition of both PCNA and FEN 1. This pathway is preferentially chosen in proliferating cells, where the choice of polymerase also changes from Polβ to Pol δ/ε (22).

1.2.1.3 - Double strand break repair

The safeguards against DNA damage mentioned previously are cell cycle checkpoints and in terms of DSBs they work to halt the transition from one phase to the next. When a DSB is detected, it triggers the DNA damage response (DDR), a process that works to maintain genomic stability by a variety of pathways including DSB repair, cell cycle arrest and apoptotic signal transduction (23).

Figure 1.4 shows the mammalian cell cycle. The point in the cell cycle determines the DSB repair pathway choice for its repair. If the damage is detected either in G_0 or G_1 then the DSB will be predominantly repaired using the non-homologous end joining (NHEJ) pathway. If however the damage is detected in the G_2 phase, homologous recombination (HR) will be employed. This preference for HR is due to having a newly synthesised homologous strand very close to the damaged DNA as replication has just concluded. With no homologous strands nearby in G_0 or G_1 the NHEJ pathway is preferred (23, 24). The proteins highlighted in the figure are known to play roles as checkpoint proteins, halting the cell cycle when aberrant DNA is detected.

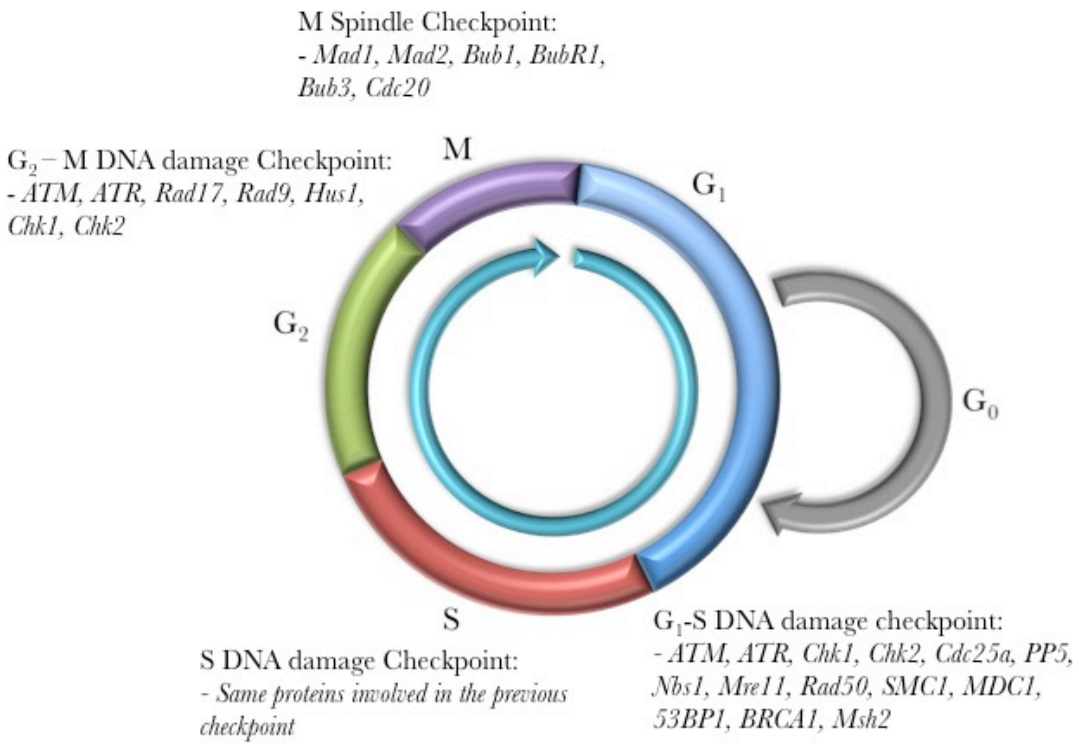


Figure 1.4 - A schematic of the mammalian cell cycle. It begins in the Gap 1 (G_1) phase where the cells physically grow in size and prepare for the impending DNA synthesis. A checkpoint exists here to ensure no DNA damage is present. Gap 0 (G_0) is a resting state of the cell, where it leaves the cell cycle and takes part in no further growth or replication. The synthesis (S) phase then begins and is where DNA replication occurs. A secondary gap phase then occurs (G_2), where the cells continue to grow once more. Again a checkpoint is in place here to ensure all newly synthesised DNA is correct, ready for mitosis. The mitosis (M) phase then begins, here all cell growth stops and focus is changed to division into two daughter cells. There is also a checkpoint midway through mitosis to ensure division is proceeding correctly (25, 26).

1.2.2 – Homologous Recombination

During either S or G_2, the checkpoint kinase Chk1 plays a key role in halting cell cycle progression when a DSB is detected, either as a result of an exogenous factor such as IR, or a DNA replication error such as a stalled replication fork (27). Once hindered in G_2, HR begins with the MRN (Mre11, Rad50, Nbs1) complex detecting the DSB and signalling for Rad52 to resect, in a 5'-3' direction, the DNA strands away from the break site. At this point these overhangs invade an adjacent double helix, with which it shares sequence homology

(12). Rad51 helps with the strand invasion and DNA polymerase then extends the strands. Resolvase A then cleaves the interwoven DNA strands, known as Holliday junctions, into crossover and non-crossover products. Crossover products involve exchange of the homologous strand by cutting the inner strand of one Holliday junction and the outer strands of another. Non-crossover products are formed when both Holliday junctions are cut in the same plane. After the junction has been resolved the backbone is reformed using DNA Ligase (12, 28).

1.2.3 – Non-Homologous End Joining

As mentioned previously, when the cell cycle progression is halted in either G_0 or G_1 then the DSB will be predominantly repaired using the non-homologous end joining (NHEJ) pathway. The halting of the cell cycle whilst in either G_0 or G_1 has been shown to be performed by p53 and this signal is sent by DNA dependent protein kinase (DNA-PK) catalytic sub unit (DNA-PKcs) (29). A schematic of the pathway can be seen in figure 1.5. Whether the source of the DNA damaging element is endogenous or exogenous the first step is detection of the break. This is performed by the heterodimer Ku70/80. It is made up of two sub units, firstly Ku70, which has a MW of 69 kDa in humans and Ku80, which has a MW of 83 kDa in humans (30). The Ku70/80 forms a dyad symmetrical molecule and binds to the DNA in a tight ring formation at a break site (31, 32), in a non-structure specific manner, meaning that it can interact with blunt end DNA as well as DNA with both 5' and 3' overhangs. The sequence is also not important to the binding (33) due to its inherent mode of action.

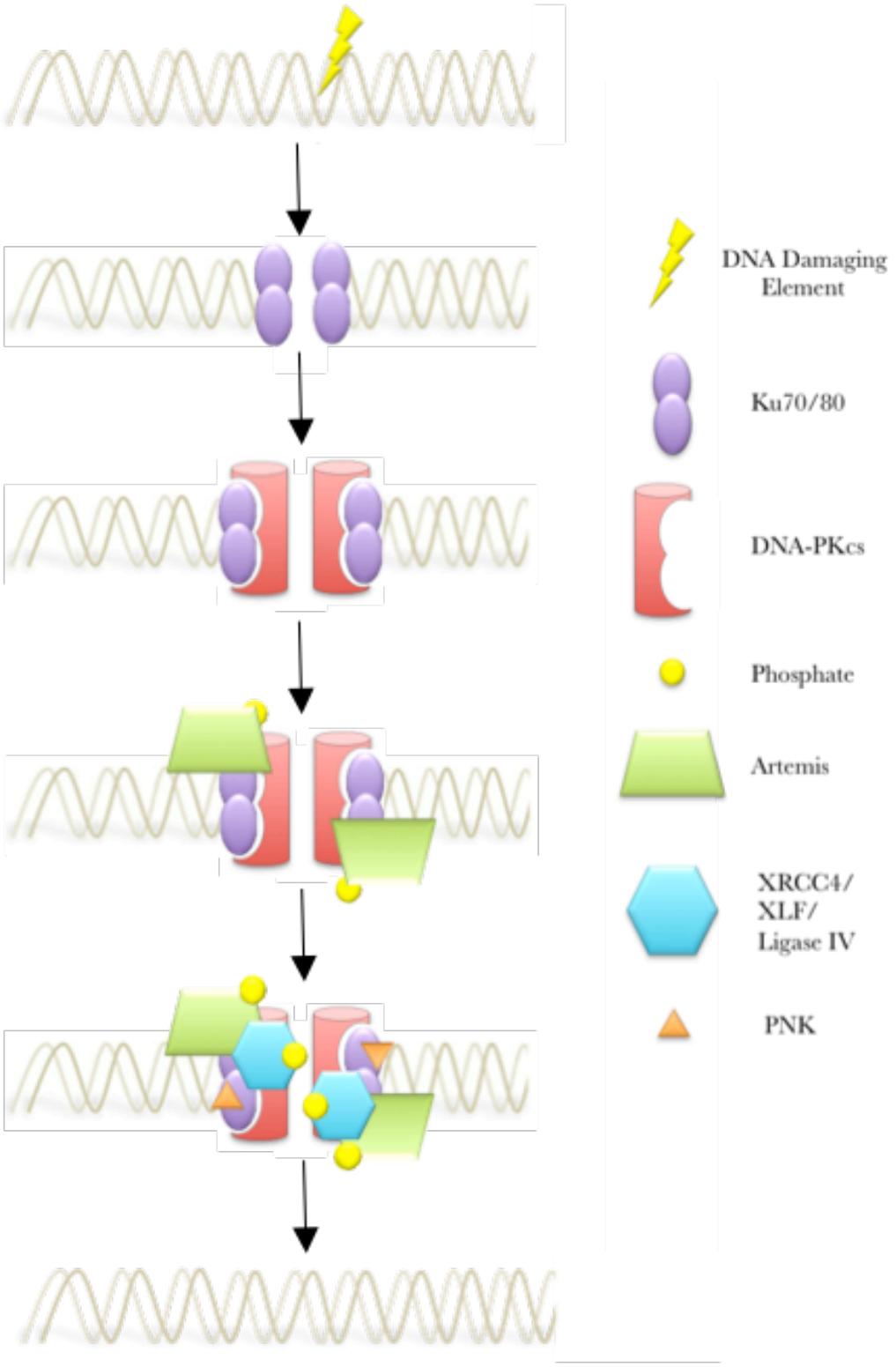


Figure 1.5 - A schematic of the non-homologous end joining pathway. It involves the break recognition by Ku70/80 and the subsequent recruitment of DNA-PKcs, which acts to phosphorylate both itself and downstream factors such as Artemis, XRCC4, XLF and Ligase IV that are essential for the repair of DSBs.

Although the Ku heterodimer is the factor involved in the DSB detection and DNA-PKcs recruitment, DNA-PKcs has been shown to bind to the DNA ends in its absence, when at high concentrations, however the activity of DNA-PKcs is diminished in the absence of the Ku heterodimer (33, 34). It has also been shown that the DNA ends are required in order to elicit an interaction between the Ku70/80 and DNA-PKcs.

The DNA-PKcs is recruited *via* a helical domain present on Ku80 (32), at which point it interacts with both the Ku70/80 as well as the DNA (33, 35). It has been shown that the DNA-PKcs is truly DNA dependent due to the fact that in the absence of DNA, there is no catalytic activity (36). This DNA-PKcs binding to the Ku70/80 forms the complexed, active holoenzyme DNA-PK. The recruitment and binding of DNA-PKcs to the Ku70/80 triggers the translocation of Ku70/80 approximately 10bp (one helical turn) inward from the break allowing the DNA-PKcs to access the break site (37). The next protein that is recruited is Artemis, which has been shown to possess 5'-3' exonuclease activity on its own, but works as an endonuclease when complexed with DNA-PK. It uses this functionality to remove any 5' or 3' overhangs present at the break site, preparing them for ligation (38). This endonuclease activity is dependent on both DNA-PK autophosphorylation and the phosphorylation of Artemis by DNA-PK (39). After Artemis has resected the overhangs present, terminal hydroxyl groups remain on both the 5' and 3' ends, which cannot be ligated to one another. Polynucleotide kinase (PNK) is then recruited to the break site and it also plays a role in preparing the ends of the repaired DNA for ligation by replacing these hydroxyl groups with 5' and 3'-phosphate groups, which requires the presence of both DNA-PKcs and X-ray cross complementing gene 4 (XRCC4), the latter of which also binds at this point (40).

The XRCC4/Ligase IV/XLF complex is involved in ligating the repaired ends of the DNA together. Each of these proteins has been shown independently to interact with other proteins involved in the NHEJ pathway, already bound to the break site. The C-terminus of DNA Ligase IV has been shown to interact with Artemis whilst in complex with DNA-PK

(41, 42). The XRCC4/Ligase IV complex has been shown to interact with Ku70/80 both in the presence and absence of DNA-PKcs, but in its presence the interaction is considerably enhanced (42, 43). It occurs by DNA Ligase IV and Ku70/80 directly interacting with one another whilst the XRCC4 acts to stabilise the entire complex. XRCC4-like factor (XLF) is also present in the complex and is thought to promote NHEJ through regulating the activity of the XRCC4/Ligase IV complex (44).

Once the damage has been repaired, the DNA-PKcs has to break away from the complex. This is achieved through autophosphorylation at the ABCDE cluster, triggering a conformational change in the protein, abrogating any affinity for the DNA (Fig. 1.6) (45, 46).

1.3 DNA-Dependent Protein Kinase

DNA-PKcs is a serine/threonine protein kinase that is a member of the PhosphatidylInositol-3-OH Kinase (PI3K)-related Kinase family (PIKK) and it consists of a single polypeptide chain of 4128 amino acids, providing the main functionality to the DNA-PK complex (47). Other members of this family include ATM (Ataxia Telangiectasia Mutated protein), ATR (Ataxia Telangiectasia and Rad3 related protein) and mTOR (mammalian target of Rapamycin). ATM and ATR both play significant roles in signal transduction through the cell cycle checkpoints mentioned earlier (48) and mTOR also plays a role in cellular signalling, regulating cell growth and proliferation (49).

It has been shown that cells containing mutations in DNA-PKcs have increased radiosensitivity, indicating its importance in the repair of DSBs as a result of IR. Its importance has also been shown with scid (severe combined immunodeficiency) mice, whereby DNA-PKcs is mutated, leading to it being unable to carry out the V(D)J recombination required to generate antibodies during development (11, 50).

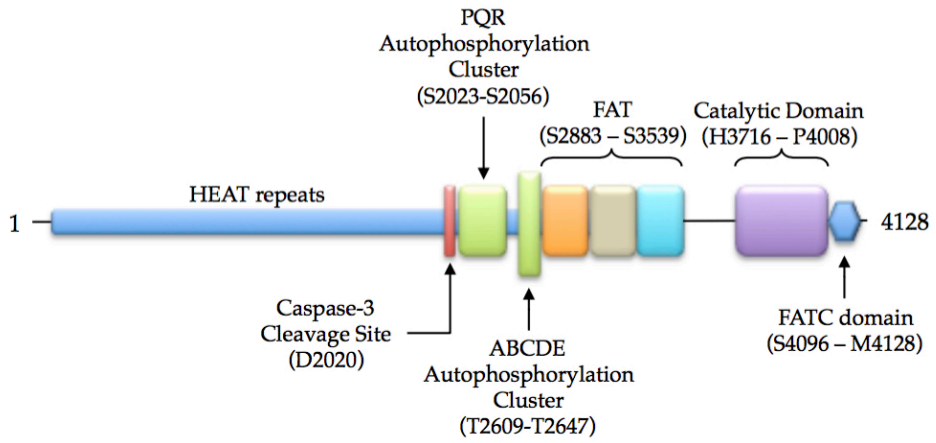


Figure 1.6 - Linear sequence and conserved domains of DNA-PKcs: Red - Caspase-3 cleavage site; Green - Two functionally relevant autophosphorylation clusters, PQR (S2023-S2056) and ABCDE (T2609-T2647) (51, 52); Orange – FKBP-12 Rapamycin associate protein domain; Beige – Ataxia Telangiectasia mutated protein domain; Cyan – Transformation/Transcription domain-associated protein domain – All three of which form the FAT domain; Purple – Catalytic domain that is a member of the PI3K family.

Highlighted in figure 1.6 are two functionally relevant autophosphorylation clusters (PQR and ABCDE) within the sequence. Present at D2020 is the Caspase 3 cleavage site. DNA-PKcs is a substrate for Caspase 3 prior to apoptosis (53). The FAT (FRAP, ATM, TRRAP) region of residues spans residues S2883 to S3539. The FRAP (FKBP12-Rapamycin Associated Protein) also known as mTOR (mammalian Target Of Rapamycin) plays a key role in cell cycle progression (54). ATM (ataxia telangiectasia mutated protein) plays a role in the phosphorylation of downstream factors such as the histone H2AX, which is triggered in response to the detection of DNA DSB's (55). TRRAP (TRansformation/tRanscription domain-Associated Protein) is a 'pseudokinase', lacking the conserved residues required for successful ATP binding and turnover. It works by recruiting several histone acetyltransferase complexes to chromatin, to be used during transcription, replication and repair of the DNA DSB (56).

The kinase (catalytic) domain codes for the region that binds ATP in order to confer its enzymatic activity and phosphorylate the Ku70/80. It is also one of the regions where the DNA-PKcs can autophosphorylate and dissociate from the newly repaired DNA double strand. The catalytic domain is directly connected to the C-terminal end of the sequence (FATC). It has been shown to be essential to the function of the enzyme, as a mutation present in the T-loop, a phosphorylation site within the kinase domain at T3950, leads to total loss of function (57).

Initial in depth structural analysis using a combination of negative stain electron microscopy (EM) and cryo-EM led to the determination of a 13 Å structure from Rivera-Calzada *et al* (58) They determined binding modes and orientations between DNA-PK and double stranded DNA. They corroborate the finding that the DNA-PKcs interacts with the DNA ends triggering a significant increase in catalytic activity. The evidence for which was a conformational change, brought about by the DNA binding. The mechanism of this DNA binding involved bringing the palm and head regions of the protein within close proximity of one another. The palm region of the EM structure was determined to map to the N-terminal HEAT repeat region of the protein, with the head region containing the C-terminal catalytic domain (58).

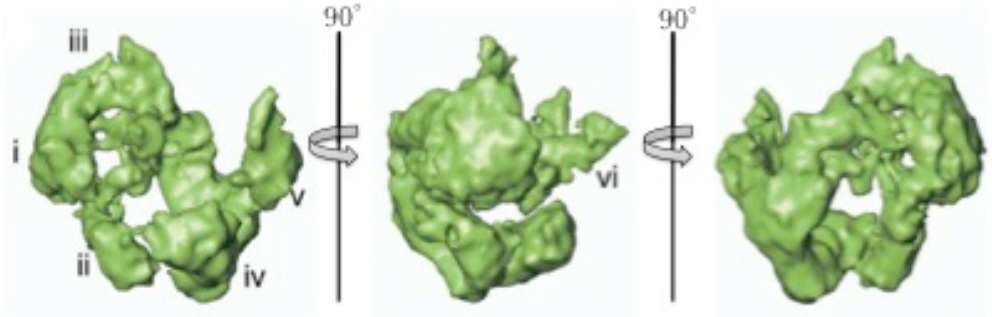


Figure 1.7 – 13Å three dimensional cryo-EM structure of DNA-PK published by Rivera-Calzada et al (58) showing 90° rotated views of the reconstructed volume. (i) – Nucleus within the head region (ii) – large globular region, (iii) – tubular structure, (iv) – distal claw (v) - proximal claw (vi) – curved stalk.

The EM studies also found that the DNA bound within the central cavity, with one end of the DNA binding within the inner portion of the HEAT repeats, between the head and palm of the protein. This work also correlates with the putative DNA binding domain present in the 6.6 Å structure highlighted in figure 1.7 (47, 58).

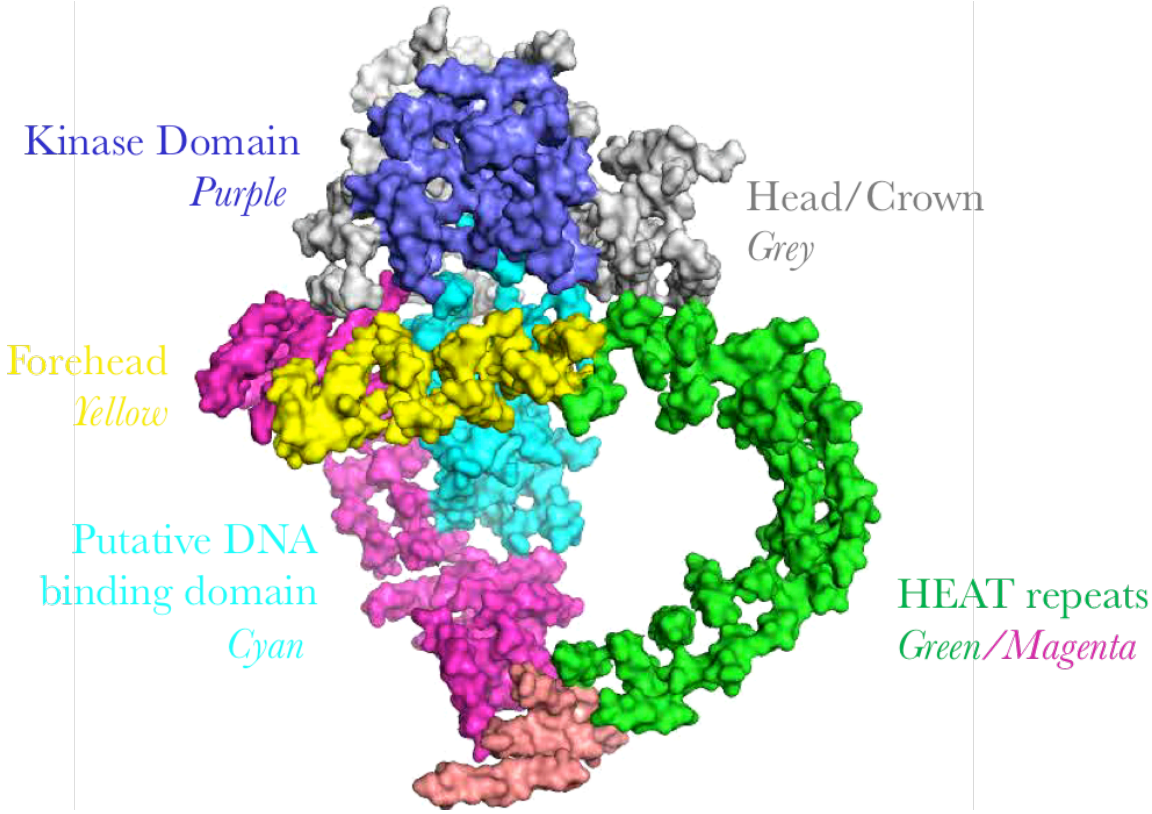


Figure 1.8 - Annotated crystal structure from Sibanda *et al* (47) showing the poly-alanine chain structure solved to 6.6 Å for DNA-PKcs. PDB ID: 3KGV. Figure prepared using Pymol (59).

The structure seen in figure 1.8 is a 6.6 Å representation of DNA-PK, isolated and purified to near homogeneity from HeLa S3 cells. At this resolution, side chain information cannot be determined. This meant that the structure was made up of alanine helices placed within the visible rods of electron density. It does show however that it is approximately 160 Å in length, with a central ring diameter of approximately 120 Å (47). The position of the kinase domain within the structure was determined by superimposing the kinase structure of PI3Kγ onto the head/crown region (47). The sequence of the PI3K/PI4K region from

DNA-PKcs was aligned to the same region from PI3Kγ using the online software Clustal Omega, and the result indicated that the two regions share 32% sequence identity (60). Highlighted in cyan is the putative DNA binding domain. The green and magenta regions of the structure represent the HEAT repeats that are arranged anticlockwise from the pink gap region shown in the structure and circumnavigates the entire structure before reversing direction on the other side. This gap is theorised to be the point at which the two portions of repeats could move apart allowing for the release of DNA-PKcs from the DNA (61), this being the structural change mentioned previously that is brought about by autophosphorylation.

As previously mentioned, the catalytic activity of DNA-PKcs is essential to confer its activity but its exact role is not yet fully understood (62). The DNA-PKcs has been shown to phosphorylate each of the factors involved in NHEJ, but it has also been shown that none of these phosphorylation events are essential for the pathway to proceed. The only phosphorylation events that have been shown to be necessary are, in fact, the autophosphorylations (36, 37).

Autophosphorylation occurs at many of the sites throughout the sequence, which will be discussed in chapter 3, however there are some sites that are very important to the progress of NHEJ. One of the first sites to be phosphorylated is T2609, which is found within the ABCDE cluster and this site promotes end processing during NHEJ (36, 37). It is also the site that, once phosphorylated, triggers the conformational change required to dissociate from both the Ku70/80 and the repaired DNA (36). There are two phosphorylation sites that serve to inhibit repair *via* NHEJ. The T3950 loop as well as T946, present within the JK cluster that, when phosphorylated, both inhibit NHEJ and promote HR, however the former does this through inactivation of the DNA-PKcs, whereas the latter does not inactivate the enzyme (37, 63).

The EM studies mentioned previously also indicate an importance of these autophosphorylation sites throughout the structure in the various clusters (32, 58). The researchers determined that, based on the placement of these sites within the structure, they cannot be accessed *via* the active site of their respective molecule. This then lead to the conclusion that the sites have to interact with an alternative DNA-PKcs molecule in a trans orientation. This is thought to be the interaction that brings the two broken ends of DNA into close proximity with one another in order for further processing of the DSB by downstream NHEJ factors (32, 58).

DNA-PK has also been shown to play a role in the protection of the ends of telomeres. After DNA replication it is found to bind to the telomere ends forming a protective terminal structure in order to prevent end-to-end fusions. The activity of DNA-PK on the telomere has been shown to only act on the product of leading strand synthesis rather than lagging strand synthesis when single chromatids are involved (64, 65). The activity of DNA-PK in this pathway has been shown to be solely possible due to the kinase activity of the enzyme, due to the fact that DNA-PKcs knockouts were not as compromising to chromosome end protection as a kinase inhibitor of DNA-PK (64, 66). These end-to-end fusion products have been shown to contribute significantly to chromosomal instability.

1.3.1 – A link between DNA-PK and cancer.

At the point a cell becomes cancerous its growth is rapid and unregulated. Although traditional cell cycle checkpoints can be faulty, the cancerous cells do employ DNA repair pathways such as NHEJ to repair DSB's. Due to the high proliferative rate of cancer cells, any introduction of DNA lesions through various cancer therapies has a higher probability of being repaired. In cancer therapy, a patient is exposed to either IR for radio-therapeutic purposes or to cytotoxic drugs for chemotherapeutic purposes. Both of these elements are

cytotoxic, promoting the formation of lesions, which will eventually lead to the formation of an ever-increasing population of DSB's.

Experiments have been conducted on DNA-PKcs, where mutations within the kinase domain were introduced, subsequently leading to the failure of the pathway to restore the DSB back to a fully functional DNA double helix [9]. If a small molecule inhibitor were formulated to either occupy or block the catalytic domain, then it would stop the DNA-PKcs performing its functions of both phosphorylation and recruitment, effectively stopping the DNA-PKcs being a functioning part of the NHEJ pathway. Due to the fact that it is such an integral part of the pathway this inhibition would also halt the full repair pathway, significantly hindering the cell being able to repair the DNA DSB's thus potentiating the effect of the therapy.

1.3.2 –DNA-PK Inhibition

DNA-PKcs is an interesting target to study in terms of potential use in combination with existing chemo- and radio-therapies. By inhibiting DNA-PK, its role within NHEJ, as well as V(D)J recombination, is inhibited. The former of which makes the cells prone to synthetic lethality. Whereby if the product of gene A is knocked out, which in this example can be DNA-PK, normal cells can survive due to the presence of various other factors and repair pathways that allow the cell cycle to continue. In cancer cells however, mutations in the DSB repair mechanisms already exist causing genomic instability. If the same product in these cells is then knocked out, the lack of alternative repair mechanisms mean the cell is considerably more likely to die.

1.3.2.1 – Small molecule catalytic inhibitors

An ATP-analogue inhibitor will competitively inhibit ATP binding, thus restricting the downstream phosphorylation required in the NHEJ pathway. This particular mode of action

however can lead to significant selectivity issues as all of the approximately 500 various kinases coded for within the human genome require a similar binding mode of ATP. To date, several small molecule inhibitors of DNA-PKcs have been formulated through combinatorial chemistry with widely varying results (67, 68). Wortmannin is an example of a non-specific DNA-PK inhibitor (68), but this trend can be further expanded upon with results published by Bain *et al* (69) indicating that over two previous studies, totalling 38 different supposedly specific kinase inhibitors, a high number of these affected too many alternative kinase targets to be accurately called specific. Hollick *et al* (68) however claim to have found, through their use of structure activity relationships (SARs) a compound, NU7442, that is highly specific to DNA-PK with a half maximal inhibitory concentration (IC_{50}) of 14nm, based on the inhibition of serine phosphorylation.

This use of SARs and homology modelling in drug design has been utilised due to a lack of any accurate structural information. This is due to the fact that the highest resolution published structure, as previously mentioned is 6.6 Å (47), which is far too low, in terms of resolution, to determine any interactions a potential drug candidate would have with side chains present within the kinase domain. With more detailed structural information, at a higher resolution, rational structure-based drug design could be implemented. This would then lead to a more effective use of time and resources in the production of small molecule inhibitors.

1.3.2.2 – Inhibitory peptides

One negative aspect of catalytic inhibition of DNA-PK is the inadvertent inhibition of its role within alternative pathways, such as the V(D)J recombination pathway, reducing diversity within the cellular immune system. By working on an alternative inhibition target this can theoretically be negated. As mentioned previously, the autophosphorylation at both the ABCDE and PQR clusters of DNA-PKcs is integral to its function within NHEJ. Sun *et*

al have generated a series of small peptides that are theorised to bind to the ABCDE cluster present in DNA-PKcs. By binding at this point all subsequent phosphorylation events are inhibited, with T2647 being postulated as one of the most important phosphorylation sites. They do also theorise that the inhibitory peptides could be competing with DNA-PKcs substrates blocking their access to the kinase. (70).

1.4 Archaea

Life on Earth can be categorised by where on the phylogenetic tree it belongs. The tree is constructed based on the three domains of life, the Eukarya, the Prokarya and the Archaea (71) and this classification is determined based on ribosomal RNA (rRNA) genes and how they relate to one another (72, 73). The kingdom of archaea can then be divided into the two main recognised phyla: Crenarchaeota, which can be described as thermophilic, some existing at temperatures exceeding 110 °C, with a tendency to act as sulphur reductants (74); and Euryarchaeota, which is a considerably broader phylum, containing significantly more morphologies (71, 74). Archaea are a group of prokarya that bear strong resemblance to bacteria in their physical size and shape, but at the molecular and genetic levels they differ. Firstly they can be characterised by an ability to thrive under extreme conditions. These extreme conditions can include low or high pH (acidophiles and alkalophiles respectively), temperature (hyperthermophiles), pressure (piezophiles), salinity (halophiles), and often, this is in combination with anaerobic conditions.

Secondly, at the genetic level, archaea actually have many features in common with eukaryotes. In the context of DNA, Eukarya posses a nucleus that houses the cell's genetic information, whereas prokarya do not encapsulate their DNA, instead it is present throughout the cellular matrix. The prokaryotic genomes are considerably smaller than the genomes of Eukarya. As a result the DNA is small enough to be contained within a single

circular chromosome. This is in contrast to the eukaryotic DNA that is linear and too large to fit within the nucleus without being packaged, firstly around a nucleosome and then coiled around itself in a series of loops forming `chromosomes. That being said, the majority of Archaeal proteins involved in replication, transcription and translation have homology to their eukaryotic counterparts, whereas their relationship to the bacterial proteins are far more distant (75).

1.4.1 – Pyrococcus Abyssi

The research presented in this thesis was performed using FEN 1 from the euryarcheon *Pyrococcus abyssi (Pab). Pab* is a hyperthermophilic piezophile, meaning that it thrives under high pressure and high temperature. Originally discovered 2 km below sea level in the North Fiji basin, in the south-west Pacific ocean, it grows in temperatures ranging from 67°C to 102°C and at pressures ranging from atmospheric pressure to an optimum pressure of 20 MPa (0.2 kBar) and an optimum temperature of 96°C (76). Studying archaea such as *Pab* can help to ascertain what genotypic and phenotypic aspects of the organism are required to enable it to exist in conditions that would kill other organisms. In the context of this research, comparing both ambient pressure and high-pressure X-ray crystal structures, what structural changes can occur within the specific DNA replication protein, Flap Endonuclease 1 (FEN 1), that allow the organism to not only survive, but thrive.

1.5 DNA Replication

DNA replication is the process whereby the genetic information in a proliferating cell is copied from a template strand to form two newly synthesised daughter strands, identical to the original sequence. It is a vitally important step in the life of a cell and faithful and accurate replication is necessary to ensure the correct genetic information is passed to

daughter cells. Errors can occur during this process, and the repair pathways mentioned in chapter 1.2 are present to ensure that mistakes made during replication are not made permanent for future generations.

DNA replication occurs when a series of proteins come together to form what is termed the replisome, a complex molecular machinery that actually performs the replication. The replication itself can be broken down into three processes. Firstly comes the initiation step. This begins at the origin of replication (*oriC*), with the formation of a pre-replication complex, which plays a role in the recruitment of further downstream proteins. The next stage is elongation, where the new daughter strands are synthesised, using the initial DNA as a template. This synthesis only ever occurs in a 5'-3' direction, which means that the leading strand (with 5'-3' polarity) is synthesised in a continuous manner, but the lagging strand (with 3'-5' polarity) has to be synthesised in a discontinuous manner, maintaining the correct directionality. This discontinuous synthesis makes use of short fragments of RNA-primed DNA, known as Okazaki fragments. The RNA primers present on these fragments are then removed and the two adjacent fragments are ligated to one another in a process known as Okazaki fragment maturation. The final step in replication is termination, occurring when the replisome encounters a termination site within the sequence, stopping the replication fork from proceeding any further.

1.5.1 – Replication in Prokaryotes

Extensive research has been carried out to determine how DNA replication occurs in prokaryotes using *E. coli* as a model organism. There is one origin of replication termed the *oriC*, which is a unique 245 bp sequence in the genome that directs assembly of the replisome onto the DNA to generate the replication fork (77). A DNA unwinding element (DUE) is also present at the *oriC*. This is a binding site for DnaA, whereby a DnaA-ATP complex multimerises and then melts the DNA allowing the helicase DnaB to access the DNA *via* the

helicase loader DnaC (78). At this stage the primase DnaG synthesises the RNA primers for both the leading strand, and for the Okazaki fragments needed for lagging strand synthesis. The polymerase holoenzyme Pol III HE then assembles the new DNA strand (79).

1.5.2 – Replication in Eukaryotes

DNA replication in eukaryotes is considerably more complicated and diverse than bacterial replication. There are a variety of factors that make this the case; firstly there is a larger network of associated proteins and complexes required. Eukaryotic cells take part in the cell cycle shown in figure 1.4, whereas bacteria use a far simpler mechanism of binary fission. In binary fission the DNA is first replicated and the two strands migrate to opposing sides of the cell, in the absence of spindles. Once migrated, the cell lengthens and the plasma membrane constricts in the central portion of the cell forming a new, genetically identical cell. The genetic information of a eukaryotic cell is replicated once *per* cell cycle. This replication occurs during the S phase, but there are several steps that are essential to the replication that occur during the G_1 resting phase.

The first step that occurs early in the G_1 phase, analogous to prokaryotic replication, is the formation of the pre-replication complex (pre-RC). This involves a complex of factors including the heterohexamer origin recognition complex (ORC) as well as the proteins: cell division cycle 6 (CDC6); Chromatin licensing and DNA replication factor 1 (CTD1) and the heterohexamer mini-chromosome maintenance (MCM). They bind to *oriC* sites present in the genome. In bacteria there is one *oriC* however in eukaryotes there are as many as several thousand on each chromosome (80). This is simply due to the size of the eukaryotic genome. An example is the human genome, which is approximately 3.2 billion bp, compared with the genome sequence for *E.coli*, which is approximately 4.6 million bp.

The transition from G1 phase to S phase begins directly after the formation of the pre-RC, at the point at which a pair of kinases, cyclin-dependent kinase (CDK); and Dbf (Dumb-bell former) dependent kinase (DDK), activates it. This activation results in MCM mediated DNA unwinding and binding of both replication protein A (RPA) and DNA polymerase, which work to initiate DNA synthesis (80, 81).

1.5.3 – Components of the archaeal replicative machinery

The archaeal replicative process shares aspects with both the eukaryotic and bacterial systems. Some archaea have one origin of replication, like bacteria, as is the case with *Pab*, which was also the first origin to be determined (82). Other archaea can have more, like Eukarya, such as *Sulfolobus solfataricus*, which has two (83), or *Methanocadococcus jannaschii,,* which has multiple origins (84).

The origins are large AT-rich stretches of DNA that are essential for function. They are frequently found in close proximity to the genes encoding for the archaeal homologues of both CDC6 and Orc1, mentioned previously as initiator proteins required for formation of the ORC, responsible for the formation of the pre-RC (84, 85).

In the context of *Pab*, the ORC bind to the single origin using the catalytic activity derived from its AAA+ ATPase domain. As mentioned previously, eukaryotes bind the MCM at the pre-RC, mediated by the CDC6 within the ORC. This pathway has not been resolved in archaea, but Shin *et al* postulate that based on their sequence similarity, the archaeal homologue CDC6-2 does in fact mediate the loading of MCM on to the ORC (86).

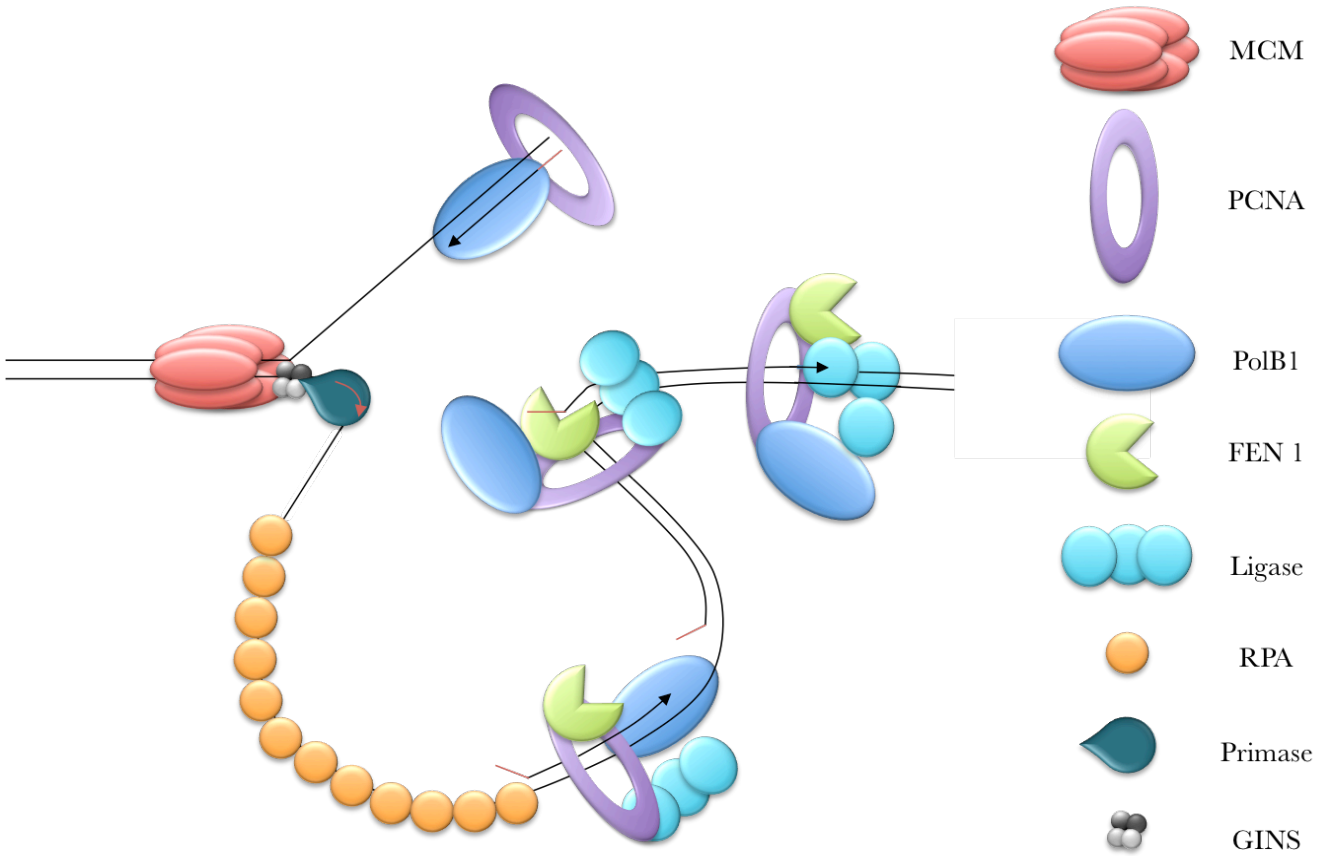


Figure 1.9 – A model of the architecture of the archaeal DNA replicative machinery. The MCM helicase unwinds the DNA allowing for the leading strand to be synthesised continuously whilst the lagging strand is synthesised discontinuously. Here the formation of the okazakisome is initiated, where PCNA mediates interactions between the DNA and the polymerase, lengthening the DNA, the FEN 1, cleaving the displaced RNA flap, and the ligase, sealing the nick between adjacent Okazaki fragments.

The GINS (Go-Ichi-Ni-San – five-one-two-three in Japanese) complex is directly involved in the interaction between the MCM and the primase. Unlike in eukaryotes, where the primase DnaG directly interacts with the helicase DnaB, in archaea there has been shown to be no interaction between the primase and the helicase MCM (87). Due to the fact that polymerases, such as PolB1, are unable to generate new DNA *de novo*, they require the use of a primed sequence. Both the small and large primase sub units are responsible for generating these short sequences of RNA that initiate the polymerases on both the leading and lagging strands.

The Okazaki fragments then undergo maturation, a process that involves removal of this RNA flap. FEN 1, mediated by its interaction with the proliferating cell nuclear antigen (PCNA), performs this flap removal. Another protein that interacts with PCNA is the ligase, that seals the nick between adjacent Okazaki fragments leading to a continuous strand of nascent DNA (88). RPA binds to the single stranded portions of the DNA, as shown in the figure in order to protect it from nuclease attack or chemical modification during the replicative process. It is a heterotrimer, made up of three monomers that are approximately 41 kDa, 32 kDa and 14 kDa (89).

	Replication Components				Replication Mechanisms
Domain	Pre-RC	Initiation	Elongation	Termination	
Eukarya	ORC, CDC6, CTD1, MCM	CDK, DDK, CDC45, GINS, MCM,	GINS, PCNA, RPA, pol ε, pol α, pol δ	FEN 1, Dna2, DNA Ligase	Occurs inside nucleus, many origins of replication, very slow replication.
Bacteria	-	DnaA, DnaB, DnaC,	DnaG, Pol III HE	Topo IV	Occurs in cytoplasm, single origin of replication, very rapid replication.
Archaea	ORC, CDC6-2, Cdt1, MCM		Primase, GINS, PCNA, RPA, PolB1	FEN 1, Dna2, DNA Ligase	Occurs in cytoplasm, various levels of replication origins, very rapid replication.

Table 1-2 - Highlighting the similarities and differences in the process of DNA replication across the three domains of life.

1.5.4 – Flap Endonuclease 1

FEN 1 is a highly conserved structure specific endonuclease that works by removing any un-annealed 5' flaps present as a result of lagging strand DNA synthesis. The maturation of Okazaki fragments and the role of FEN 1 is functionally conserved across eukaryotes, prokaryotes and archaea (90). The Okazaki fragments observed in both *Pab* and other Eukarya are the same length (82) and taking this into consideration makes *Pab* an excellent model to study DNA replication, since the replication machinery is simpler but shares many

similarities with both Eukarya and bacteria. FEN 1 has been shown to be essential to survival of the cell. Homozygous knockouts have been shown to be lethal in mice and heterozygous knockouts have been shown to lead to rapidly accelerated tumour growth due to incorrect DNA replication (91). Traditional endonucleases are sequence specific, but FEN 1 requires a branched DNA structure, with a single 3' unpaired nucleotide, overlapped with a sequence of 5' unpaired DNA that can vary in length, but if it is longer than 5 to 7 nucleotides, RPA will bind and this portion of the nucleic acid will have to be cleaved *via* Dna2 before FEN 1 can perform its function (92). This particular pathway, involving the use of RPA as a single stranded binding protein and Dna2 mediated cleavage, is termed the 'long flap' pathway (93). PolB1 displaces the RNA primer used to initiate the synthesis of the upstream Okazaki fragment, and this displacement leads to the formation of the 'double-flap' required as a structure for optimal FEN 1 binding (94). FEN-1 then works by cleaving this substrate after the first base pair preceding the 5′flap to remove the flap and create a nicked DNA product (94, 95).

Occasionally, in eukaryotes, FEN 1 can dissociate from the PCNA-FEN 1-DNA ligase-DNA complex, which in turn leads to polδ mediated flap lengthening (93, 96). If this newly lengthened strand is complementary with itself, it can in fact form a hairpin structure that cannot be repaired either by the long flap pathway or the short flap pathway (a pathway that does not require intervention from either RPA or Dna2). These hairpins have been shown *in vitro* to actually be un-wound by the helicase Pif 1 allowing the polymerase to resynthesise a new Okazaki fragment by displacing anything present on the affected portion of DNA. It is currently unknown if this also occurs *in vivo* (93, 97).

As mentioned previously, in section 1.2.1.2, BER is used to repair oxidised and alkylated nucleotides present in genomic DNA. BER can be subdivided into single nucleotide BER (SN-BER) and long patch BER (LP-BER), the difference being the size of the portion of DNA undergoing repair (98). In LP-BER, the DNA polymerase adds anything from 2 to 15

nucleotides, displacing the damaged DNA strand, in the same manner as it displaces the RNA primer during Okazaki fragment maturation. This displacement leads to the generation of a 5' single stranded flap, that is a substrate for FEN 1 (98). The FEN 1 then cleaves this flap in the same manner as in Okazaki fragment maturation leading to DNA ligase sealing the nick and producing a repaired DNA product.

1.5.5 –Effects of a high-pressure system

At the high pressures that *Pab* thrives at, the pressure is initially exerted on the cell wall and membrane, which can potentially shield the internal cellular components from these extremes. When pressure is applied to a system, there has to be a change in the volume of that system to maintain equilibrium. The Gibbs free energy (G) can be defined by the following equation:

$$G = H - TS = E + pV - TS$$

Where H is enthalpy of the system, T is temperature (in kelvin), S represents the entropy of the system, E is the internal energy, p is the pressure and V is the volume.

Le Chatelier's principle states that in a given system, upon the introduction of an external stimulus, a change in another aspect of the system has to occur to oppose the external stimulus and maintain equilibrium. The equation above states that in order for the equilibrium to be maintained, a shift toward an increase in pressure will ultimately lead to a decrease in overall volume (99).

When pressure is exerted onto a protein, there are several aspects of the protein that can be altered including the density and the elasticity (100). Under pressure elasticity, the ability of a system to undergo a gradual and reversible change in its shape or density, the protein volume will decrease to levels lower than at ambient pressure, which theoretically

could hinder the protein from functioning. However under the same pressure solute motility increases relative to ambient pressure, which in theory could enhance the enzymatic ability of the protein (100). Production and abundance of internal cavities could also be intrinsic in a proteins ability to perform at high pressure. Functional assays would need to be performed at high-pressure to determine if the protein was in fact being subjected to the pressure, or whether the cell wall and membrane were shielding internal components from this pressure. In the context of FEN 1 in *Pab*, One such assay could involve the use of either a radio- or fluorescently- labelled DNA strand that, upon successful cleavage from the engineered 'flap', could be detected. This assay could be performed both *in vivo* and *in vitro* at both ambient pressure and at optimum pressure for *Pab,* to determine where the pressure is acting in this particular system.

1.6 Scope of this Thesis

1.6.1 – DNA-PKcs

The principal aim of this research was to determine a high resolution X-ray crystal structure of human DNA-PKcs, both *apo* and in complex with a variety of established and novel inhibitors, for the purposes of rational structure-based drug design. This thesis will discuss the work performed with this aim in mind, but ultimately working toward generating a soluble, correctly folded, stable protein for the purposes of biophysical and structural characterisation. Chapter 3 will delve into further detail regarding the attempts made to increase the solubility of the protein, as well as further attempts at construct manipulation to achieve this same goal. It will also go into further detail regarding the eventual denaturation and refolding of inclusion bodies and an eventual optimised protocol, with the subsequent aim of future work being scale-up of the protocol leading to structure determination.

1.6.2 – FEN 1

The aim of this portion of research was to determine an X-ray crystal structure for FEN 1 from *Pyrococcus abyssi*. Being from an archaeal system, the replicative machinery is similar to that of eukaryotes, but simpler, making it an ideal system to greater elucidate the structure and function of FEN 1. Another reason that *Pab* is a good model system is due to its behaviour as a piezophile. Taking advantage of this trait could potentially allow for the study of proteins at high pressure, more specifically performing high-pressure macromolecular crystallography (HPMX) to determine any changes from ambient to high pressure structures, possibly indicating reasons for thriving in these conditions. This thesis will go into further detail regarding some biophysical characterisation as well as direct comparisons with homologs from both other archaea such as *Sulfolobus solfataricus* and *Pyrococcus furiosus* and from human. Chapter 4 will cover these details as well as the work carried out in an attempt to co-crystallise *Pab* FEN 1 with DNA to determine its binding mode *in vitro*. This thesis will also shed light on preliminary high-pressure X-ray crystallography experiments, possible implications of this research and what it could mean in the future for the field of HPMX as a whole.

Chapter 2

Materials & Methods

2.1 Materials

2.1.1 – Media

All liquid media was stored at room temperature until required for use in experiments, at which point it was incubated at the desired temperature prior to supplementation with additives: Either antibiotics, cultures or inducing agents. All media was autoclaved prior to use. Solid media was obtained through the addition of agar to stocks of the aforementioned liquid agar. Table 2.1 contains ingredient information for all media used throughout the study.

2.1.2 – Antibiotics & Inducers

Antibiotics used throughout this study include Ampicillin/Carbenicillin, Chloramphenicol, Tetracycline and Kanamycin in various combinations. The formulations for stock solutions of each of these antibiotics can be found in Table 2.2. When the use of an inducer was required, the chemical Isopropyl-β-D-1-thiogalactopyranoside (IPTG) was used. Its stock formulation can also be found in Table 2.2.

Media	Composition
***Luria Bertani* (LB)**	1% Bacto-tryptone, 0.5 % yeast extract, 1 % NaCl; pH adjusted to 7.5 with NaOH
Terrific Broth (TB)	1.2% Bacto-tryptone, 2.4% yeast extract, 0.4% Glycerol, 10% 0.17M KH_2PO_4 and 0.72M K_2HPO_4
LB Agar	1 % Bacto-tryptone, 0.5 % yeast extract, 1 % NaCl, 1.5 % Bacto-agar; pH adjusted to 7.5 with NaOH
SOC	2% Bacto-tryptone, 0.5% yeast extract, 0.5% 5M NaCl, 0.25% (v/v) 1M KCl, 1% (v/v) 1M $MgCl_2$, 1% (v/v) 1M $MgSO_4$, 2% (v/v) 1M Glucose
Auto-Induction	1% Bacto-tryptone, 0.5% yeast extract, 0.33% $(NH_4)_2SO_4$, 0.68% KH_2PO_4, 0.71% Na_2HPO_4, 0.05% Glucose, 0.2% α-lactose, 0.015% $MgSO_4$, 0.003% Trace elements

Table 2-1 - A list of media used throughout this study and their corresponding compositions. Concentrations indicated as percentages are weight/volume (w/v) unless otherwise stated.

Antibiotic/ Inducer	Acronym	Solvent	Stock Conc^n	Working Conc^n
Ampicillin	Amp	H_2O	100 mg ml^{-1}	100 µg ml^{-1}
Carbenicillin	Carb	H_2O	100 mg ml^{-1}	100 µg ml^{-1}
Chloramphenicol	Cam	Ethanol	50 mg ml^{-1}	50 µg ml^{-1}
Kanamycin	Kan	H_2O	50 mg ml^{-1}	50 µg ml^{-1}
Tetracycline	Tet	Ethanol	12.5 mg ml^{-1}	12.5 µg ml^{-1}
Isopropyl-β-D-1-thiogalactopyranoside	IPTG	H_2O	1 M	1 mM

Table 2-2 - A Table of antibiotics and inducing agents at both stock and working concentrations (Conc^n) that were stored at -20°C until required.

2.1.3 – Purification resin

All resin and pre-packed columns used throughout this study was purchased from GE Healthcare and used according to their guidelines, unless otherwise stated.

2.1.4 – Crystallization reagents

For the purposes of screens testing various crystallisation conditions, various manufacturers were used. These included Molecular Dimensions HT-96 format, MIDAS™, Morpheus®, JCSG-plus screens. They also included Hampton Research Crystal Screen 1 & 2. AstraZeneca in-house screens were also used. The 'Original MRC Crystallization Plate' (SWISSCI) was used for all small scale screening in a sitting drop conformation. Larger scale hanging drop experiments were performed using 24-well Linbro plates (Hampton Research) with siliconised glass cover slides (Hampton Research).

2.1.5 – Enzymes

Restriction Enzymes used throughout the study were all purchased from New England Biolabs (NEB). GoTaq® and *pWo* polymerases were purchased from Peqlab. All were used in accordance with their corresponding manufacturer's guidelines.

Lysozyme, lyophilized from chicken egg white, was purchased from Sigma Aldrich and used, unless otherwise stated, according to their guidelines.

Trypsin for the purposes of peptide mass fingerprinting was kindly provided by Dr Andy Cronshaw, University of Edinburgh, and used in accordance with his protocols.

2.1.6 – Oligonucleotides

Oligonucleotides were either purchased from Sigma Aldrich or from GeneArt. Both came in dry form and were resuspended in nuclease free H_2O to a stock concentration of 100μM unless otherwise stated.

2.1.7 – Buffers

All buffers and solutions were stored at room temperature unless otherwise stated. All reagents and chemicals were added to double distilled, sterile H_2O (Milli-Q®)

EDTA

0.5M stock solutions were prepared and adjusted to pH 8 using concentrated NaOH, and then filtered through a 0.22µm filter membrane for storage.

Inoue Transformation Buffer

10.8g $MnCl_2.4H_2O$, 2.2g $CaCl_2.2H_2O$, 18.6g KCl, made up to 800 mL with filter sterilised H_2O stored at -20°C until required for competent cell production.

TAE (50x) Buffer

2M Tris base, 0.05M EDTA and 5.7% (v/v) glacial acetic acid made up to 1000 mL with filter sterilised H_2O stored at room temperature until required. The working concentration of TAE was 1x.

TGS (10x) Buffer

60.4g Tris base, 288g Glycine, 20g Sodium dodecyl sulphate (SDS) made up to 2 L with distilled water and filter sterilised. The working concentration of TGS buffer was 1x.

Western Blotting Buffer

60.4g Tris base, 288g Glycine made up to 2 L with distilled water and filter sterilised. The working concentration was 1x.

2.2 Methods

2.2.1 – Molecular Biology

2.2.1.1 – Plasmid Purification

In order to generate the purified plasmid DNA, it first had to be isolated from the bacterium. The *E. coli* strain that contained the desired plasmid was grown overnight in 5 mL of LB broth at 37°C. The plasmids were then extracted using a QIAprep® spin miniprep kit from Qiagen, using 250µL buffer P1, 250µL buffer P2 and 350µL N3, according to their guidelines. The plasmid DNA bound to the purification column was eluted using 50 µL nuclease free H_2O that had been heated to 70°C and then once eluted, it was stored at -20°C until required.

2.2.1.2 - Polymerase Chain Reaction

All polymerase chain reactions (PCR) were performed using a thermal cycler. Either GoTaq® Flexi DNA polymerase (Promega) was used, or in situations where the product was proceeding further with cloning, a polymerase possessing 3'-5' exonuclease activity, (*pWo* – Promega) was used. All reactions were performed according to the enzyme manufacturer's guidelines. The melting temperature (Tm) of the oligonucleotides used was calculated by their respective manufacturers and unless otherwise stated, annealing temperatures were 5°C below the lowest Tm in the reaction.

2.2.1.3 – Cloning

The purified DNA product of the PCR reaction was then digested by restriction endonucleases, used according to their manufacturer's guidelines. This digested DNA was then run on a 0.8% agarose gel (further information in section 2.2.2.2) at which point the desired band was excised and purified using a QIAquick Gel Extraction Kit (Qiagen) prior to

ligation. When linearising vector DNA, 1µL Calf Intestinal Phosphatase (CIP) (NEB) was added at the start of the reaction and again after 60 minutes in order to remove 5' phosphates from the digested DNA. This was in order to remove background caused by the vector ligating with itself forming an 'empty vector', resulting in false positives. Ligation was performed with T4 Ligase (NEB) overnight at 16°C. Ligation products were then transformed into competent *E.coli* XL1-blue strain (*endA- recA- hsdR*). *endA-* refers to a lack of endonucleases improving the quality of the miniprep DNA. *recA-* means the cells are recombination deficient, thus improving the stability of the insert. *hsdR* refers to a mutation preventing the cleavage of cloned DNA *via* the *EcoK* endonuclease system. Once transformed these cells could then act to propagate the plasmid DNA.

2.2.1.4 – Expression Vector

All DNA-PKcs constructs produced were ligated into the expression vector pTWO-E. This vector is a modified pET17-b vector (Novagen, Merck Chemicals) that contains an additional PreScission protease cleavable N-terminal 6xHis tag, designed for optimal expression in *E.coli*.

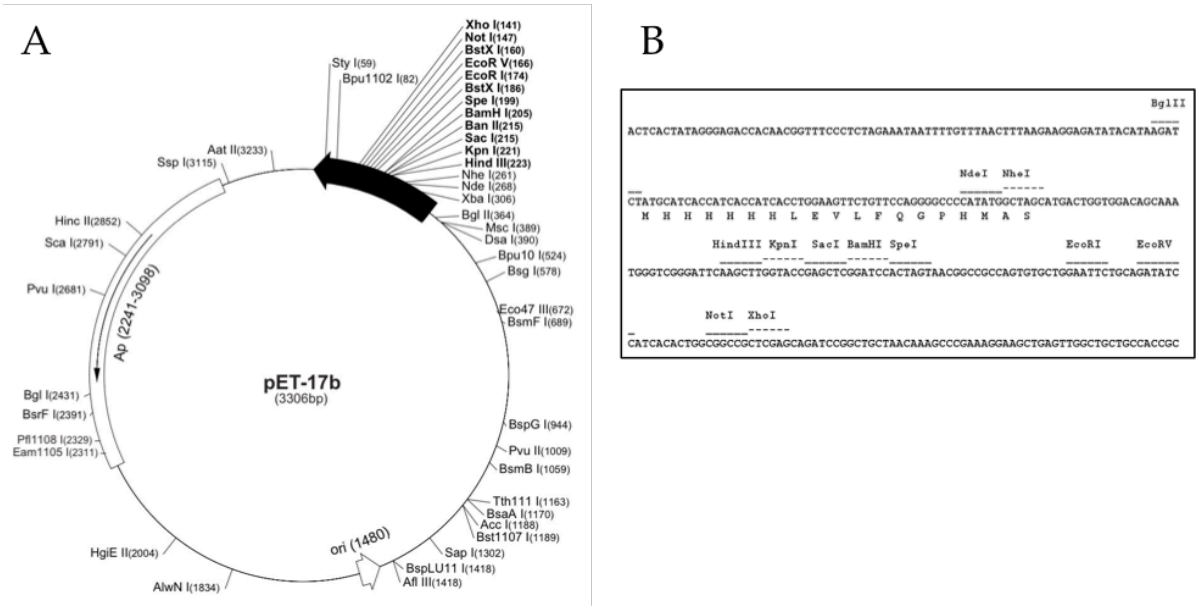


Figure 2.1 (A) - Vector map of pET17-b, the vector that pTWO-E is built upon. The modifications take place at the start of the multiple cloning site (MCS) highlighted, indicating the difference between the two vectors. All cloning performed using this vector made use of the restriction site NdeI for the digestion/ligation (101).

2.2.1.5 – BigDye® DNA Sequencing

The sequencing was performed at the GenePool, University of Edinburgh (https://www.wiki.ed.ac.uk/display/GenePool/Home). The plasmid DNA was sent with either the forward primer, (T7 or construct specific), or with the reverse primer (STO720 or construct specific) at which point the GenePool facility would perform the PCR and carry out the sequencing reaction using a BigDye® terminator v3.1 cycle sequencing kit (Applied Biosystems) according to the manufacturer's guidelines. Samples were then analysed using an ABI PRISM 3100-Avant Genetic Analyser.

2.2.1.6 – Codon Optimised DNA

All codon optimised DNA used throughout this study was designed using GeneArt's online software 'GeneOptimizer®' to modify codon usage from *Homo sapiens* to *E.coli*

2.2.1.7 – Annealing DNA

All oligos were provided by Sigma Aldrich and were resuspended to a final concentration of 300 μM. The three strands that were required to form the structure specific 'flap' were the upstream, downstream and template, further details of which will be discussed in chapter 4. The downstream and template strands were combined in a 1:1 ratio with annealing buffer (50 mM Tris-HCl pH 8.8, 50 mM NaCl, 1 mM DTT) and placed in a thermal cycler at 90°C for 5 minutes at which point the temperature dropped to 85°C. At this point an equivalent amount of upstream oligonucleotide was added (for a final ratio of 1:1:1) and it was left for another 5 minutes. This continued at 75°C, 70°C, 65°C all with 5 minute intervals at which point the thermal cycler was set to 10°C and this was the end of the annealing process.

2.2.2 – Microbiology

2.2.2.1 – Competent Cell Production

A scraping of a previous stock of the competent cells required was taken and plated onto a fresh LB agar plate with the appropriate antibiotic (or lack thereof) and grown for 18-24 hours at 37°C.

One colony was then removed from the plate and used to inoculate 25 mL SOC media in a 100 mL flask. This culture was then incubated for 6-8 hours at 37°C in an InFors HT Multitron shaking incubator at which point three individual 1 L flasks were inoculated with this starting culture. These flasks contained 250 mL LB media. The first flask was inoculated with 10 mL starting culture, the second flask had 4 mL added and the third had 2 mL added. All were then allowed to grow overnight at 18°C.

The following morning the OD_{600} measurements were taken and were monitored until one of the flasks had an OD_{600} of 0.55. At this point the corresponding flask was removed

and placed in an ice bath for 10 minutes. Cells were then harvested at 4000g using an Eppendorf 5810 R Centrifuge for 10 minutes at 4°C. The media was removed and the tube containing the whole cell pellet was inverted on paper towel to remove any excess media. The cell pellet was then resuspended in 80 mL ice cold Inoue Buffer and the centrifugation was repeated along with the drying of the pellet.

The cells were then resuspended in 20 mL ice cold Inoue Buffer. At this point 1.5 mL DMSO was added and the mixture kept on ice for 10 minutes. 100 µL aliquots were then dispensed into sterile micro-centrifuge tubes and flash-frozen in liquid N_2. The cells were then stored at -80°C until further use.

2.2.2.2 – Heat Shock Transformation

An aliquot of the desired strain of cells was removed from the -80°C freezer and allowed to thaw on ice for 20 minutes. A 1 µL aliquot of miniprep DNA corresponding to the desired construct was taken and placed into a fresh 1.5 mL eppendorf tube and left on ice. The competent cells were then added to the miniprep DNA and allowed to incubate on ice for 30 minutes. The cell/DNA mixture was then placed into a water bath at 42°C for a specific length of time, which varied according to the cell strain used. The times used were chosen based on manufacturer's guidelines. A list of cell strains used throughout this study can be found in Table 2.3.

Cell Strain	Genotype	Heat Shock Time
XL1-Blue	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac* [F′ *proAB lacI*^{q}*ZΔM15* Tn*10* (Tet^{r})].	45”
DH5α	*fhuA2 lac(del)U169 phoA glnV44 Φ80′ lacZ(del)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*	20”
Rosetta pLysS	F- *ompT hsdS*B(rB- mB-) *gal dcm* (DE3) pLysS (camR)	30”
BL21star	F- *ompT hsdS*$_B$ ($r_B^- m_B^-$) *gal dcm rne131* (DE3)	30”
BL21gold	F$^-$ *ompT hsdS*(r^- m^-) *dcm*$^+$ Tet^r *gal* λ(DE3) *endA* Hte	30”
BL21-AI™	F– *omp*T *hsd*SB(rB–, mB–) *gal dcm ara*B::*T7RNAP-tetA*	30”

Table 2-3 - List of cell strains used throughout this study.

After heat shock the cells, now containing the plasmid DNA, were then allowed to incubate on ice for a further 15 minutes. At which point 500 µL SOC media was added and the cells were incubated at 37°C for 60 minutes in a shaking incubator. After this period 200 µL and 400 µL of the mixture were plated onto two fresh LB agar plates with corresponding antibiotic and placed, inverted, in a stationary incubator set at 37°C for 18-24 hours. Plates were then stored at 4°C for a maximum of 1 week.

2.2.3 – Electrophoresis

2.2.3.1 – SDS-PAGE

Sodium Dodecyl Sulphate–Polyacrylamide Gel Electrophoresis (SDS-PAGE) was performed using tanks manufactured by either BioRad or Invitrogen. They were run at constant amperage of 50 mA for various lengths of time. When casting gels, 10%, 12%, 15% or 20% acrylamide was used. Solutions contained filter sterilised H_2O, Acrylamide at the

given concentration, Tris-HCl (1M at pH 8.8 resolving) (1.5M at pH 6.8 stacking), 10% SDS, 10% Ammonium Persulphate (APS), Tetramethylethyldiamine (TEMED). Pre-cast gels were either Any kD™ Mini-Protean® TGX™ (Biorad) or Novex® 4%-12% Tris-Glycine 1.0 mm mini gels (Invitrogen). Protein standards were either Precision Plus Protein™ Dual Colour Standard (Biorad) or SeeBlue® Plus2 Pre-Stained Standard (Invitrogen). After running the gels they were stained with InstantBlue protein stain (Expedeon) for 30 minutes and destained in filtered, sterilised H_2O.

2.2.3.2 – Agarose Gel Electrophoresis

Agarose gels were run at constant amperage of 50 mA for various lengths of time. The gels were cast at either 0.8% or 1% w/v agarose in 1x TAE buffer. Electrophoresis running buffer consisted of 1x TAE buffer. DNA standard markers were either 1kb DNA Ladder (NEB) or 100bp DNA Ladder (NEB). Visualisation was performed using a high-sensitivity charged couple device (CCD) gel imaging system: a GelDoc™ XR+ System (Biorad).

2.2.3.3 – Western Blot

Western blots were performed immediately after analysis by SDS-PAGE. Membranes used to trap the proteins during transfer included Hybond-P PVDF membrane (Amersham), and Hybond ECL nitrocellulose membrane (Amersham). All transfers were performed using a PerfectBlue™ Tank Electro Blotter Web™ S (Peqlab) at 300 mA for 120 minutes in western blotting buffer. When using PVDF membrane, the membrane was incubated for 10 minutes in 100% ethanol prior to running. Sponge and filter paper was used either side of the gel/membrane, both of which had been incubated in western blotting buffer forming the 'sandwich'.

After running, the membrane was removed and placed into a solution of 5% w/v Skim Milk Powder (Sigma Aldrich) in Phosphate Buffered Saline with 0.2% v/v Tween 20 (PBST) and incubated for 60 minutes at room temperature. At which point two washes were

performed with PBST only and then a solution of PBST in combination with a primary antibody was added to the membrane and another incubation of 60 minutes at room temperature was performed. If a secondary antibody was required the wash step was repeated and the new antibody added to a fresh PBST solution for further incubation. A list of antibodies used throughout the study can be found in Table 2.4.

If the antibody was conjugated to alkaline phosphatase (AP) then the membrane was developed with Sigma*Fast*™ BCIP®/NBT (Sigma Aldrich) tablets resuspended in PBST (1 tablet / 10 mL PBST) and imaged using GelDoc™ XR+ System (Biorad). If the antibody was conjugated to Horseradish Peroxidase (HRP) then the membrane was developed with ECL Prime Western Blotting Detection Reagent (Amersham) and imaged using Kodak BioMax MR-1 X-ray film and a standard X-ray film developer.

Antibody	**1º / 2º**	**Host**	**Conjugation**	**Stock Concⁿ**	**Working Concⁿ**
α- 6xHis	1º	Mouse	AP/un-conjugated	1 mg/ml	0.5µg/ml
α – DNA-PKcs	1º	Rabbit	un-conjugated	0.3 mg/ml	0.5µg/ml
α – GST	1º	Goat	un-conjugated	1 mg/ml	0.5µg/ml
α – goat	2º	Donkey	HRP	1 mg/ml	0.1 µg/ml
α – rabbit	2º	Mouse	HRP	0.8 mg/ml	800 ng/ml
α – mouse	2º	Rabbit	HRP	2 mg/ml	0.2 µg/ml

Table 2-4 - List of antibodies used throughout this study.

2.2.4 – Protein Production

After transforming the competent cells with the desired construct, a single colony was picked from the LB agar plate and used to inoculate 5 mL LB media with corresponding antibiotic, for selection purposes. This was then allowed to grow for 6-8 hours at 37°C in a shaking incubator. This culture was then used to inoculate 100 mL of LB media containing the same antibiotic, which was then grown overnight at 37°C.

5 mL of this was taken and added to a 2 L conical flask containing 500 mL of media (TB, LB or AI) and antibiotic. The ratio of 1:4 volume of media: volume of flask was maintained throughout all experiments. The culture was then incubated for 1 hour at 37°C in the shaking incubator at which point the OD_{600} was measured, and measurements were taken repeatedly every 30 minutes to assess the doubling time of the organism. When the cells reach mid-logarithmic scale of growth and the optimal time to induce the growth of the protein, IPTG was added to a final concentration of 1 mM unless otherwise stated. This phase of cell growth is determined by measuring the $OD_{600.}$ Depending on the media this value can vary and is affected by the maximum optical density that can be reached with the media. LB media has a maximum OD_{600} of between 2-3 and it is between 0.6 and 0.8 that logarithmic growth occurs. TB however can reach a maximum OD_{600} of between 5-8, and therefore the logarithmic scale extends much higher to between 0.6-1.2. Initial time course experiments determined the optimal induction time and this time (either 1 hour, 2 hours, 3 hours, 4 hours or overnight) was then used for all subsequent experiments.

Once the optimal induction time had been reached the cells were harvested using a Beckman Coulter Avanti J-30I High Performance Centrifuge at 8000g for 30 minutes at 4°C The pellets were then collected and stored at -80°C for future use.

2.2.4.1 – Buffer Scouting

When solubility was an issue with a target protein, the buffer constituents and parameters were subjected to buffer scouting to increase yields of soluble protein. In all cases the cell pellet was initially lysed in 20 mM HEPES pH 7.5, 200 mM NaCl, 10 % Glycerol, 5 mM $MgCl_2$ to determine a baseline level of solubility. At this point the buffer scouting would begin. Five different buffer constituents were chosen: HEPES (buffering range pH 6.8-8.2), Tris (buffering range pH 7.5-9.0), MES (buffering range pH 5.5-6.7), PIPES (buffering range pH 6.1-7.5) and CAPS (buffering range pH 9.7-11.1). These buffers were all made to a final concentration of 20mM and were chosen in order to cover pH ranges both more acidic and more basic than the theoretical pI of the protein, which for DNA-PKcs was 8.44. The NaCl (200mM), glycerol (10%) and $MgCl_2$ (5mM), were all kept constant, as well as the temperature, all buffer scouting experiments were performed at room temperature. A successful buffer condition was defined as the one that had either the largest yield of soluble protein or the largest amount in the supernatant relative to its pellet fraction.

2.2.4.2 – Cell Lysis

When lysing cells with lysozyme, firstly the cell pellet was resuspended in 20mL buffer, determined after initial buffer scouting, at which point the lysozyme mixture was added, which consisted of Tris-HCl pH 8.0, 1.2% v/v 0.5 M EDTA pH 8.0, 1% w/v lysozyme lyophilized from chicken egg white (Sigma Aldrich). Once this lysis solution was added, the cell pellet/lysozyme solution was incubated with rotation for 60 minutes at room temperature.

When lysing the cells *via* sonication the cell pellet was resuspended in the desired buffer and kept on ice. At which point it was placed in to a Soniprep 150 sonicator with the MSE 9.5 mm diameter probe inserted into the solution. With the cell suspension remaining

on ice, the cells were sonicated in 3 x 30 second bursts, with 30-second intervals in between unless otherwise stated.

When using a French press, the Constant Systems Ltd 1.1 kW TS Cell Disruptor was used. After washing with filtered, sterilised H_2O and detergent, the sample was placed into the reservoir and passed through the system according to the manufacturer's guidelines. The lysed product was then collected.

Prior to any cell lysis, cOmplete EDTA-free protease inhibitor cocktail (Roche) was added to the buffer according to the manufacturer's guidelines.

After all cell lysis, the mixture was clarified *via* ultracentrifugation at 25,000g for 60 minutes at 4°C and the pellet and supernatant were separated for future analysis.

2.2.5 – Protein Purification

2.2.5.1 – Affinity

All purifications with pre-packed columns were used alongside either an ÅKTAprime or ÅKTAexpress (GE Healthcare) and their associated software. All columns and purification systems were used according to the manufacturer's guidelines.

Ni-Sepharose:

Affinity purification was performed either with resin not bound to a column matrix, which will be discussed further in Chapter 2.2.4.3, or with pre-packed Ni-Sepharose 1 mL and 5 mL 'HisTrap' Columns (GE Healthcare) when using a protein designed to have an N-terminal 6xHis tag.

The supernatant contained 10mM imidazole in order to minimise non-specific binding, and was passed over the pre-equilibrated column during the binding process. Two distinct wash steps were performed, firstly at 50mM imidazole, secondly at 100mM imidazole to further remove non-specific or weak interactions with Ni-sepharose resin.

At this point the elution stage began and this was performed with an increasing concentration of imidazole either *via* gradient or step-wise. When performing gradient elution methodology, the initial imidazole was 100mM imidazole, increasing to 500mM over 40 column volumes (CV). This then resulted in elution of the purified tagged protein to be later analysed by SDS-PAGE and western blot, at which point positive fractions were taken forward in the purification protocol.

When a step-wise elution methodology was used, the imidazole concentration was 100mM, with periodic increases in concentration over 10 CV, the intervals of which varied from 100mM to 200mM.

Glutathione-Sepharose:

When using an N-terminal GST tag, Glutathione-Sepharose 1 mL 'GSTrap' columns (GE Healthcare) were used. The competing ligand with this resin is reduced glutathione.

The supernatant contained 10mM reduced glutathione in order to minimise non-specific binding, in a manner similar to previously discussed affinity purification methods, and was passed over the pre-equilibrated column during the binding process. The wash step discussed in previous affinity purification methods was omitted when purifying GST-tagged proteins.

Elution was performed with an increasing concentration of reduced glutathione competing ligand either *via* gradient or step-wise. When performing gradient elution methodology, the initial concentration of reduced glutathione was 10mM, finishing at 20mM reduced glutathione, over 40 column volumes (CV). During this research gradient elution was the only method used when using this particular resin. All subsequent analysis was performed in the same manner as previous affinity chromatography experiments.

2.2.5.2 – Ion Exchange

Anion exchange was performed with pre-packed 1 mL HiTrap Sepharose columns. Strong anion exchange resin Q was used along with weak anion exchange resin DEAE. Cation exchange was performed with the same 1 mL HiTrap Sepharose. However for weak cation exchange CM resin was used and SP resin was used as a strong cation exchanger. All columns were manufactured by GE Healthcare and used according to their guidelines.

Making use of the theoretical isoelectric point (pI) of the protein at a given pH allows for prediction of binding to a particular ion exchange resin. Theoretically, when the pH of a solution is above the pI of the protein, the net surface charge on the protein will be negatively charged and therefore bind positively charged (anionic) resin and *vice versa*. Although this is not always the case binding affinity experiments were carried out on a small scale, which will be discussed further in Chapter 2.2.4.3.

The supernatant was passed over a pre-equilibrated column at 20-50mM NaCl in the buffer, resulting in a weak ionic strength. Once bound, the salt concentration was raised to 100mM to act as a wash step. At which point an increasing salt concentration was applied either in a gradient or step-wise fashion as mentioned previously. This increase in ionic strength of the buffer weakens the interaction of the protein to the resin and it is eluted at a given salt concentration. Fractions were then analysed by SDS-PAGE and any containing the desired protein were pooled and taken forward through the remaining purification protocol.

In some respects Heparin Sepharose can be identified as both an affinity resin and an ion exchange resin. DNA binding proteins will have an affinity for the resin but it will also act as a cation exchange column and bind positively charged molecules.

2.2.5.3 – Batch Purification

Before developing the purification protocol, optimal binding affinity had to be determined, and this was done through the use of resin not packed in a column. Ni-Sepharose, GST-Sepharose, Heparin Sepharose, Q-Sepharose, DEAE-Sepharose, CM-Sepharose and SP-Sepharose were all purchased from GE Healthcare and used according to their guidelines.

An amount of unbound resin that is stored in 20% ethanol was taken and dispensed either into an eppendorf tube or a falcon tube depending upon the scale of the experiment. Small scale experiments used 100µL resin, whilst the largest scale experiments used 1mL of resin. It was then centrifuged at 800g for 5 minutes and the 20% ethanol was removed. An equal amount of equilibrating buffer was then added to the resin and it was resuspended briefly before another round of centrifugation. This was repeated at least 3 times in order to full equilibrate the resin to the desired conditions.

The protein sample was then applied to the resin and it was incubated for 20 minutes with rotation at either at 4°C. The tube was then centrifuged at 800g for 10 minutes and the flow-through collected. Wash buffer was then added in exactly the same manner. It was either identical to the loading buffer, or the concentration of competing ligand was increased, to either 100mM NaCl for ion exchange, 40-100mM imidazole for Ni^{2+} affinity chromatography or 10mM reduced glutathione for GST-affinity chromatography. After centrifugation this wash fraction was removed. 500µl elution buffer was then added to the resin. This buffer contained the highest concentration of competing ligand required to elute all of the adsorbed particles (1 M NaCl, 500 mM imidazole, 20 mM reduced glutathione). After incubation and centrifugation this sample was collected and all fractions were analysed *via* SDS-PAGE and western blot to determine the optimum binding conditions for the protein in question.

2.2.5.4 – Size Exclusion Chromatography

All size exclusion chromatography (SEC) was performed on either a Superose 6 10/300 GL pre-packed Sepharose column (GE Healthcare) or a Superdex 200 10/300 GL Sepharose column (GE Healthcare). All SEC was performed alongside either an ÅKTAprime or ÅKTAexpress (GE Healthcare) and their associated software. Both columns were used according to the manufacturer's guidelines.

Initially the column was equilibrated with 2 CV of buffer. This buffer changed depending on the preceding step. SEC performed after Ni^{2+} affinity chromatography contained 20mM HEPES, 50mM NaCl, 20mM imidazole. After ion exchange the buffer contained 20mM HEPES, 400mM NaCl. The protein solution was then concentrated to a final volume of either 200 µL. The loop was emptied with 3 loop volumes (LV) onto the column and then removed from the flow path. The flow rate was kept at 0.5 mL / minute for the entirety of the run. The protein present in the solution was then separated based on its hydrodynamic radius.

A protein with a small hydrodynamic radius will access a considerably larger amount of the available pores present in the Sepharose beads than a protein with a large hydrodynamic radius. Therefore larger proteins are eluted more quickly than than the smaller ones due to having a shorter flow path. This then leads to separation of proteins according to their size and allows for the technique to be used as a polishing step to remove impurities such as oligomers or aggregates from the desired protein which, when eluted, should be contained within a suitable buffer at homogeneous size. The UV absorbance, measured in mAu (milli-Absorbance units), was used to identify peaks being eluted from the size exclusion column. 0.5mL fractions were collected and analysis of these fractions was then performed *via* SDS-PAGE and western blotting. This analysis allowed the determination of which fractions contained the purest target protein, which were subsequently collected, pooled and concentrated for further experimentation.

2.2.5.5 – Purification in Denaturing Conditions

Throughout the study, some protein had to be obtained by attempting to solubilise proteins from inclusion bodies, present in the insoluble pelleted fraction after lysis and centrifugation. This involved several rounds of washing to remove as many impurities from the inclusion bodies as possible. These impurities included cell membranes and membrane proteins.

Firstly the pellet and supernatant were separated as mentioned in Chapter 2.2.4.1. The pellet was then resuspended in the same lysis buffer at a ratio of 1:25 w/v (pellet : buffer). Resuspension was performed using a CAT X120 Homogeniser (Bennet Scientific Limited) to create a fully homogenous solution. This solution was then centrifuged at 25,000 rpm for 30 minutes.

This supernatant was then removed and kept for analysis. The pellet then underwent 6 further wash steps with changes to the buffer each time. A wash being defined as resuspension of the pellet in buffer with the homogeniser mentioned previously, centrifugation to collect the inclusion bodies present in the pellet fraction and removal of the supernatant for analysis. Washes 2 and 3 used lysis buffer containing 0.1% (v/v) Triton™ X-100 (Sigma Aldrich) in order to solubilise the membranes and membrane proteins. Washes 4 and 5 used lysis buffer containing 1M NaCl and 0.1% (v/v) Triton™ X-100, in an attempt to try to remove nucleic acid contamination.

The final wash used lysis buffer only. This was done to remove traces of detergent and normalise the salt concentration, so that they could not otherwise affect how the protein reacted to the denaturant. The final inclusion body pellet was resuspended in a solution of lysis buffer containing either 8M Urea (Sigma Aldrich) or 6M Guanidine Hydrochloride (Sigma Aldrich) to a final w/v ratio of 1:25. The pellet was fully resuspended using the homogeniser and left to incubate overnight at 4°C with rolling to fully denature all proteins

still remaining in the pellet. The denatured protein mixture was then centrifuged at 25,000 rpm for 60 minutes and the solubilised fraction and pellet were separated.

Samples were then analysed *via* SDS-PAGE and western blot before being taken forward for further purification and renaturing.

2.2.5.6 – Refolding

After denaturing the protein, the denaturant had to be removed from the buffer in a controlled manner as to allow the secondary and tertiary structures to form in the correct manner, leading to a stable, functionally active refolded protein. There were 3 methods of refolding employed during this research: 'on-column', dilution and dialysis.

When the protein was refolded 'on-column', the protein-denaturant solution was loaded on to an equilibrated 10 mL HisTrap column at room temperature. At this point the concentration of denaturant was decreased over a shallow gradient of 60 CV, from either 8M urea or 6M guanidine hydrochloride to a final concentration of 0 mM at a flow rate of 1 mL/minute, allowing the protein to slowly fold into a secondary and tertiary conformation around the scaffold of the interaction of the 6xHis tag with the Ni-sepharose resin.

When refolding was performed *via* dilution, it was done so as an intermediate step, and technically cannot be considered a refolding step as it is being performed in the presence of a denaturant, although at concentrations lower than their optimum for denaturing protein. This involved rapidly diluting a protein sample into a lower concentration of an experimentally determined alternative denaturant (2 M Guanidine), which could then be further processed by dialysis, thus removing this lower concentration of denaturant. Dilution of protein sample into refold buffer was performed using a Mosquito LCP® Liquid Handling Robot (TTP Labtech) for the small scale experiments, and with a pipette at the larger scale. The ratio of protein : refold buffer was 1:50, which equated to 24 nL protein : 1.2 µL refold buffer for initial screens, 100 µL protein : 5 mL buffer for the intermediate scale and 1 mL

protein : 50 mL buffer at the largest scale performed, all of which were performed at 18°C. The protein concentration at the initial stage of refolding was 5 mg/mL. The refolds were also performed both in the presence and absence of ligands to determine their effects.

When refolding protein *via* dialysis, it was always performed with Dry Spectra/Por® regenerated cellulose membrane (Spectrum Labs) with a pore size of 6-8 kDa and varying flat widths depending on the final volume. All dialysis was performed at 4°C with stirring, at a ratio of protein : dialysis buffer of 1:500, with 50 mL protein, at 2 M guanidine, to 25 L dialysis buffer. This consisted of 20mM Tris, pH 8.1, 50 mM NaCl. There were a total of 4 changes of dialysis buffer over the course of the experiment, with at least 8 hours in between changes to allow enough time for equilibration. This therefore reduced the concentration of denaturant from 2000 mM to approximately 32 pM.

2.2.6 – Biophysical Characterisation

2.2.6.1 – Dynamic Light Scattering

Dynamic Light Scattering (DLS) was used during this study in correlation with SEC previously mentioned in order to determine the hydrodynamic size of the proteins being tested. Experiments were carried out on a Zetasizer Auto Plate Sampler (Malvern Instruments Ltd) and were done so according to the manufacturers guidelines.

Samples were prepared to a final volume of 65 µL into a 384 well plate (Malvern Instruments Ltd), at more than one concentration. For the DNA-PKcs work the concentrations used were 0.25 mg/mL and 0.125 mg/mL. For the FEN 1 work the concentrations used were 18 mg/mL and 1 mg/mL. The plate was then inserted into the machine and allowed to equilibrate to a predetermined temperature of 25°C. A standard operating procedure was then created whereby the information for each sample (solvent constituents & protein) was uploaded to the software. Each measurement was repeated a total of 3 times in order to generate reproducibility. The result for the experiment is a

measurement of the distribution of particle size within the solution in relation to both the intensity and volume, with the readout showing this distribution on a logarithmic scale, with particle sizes being shown as a diameter in nm.

2.2.6.2 – Thermal Denaturation Assay

Thermal denaturation assays were performed during this study to determine the thermal stability of the target protein. All assays were performed on an IQ5 ICycler (Biorad), a real time PCR (rt-PCR) machine whose use was dedicated to this assay. It, along with the coordinated computer software, was used according to the manufacturers guidelines.

An iQ™ 96-well PCR plate (Biorad) was used for the reactions. Each sample was placed into a well (in triplicate) to a final volume of 45 µL. 5 µL of 5x SYPRO® Orange protein stain (Invitrogen) was then added to a final concentration of 1x. 5x solution was made by combining 1 µL of stock 5000x SYPRO® + 999 µL H_2O. The starting temperature was set at 25°C, with a ΔT of 1°C / minute for initial screening experiments and a ΔT of 0.5°C / minute for more accurate results. The SYPRO® Orange protein stain binds to hydrophobic regions on the surface of the protein, which emits fluorescence upon binding to these regions. As the temperature increases and the protein begins to unfold, more binding events occur leading to an increase in fluorescence, which is detected by the CCD-based optics within the iQ5. At the melting temperature of the protein this process occurs at a much faster rate, which can be observed as a peak on the readout, which is shown as relative fluorescence units (RFU) vs time as well as the negative first derivative of the fluorescence change (-dRFU/dT).

Due to the fact that FEN 1 is a thermostable protein, the final temperature for all of the corresponding TDA experiments was set to the maximum level 95°C. The DNA-PKcs experiment temperatures didn't need to be set as high as they were for FEN 1, but still

varied, details of this will be discussed further in the results section. Results were analysed by the IQ5 software for further interpretation.

2.2.6.3 – Peptide Mass Fingerprinting

Peptide mass fingerprinting was performed to confirm the identity of proteins being purified. After purification the samples were analysed *via* SDS-PAGE and western blot. The SDS-PAGE serves to identify an estimated size of the protein in relation to a given standard run alongside. Western blotting can then confirm the identity by using the antibodies raised against them. Although these techniques can show an approximate size, as well as the presence of a specific epitope, when using the appropriate antibody raised against it, it cannot provide an absolute mass or identity. This information can be obtained however through the use of peptide mass fingerprinting. All fingerprinting was performed using Matrix Assisted Laser Desorption/Ionisation – Time of Flight (MALDI-TOF) mass spectrometry. The matrix used was α-cyano-4-hydroxycinnamic acid (CHCA) and it works by rapidly absorbing UV energy, converting it to heat and vaporising the sample. These vaporised particles are then accelerated through a drift tube toward a detector. The time taken to travel this distance is proportional to the square root of their molecular weight.

Firstly the test sample had to be prepared. The samples were analysed by SDS-PAGE and subsequently stained with Instant Blue protein stain. Instant Blue is used, rather than coomassie stain due to the lack of methanol in the solution that would otherwise fix the protein to the acrylamide gel rendering further treatment null. Once the band was stained it was excised using a scalpel as tightly as possible, with minimal excess of acrylamide as excess acrylamide can provide large peaks that can mask the peptide peaks in the trace.

This gel piece was then incubated in 300 µL NH_4HCO_3 (ABC) in 50% acetonitrile (ACN) at room temperature for 30 minutes. This was then repeated 2 further times in order

to remove SDS. The band was then incubated in 300 µL 20 mM DTT, 200 mM ABC, 50% ACN at room temperature for 60 minutes in order to reduce the protein.

The band was then washed 3 times in 300 µL 200 mM ABC, 50% ACN. Cysteines present in the protein had to then be alkylated using 100 µL 50 mM iodoacetamide (IAA), 200 mM ABC, 50% ACN at room temperature in the absence of light for 20 minutes. A further 3 wash steps with 20mM ABC, 50% ACN were then performed.

The band was then cut into 2 mm x 1 mm pieces and centrifuged at 13,000 rpm for 2 minutes. The gel pieces were then completely submerged in 100% ACN until they turned white in colour, at which point the ACN was removed and the gel pieces were allowed to dry at room temperature.

A solution of 29 µL 50 mM ABC with 1 µL 1 mg/mL trypsin (0.034 mg/mL final concentration) (Promega – provided by Dr Andrew Cronshaw) was then added to the dried gel pieces. This was kept at 4°C until the gel pieces swelled, at which point the eppendorf tube was placed at 32°C for 16-24 hours along with a blank containing the same solution of 50 mM ABC with 1 µL trypsin but no gel pieces.

After sample preparation 0.5 µL of the digestion solution was mixed with 0.5 µL of CHCA matrix solution and was spotted onto a stainless MALDI plate. The samples were then analysed on a Voyager DE-STR MALDI-TOF Mass Spectrometer (Applied Biosystems) and after using Data Explorer software (Applied Biosystems), the processed spectra were directly compared against entries within the National Centre for Biotechnology Information (NCBI) non-redundant database using the MASCOT Server (Matrix Science) to determine their identity.

2.2.7 – Crystallisation

2.2.7.1 – Principles of crystallisation

To determine the X-ray structure of a protein, one cannot rely on the scattering of X-rays from electrons within the protein, simply due to it not being powerful enough to generate a signal. To overcome this, a protein can be grown into a crystal of repeating units thus amplifying the X-ray scattering. This amplification occurs due to the nature in which a crystal is formed. A crystal is made up of a structural motif, in this instance the protein, arranged in a lattice of regularly repeating units. These regularly repeating units are termed unit cells defined by three lengths and three angles (a, b, c, α, β, γ). The unit cell will usually contain one or several copies of the molecule, termed the asymmetric unit (AU), which have symmetry elements applied to them in order to generate the unit cell. An example of the unit cell and AU can be seen in figure 2.2.

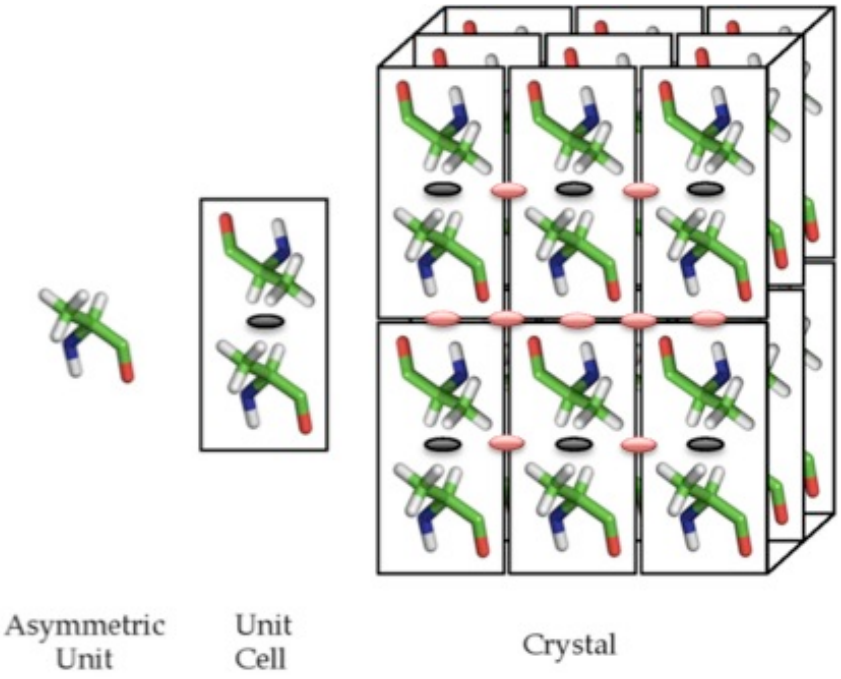


Figure 2.2 – Representation, using the amino acid Alanine, of the asymmetric unit and how it forms the unit cell, in this case containing a two-fold rotation axis perpendicular to the page, which in turn forms the crystal. Other rotation axes present within the crystal are also two-fold rotation axes. They indicate how one unit cell relates by symmetry to others within the lattice.

The regularity of a crystal means that the X-rays interact with it, producing a discontinuous diffraction pattern, meaning that the scattered X-rays are only observed at distinct points, the individual waves of which are referred to as reflections and the intensity of each of which is measured. A perfect crystal will have all unit cells perfectly aligned with one another and the diffraction from each unit cell will be perfectly aligned with that from each of the other unit cells. This will then produce a mosaicity value of 0. The mosaicity is the width of the distribution of mis-orientation angles of all unit cells in a given crystal. A high mosaicity will result in broader spots. In reality there are almost always packing imperfections in the crystal as a result of either growth defects or flash freezing. The higher the mosaicity, the larger the spread of diffraction which means a greater angle of rotation is required to collect the full intensity for each individual reflection.

2.2.7.2 – Sitting & Hanging Drop

There are a variety of crystal growth techniques, including liquid/liquid diffusion, sublimation and vapour diffusion. In the interests of brevity the only method to be discussed here will be vapour diffusion as this was the only crystal growth method used throughout the experiments.

Figure 2.3 shows a schematic of the vapour diffusion technique using either tissue culture plates, or custom-made crystallisation plates. Initial screens made use of a Gryphon LCP® Liquid Handling Robot (Art Robbins), in a sitting drop conformation. In this experiment 100 nL of protein at a concentration of 2 mg/mL was combined with 200 nL of precipitant in the drop and 1 µL precipitant in the reservoir. Commercial screens used throughout these studies included Molecular Dimensions HT-96 format, MIDAS™, Morpheus®, JCSG-plus and Hampton Research Crystal Screens 1 & 2. Larger scale studies were performed with pipettes manually in a hanging drop conformation. Once the drops were set up the sitting drop plates were sealed using EasySeal™ sheets (Molecular Dimensions) whereas the hanging drop plates were sealed using a cover slip and silicon grease.

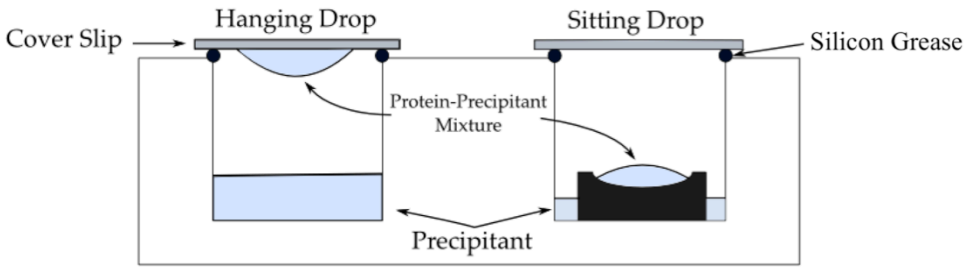


Figure 2.3 - Schematic showing the two vapour diffusion methods used throughout these experiments, hanging drop and sitting drop. Both of which work on the same principle.

When the drop was first set up, it had a concentration approximately half that of the reservoir, because the drop contains both protein solution and reservoir solution in a 1:1 ratio. When changing this ratio the actual drop concentrations of protein and precipitant can vary. Equilibration then occurred, through the vapour phase, leading to an increase in the concentration of the precipitant in the drop to roughly equal that of the reservoir. This left the reservoir concentration practically unchanged due to the vast difference in volumes of the reservoir compared to the drop. The concentration of both the precipitant and the protein in the drop however have doubled due to the evaporation of water from the drop resulting in a smaller drop volume. This subsequent increase in concentration of both the protein and precipitant changes the internal environment of the drop, a change that can be visualised with a two-dimensional phase diagram, illustrated in figure 2.4:

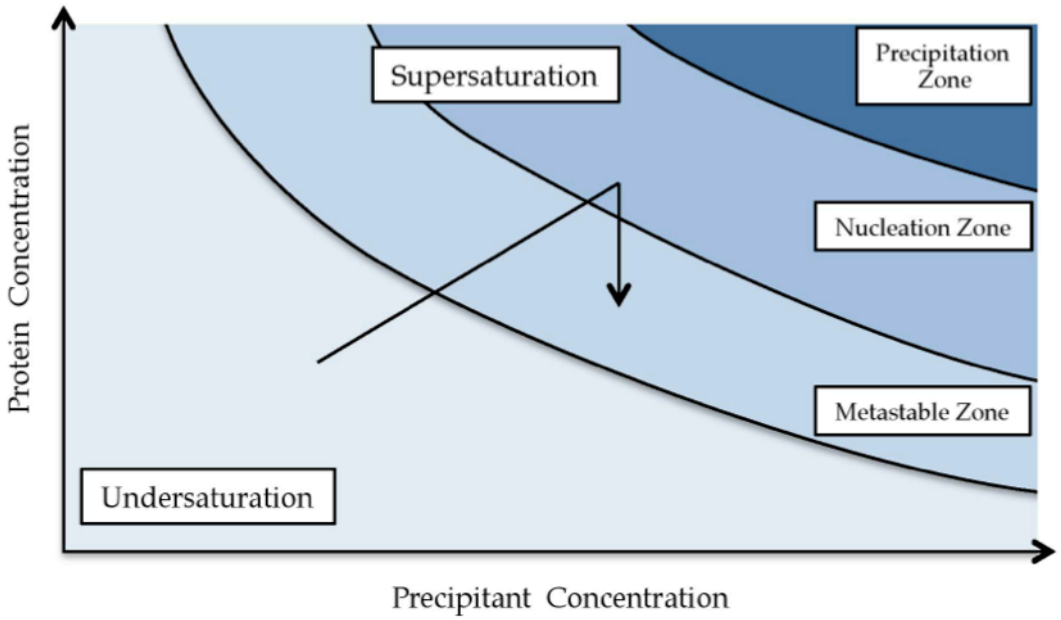


Figure 2.4 - A two-dimensional phase diagram. Below the first curved line, indicating the saturation point, is termed undersaturation, and above this curve is termed supersaturation. Supersaturation can then be divided into: the metastable zone - whereby supersaturation is too small and the nucleation rate is too slow to form crystals; the nucleation zone - where supersaturation is large enough that spontaneous nucleation can occur; and the precipitation zone – where crystals cannot form due to aggregates and precipitation forming faster than any crystals can form (102).

In a perfect scenario during the vapour equilibration, the concentration of the drop will increase so that it is no longer undersaturated but is concentrated enough to form spontaneous nucleation and small, nucleated crystals will begin to form. As they form the concentration of protein in the drop decreases moving into the metastable zone, allowing for the nucleated crystals to stabilise and grow larger into crystals that can be used to collect diffraction data. This movement is depicted by the arrow in figure 2.3. In practice however the variables are much more complicated and diverse and cannot be determined theoretically, they have to be determined in a trial-and-error methodology. In practice this trial-and-error methodology involves large scale crystal growth screens with iterative rounds of refinement to finally determine the optimum set of conditions to get suitable, reproducible crystals to be used for data collection.

2.2.7.3 – Screening

Initial screening experiments were performed to maximise the chances of successfully growing crystals. These screens allowed testing of a vast array of conditions whilst still using minimal amounts of protein. Both commercial and in-house crystal screens were used, the identities of which will be discussed in chapter 4. Using a Mosquito LCP® Liquid Handling Robot (TTP Labtech), unless otherwise stated, aliquots were taken from the 96-well Deep Well blocks containing the screen and placed into the corresponding wells in the 96-well MRC plate mentioned in the previous section. The plates were then sealed and left at 18 °C to allow the crystals to form. The drops were then examined *via* a light microscope to determine which conditions yielded good crystals or at least a good starting point for optimisation.

2.2.7.4 – Optimisation

Optimisation began with the conditions found through screening, as previously mentioned, that yielded nucleation that could be improved upon. This nucleation was in the

form of micro-crystals or needle-like crystals. Precipitation was used as a determination of poor conditions and was used to exclude crystallisation conditions from subsequent experimentation. This involved varying the concentrations of the various constituents within the precipitant solution (salt, buffer, precipitant) as well as the final concentration of the protein and other parameters such as the pH. All optimisation was performed using 24-well Linbro plates to increase the potential for greater numbers of crystals or physically larger crystals more suitable to testing due to their decreased susceptibility to X-ray radiation. All solutions were dispensed manually using standard single channel or multi channel pipettes.

2.2.7.5 – Seeding

The process of seeding involved taking a drop that contained high levels of microcrystals, and placing it into a 1.5 mL eppendorf tube. A glass capillary with a sealed end was then used to grind the crystals down. The eppendorf was then centrifuged at 13,000 rpm for 60 seconds. This solution was then either used in this state or serial dilutions were produced in order to have a seeding stock. Further optimisation plates were then set up and the drops were set up in the same manner as previous with the exception that in this instance an equal measure of crystal stock was also added to the drop.

2.2.8 – Principles of X-ray crystallography

In order to determine the structural information from the protein crystal that has been grown, X-rays that have been passed through both a monochromator and collimator to get a fine beam of uniform fixed wavelength are shot at the crystal and subsequently diffracted. These diffracted X-rays then collide with a detector, which can measure the intensity and position of the diffraction pattern, which can then be used to determine the structure of the protein.

This diffraction occurs due to parallel planes within the crystal. They are added constructively to one another to generate a single reflection, or spot on the diffraction

pattern. The conditions required for this constructive interference, and therefore a signal being detected at angle θ, is given by Bragg's Law, which then relates the diffracted spot to the lattice parameters:

$$n\lambda = 2d_{hkl}sin\theta$$

where λ is the wavelength of the X-rays used in the data collection and θ is the angle between the incident ray and the lattice planes mentioned previously, which is also equal to the reflection angle. d_{hkl} refers to the interplanar distance, of direction hkl. This means that for constructive interference each wave must travel exactly one wavelength further than the wave next to it. According to basic trigonometry, this requires that the angle of incidence of the x-ray relative to the plane must be an equal integer multiple of the wavelength.

During an x-ray diffraction experiment the wavelength is fixed, and for a set of planes within the crystal the interplanar distance is also fixed. As a result of this the crystal must be rotated to ensure conditions that satisfy Bragg's law for all reflections. Visualisation of conditions that allow detection of reflections can be done with an Ewald sphere, which is a reconstruction of Bragg's law in reciprocal space. The radius of this sphere is 1/λ from the crystal origin. The origin of the reciprocal lattice will lie on the incoming beam. A reflection from a particular plane will satisfy the conditions of Bragg's law when the lattice point lies on the surface of this sphere. By rotating the crystal in real space, the reciprocal lattice also rotates, bringing new reflections into diffracting conditions.

Physical parameters such as radiation damage can affect the intensity of the reflection. Symmetry related reflections should have identical intensities. To ensure this the data is all scaled, combining reflections from different images within the dataset and applying a scaling factor, and then merged into one dataset for further analysis.

The direction of the planes described above can be described by the Miller indices, which are written as h, k, l values. These values relate to where on the x, y and z axes the plane intercepts, an example of this can be seen in figure 2.5:

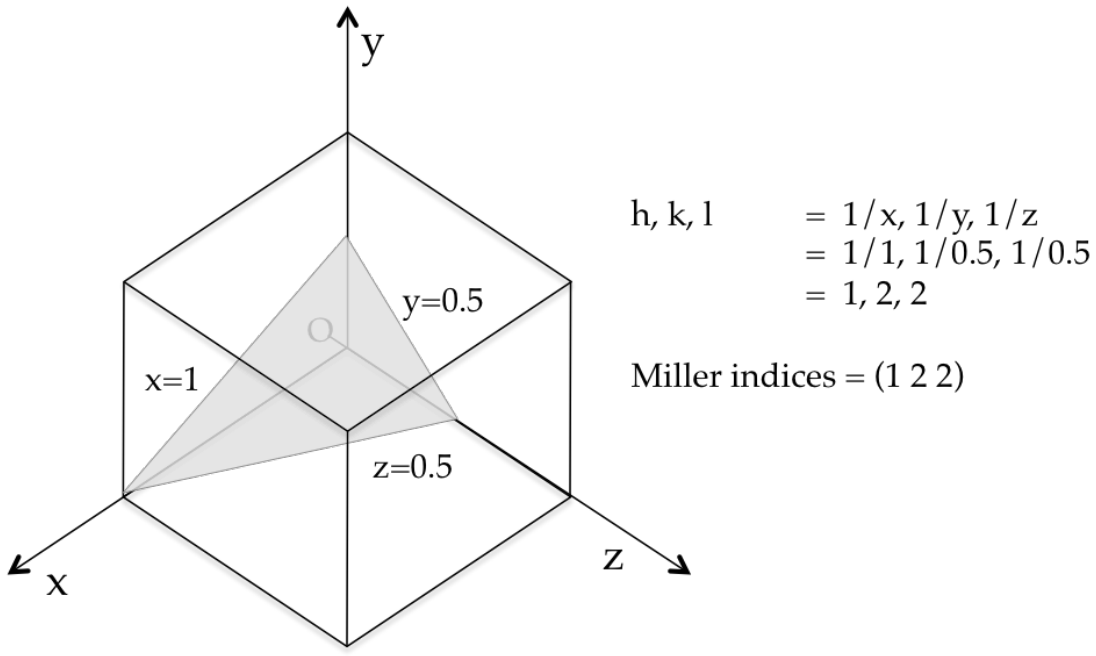


Figure 2.5 - An illustration of a plane within an orthorhombic unit cell, and how its direction can be described by the miller indices

Once the h, k, l values have been determined and the plane is known, the distance between the planes (d_{hkl}) can be determined. An example of a calculation for a cubic system can be seen below:

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}}$$

2.2.8.1 – The Phase Problem

During an X-ray diffraction experiment, the data collected comprises the intensity of the reflections, whereas the phase information (α_{hkl}) is lost. X-rays behave as waves and therefore have both amplitude and phase. The amplitude is defined as the maximum extent

of the wave whereas the phase is a particular point in the cycle of a waveform. The structure factor amplitude ($|F_{hkl}|$), is proportional to the square root of the intensity information collected but the point in the wave cycle cannot be collected. This is what is termed the 'phase problem'

The relationship between these parameters can be seen in the equation below, where the ρ(xyz) relates to the electron density at point xyz in the unit cell and the unit cell volume is represented by V:

$$\rho(xyz) = 1/V.\Sigma\,|F(hkl)|\exp(i\alpha_{hkl})\exp(-2\pi i(hx+ky+lz))$$

Of the methods used to solve the phase problem, only one will be discussed in detail in this thesis, molecular replacement (MR). This was the sole method used to solve the phase problem in this research due to an abundance of suitable models to use. Other methods used include isomorphous replacement, anomalous dispersion and molecular replacement, all of which are briefly discussed in the appendices.

MR is performed computationally, and usually requires a single, complete dataset. In order to perform MR, a PDB file is required that contains the atomic coordinates of an ensemble that resembles the target protein.

In MR one of the first variables to determine is the solvent content of the crystal. The Matthew's Coefficient is a means of determining the solvent content of a crystal and it is defined as the crystal volume *per* unit of protein molecular weight (V_M). By studying 116 crystal forms in 1968, and 226 in 1976, Matthews determined the range of solvent contents and determined a relationship between the solvent content and number of molecules in the AU (103, 104). This value can be calculated after determination of the space group and unit cell dimensions, in conjunction with the molecular weight of the protein. Once a value is obtained, an estimate can be performed, based on probabilities, to determine the number of molecules of the protein within the asymmetric unit. This information is required when

performing MR. The principle behind MR is that a Patterson map is generated for the search model that has been taken from an existing set of coordinates. This search model has to be rotated in three rotation angles, α, β, γ as well as three translation functions, t_x, t_y, t_z. By performing the rotations first to determine the suitable angles, and only performing translations on fixed angles determined from the best results from the rotation function dramatically improves computational efficiency. Both the rotation and translation functions are more efficient in reciprocal space, therefore structure factors are calculated for the search model, where the associated phase with every reflection is set to 0. This is known as the Patterson function, the equations for which is shown below:

$$\rho(uvw) = \int_x \int_y \int_z \rho(uvw)\rho(\mathrm{hx} + \mathrm{ky} + \mathrm{lz})\, dx\, dy\, dz$$

$$\rho(uvw) = \frac{1}{V} \sum_h \sum_k \sum_l |F(hkl)|^2 \, (hx + ky + lz)$$

A peak at 'uvw' in the Patterson map will correspond to a vector between atoms 'xyz' and 'x+u', 'y+v', 'z+w' in real space. The N atoms in a molecule will produce N^2 peaks in the Patterson function, with a single peak at the origin corresponding to a vector between an atom with itself, i.e. with a vector of zero length. The search model Patterson function is compared to the experimental Patterson function, where successful orientations can be identified by overlapping peaks.

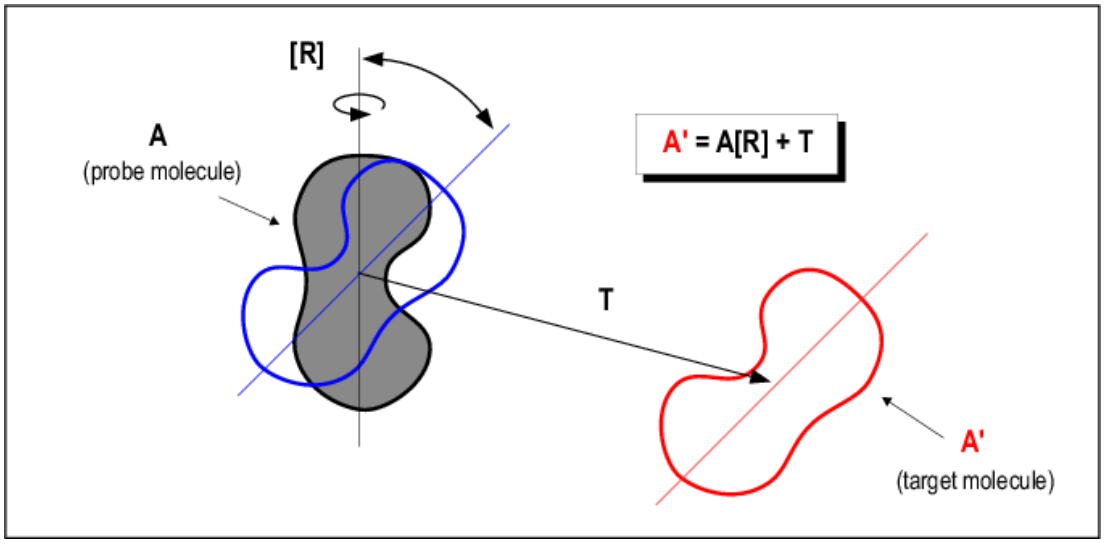


Figure 2.6 - Cartoon depicting the process of molecular replacement, where the search molecule (A), is first rotated to a position where theoretical and experimental values correlate. At which point the molecule is translated to the target molecule (A') (105)

Once these phases have been calculated and combined with the amplitude data generated during the data collection this will produce an approximate Fourier synthesis of the structure of interest that will then begin the necessary rounds of model building and refinement. The success of the MR can only be confirmed by analysis of the newly calculated phases by inspection of the electron density maps. That being said, there are two score functions that can be used as strong indicators for success, where a score is significantly higher than background, this is likely to be a unique, correct solution.

The Log Likelihood Gain (LLG) score is a probability function that compares the likelihood of obtaining an input (experimental data) based on the outcome (the model, and its structure factor determined by its position) to an outcome based on a random distribution of atoms based on the Wilson plot.

The Z-Score is a statistical test that measures how many standard deviations the LLG score is above or below the mean. A translational Z-Score above 5 is usually indicative of a correct solution (106), however rotational Z-Score is often poorer as the error is larger and several orientations are tested for each translational search.

2.2.9 – Data Collection

2.2.9.1 – High Pressure Data Collection

All data collection performed at the European Synchrotron Radiation Facility (ESRF, Grenoble, France) was at room temperature, using a Diamond Anvil Cell (DAC) (107). A schematic and picture of the DAC used can be found in Figure 4.32. The Detector used was a Mar555 (Mar Research) and all analysis was performed using XDS (108) and CCP4 (109, 110). Samples were transported *via* 1 mm capillaries with precipitant surrounding the crystal. The capillaries were sealed using wax and housed within a refrigerated unit maintaining a constant 18ºC temperature until arriving at the beam-line. The DAC was then cleaned and loaded with a small ruby, which was used to accurately determine the pressure within the cell during the experiment. The crystal was removed from the capillary directly onto a glass slide, which could be easily accessed and viewed under a light microscope. This crystal was picked using a loop and, simultaneously, 2 µL of precipitant was added to the sample chamber. The crystal was then dropped into the sample chamber and the DAC was moved under the microscope, at which point it had to be manipulated from the surrounding area to be successfully oriented within the chamber. Any bubbles present within the chamber were removed at this point using a fine needle. The cell was then closed, sealed and loaded onto the beam-line.

The DAC was then aligned to the beam centre and further positioned so that the crystal was in the X-ray beam for data collection. An initial crystal was used for a compressibility experiment, to determine the pressure at which the crystal no longer diffracts X-rays. Once this final value was experimentally determined, a pressure at approximately 50% the maximum was applied and used throughout the remaining data collection.

Pressure was measured by firing a laser beam at a ruby within the DAC as the pressure was increased. Using a calibration curve first postulated by Mao *et al* (111) a

constant pressure was achieved, monitored and regulated throughout the experiment. This calibration curve was constructed because of an observable upward shift in the wavelength of the fluorescence emitted from a ruby present within the DAC that correlates with an increase in the pressure.

Due to the position in which the DAC sits in relation to the beam, there was only a 90º rotation possible from -45º to +45º around the vertical axis. This meant that to collect a complete data set in order to produce a structure, more crystals had to be used, and they were placed within the sample chamber in as many varied positions as possible (112).

The concept and data processing of high pressure X-ray diffraction are practically identical to that of ambient pressure X-ray crystallography. The differences lie in the actual data collection and limitations that result from this. Firstly the data is collected at room temperature. This is done due to the fact that temperature and pressure are proportional, as shown by the following equation:

$$PV = kNT$$

Where P = Pressure, V = volume, N = number of molecules, T = temperature and k=Boltzmann constant. This formula states that with an increase in pressure, there is an associative increase in temperature, until equilibrium is reached. Therefore it is considerably more difficult to increase the pressure and maintain it, whilst simultaneously lowering the temperature to cryogenic levels rather than collecting at ambient temperature, although a lower temperature will greatly increase the lifetime of the crystal, as radiation damage from free radicals is greatly reduced.

2.2.9.2 – Data Collection at the Diamond Light Source.

All data collection performed at the Diamond Light Source Synchrotron, Oxfordshire, was done so on cryo-cooled crystals. The crystals were picked from their drop and if the

polyethylene glycol (PEG) concentration was high enough then the loop containing the crystal was submerged in liquid N_2. When the PEG concentration was lower in the precipitant, the loop containing the crystal was first passed through a solution of cryoprotectant to reduce the chances of damage to the crystal or the presence of ice-rings in the diffraction pattern.

The loops, once frozen, were then placed into a 'puc' and transported *via* a dewar containing liquid N_2 directly to the beam-line. Data collection was either performed in person or *via* remote access. Data processing was performed with MOSFLM (113) and CCP4 (109).

Chapter 3

DNA Dependent Protein Kinase Catalytic Sub unit

3.1 Introduction

DNA-PKcs is a protein consisting of 4128 amino acids that associates with the Ku 70/80 heterodimer to form the active holoenzyme DNA-PK, which plays a vital role in the repair of DNA double strand breaks (DSBs) through the non-homologous end joining pathway. This chapter will discuss preliminary work aimed at producing soluble domains of the catalytic domain of DNA-PKcs. By obtaining soluble domains of DNA-PKcs, structural information could be determined experimentally, provided that the protein is properly folded and can be crystallised, allowing for x-ray diffraction and data collection. With key knowledge such as this, inhibitors with greater potency and specificity could be synthesised to adjunct existing chemo- and radio- therapies. Soluble domains could also be used to develop assays to greater elucidate the function of DNA-PKcs within the NHEJ as well as a more detailed picture of its interactions and interacting partners.

3.2 Protein Production

3.2.1 – Construct Design

Beginning at the N terminus and continuing through the entire sequence is a large portion of HEAT repeats. These repeats are overlaid with several other domains throughout the sequence.

A series of constructs surrounding the kinase domain were designed, seen in figure 3.1. This construct design approach was utilised in order to determine the minimum sequence required to produce functional, correctly folded, soluble protein for the purposes of crystallisation and crystallography. The largest construct, A1, starts at P3600 and continues through to M4128.

In reference to the domain map in figure 3.1, this construct begins 60 amino acids downstream of the FAT domain and continues through to the C terminus of the protein. Highlighted in red in figure 3.1 is the minimum kinase domain as predicted by the software 'Phyre' (114) and highlights the region from H3716 through to P4008. Phyre works by comparing a query sequence to known structures deposited in the protein databank (PDB), with a comparative ranking based on the confidence allowing accurate prediction of domain boundaries deduced from known similar structures. From S4096 through to M4128 is the FATC domain. Between M3796 and L3829 is a portion of the sequence that is present in many human cell lines, but not in HeLa cells due to the exon on the cDNA being absent (115). This 'splice variant' region has been removed in the E and F series of constructs to determine its effects, if any, on stability of the protein as well as its role in folding into secondary and tertiary structure as currently it has unknown function.

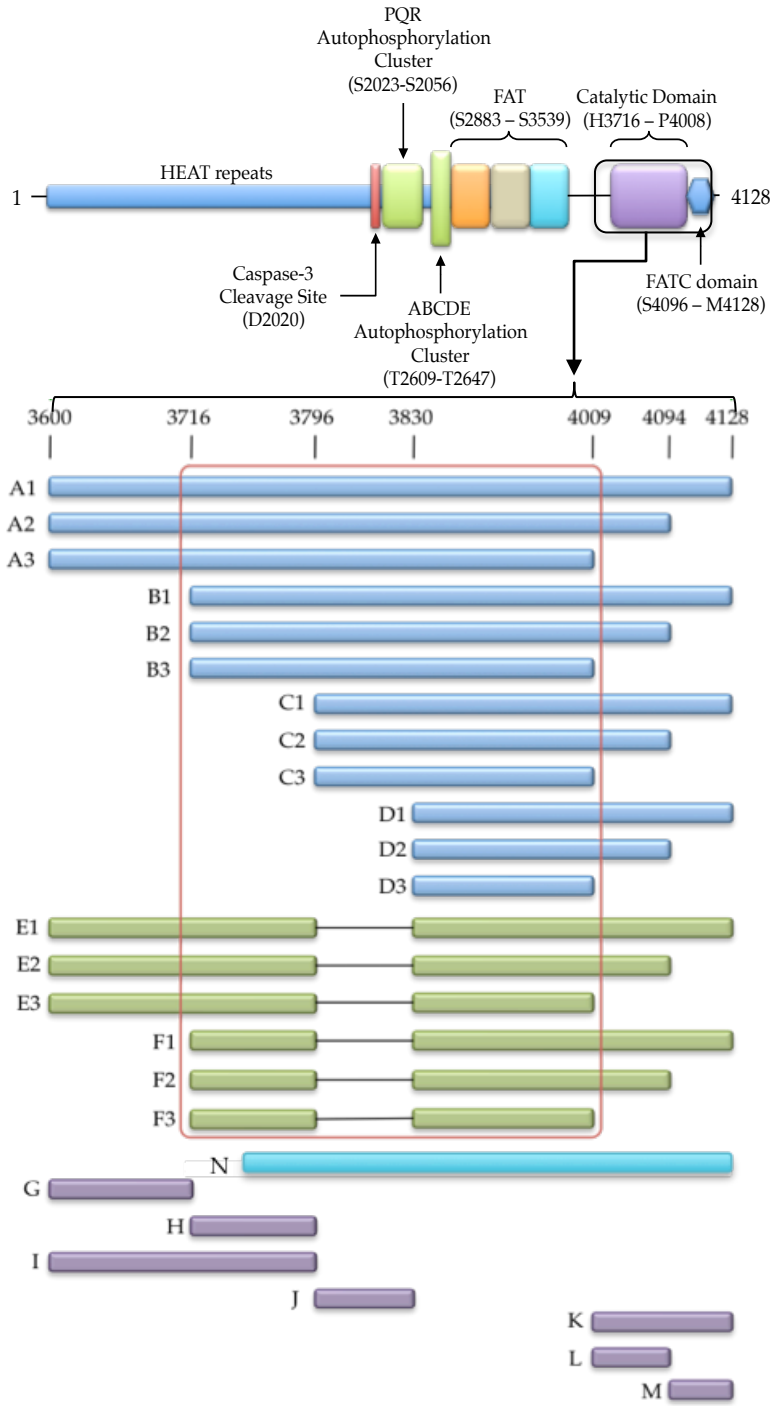


Figure 3.1- - Construct map showing the constructs designed and used throughout this study with reference to figure 1.3 indicating what regions of DNA-PKcs these constructs refer to. Cloning of constructs A1-D3 was performed by Hanaë Gourier (HG), while E1-F3 were cloned by the author using A1 as a template. Highlighted in red is the minimum catalytic domain determined by Phyre (114). Construct N was purchased through GeneArt based on domain boundaries published by Sajish *et al* (115). Constructs G-M have been included in the figure to indicate that these regions were initially tested in the interests of being thorough, however they did not progress past initial solubility studies.

3.2.1.1 – E & F Series Production

Both the E and F series of constructs were built by initially using the A1 construct as a template for the production of E1. A1 itself was designed and constructed by a previous member of the Spagnolo Lab, Hanae Gourier (HG).

The first step in generating these constructs was the production of the two fragments either side of the removed splice variant portion. This was performed with 4 different primers (figure 3.2 and table 3.1) supplied by Sigma Aldrich.

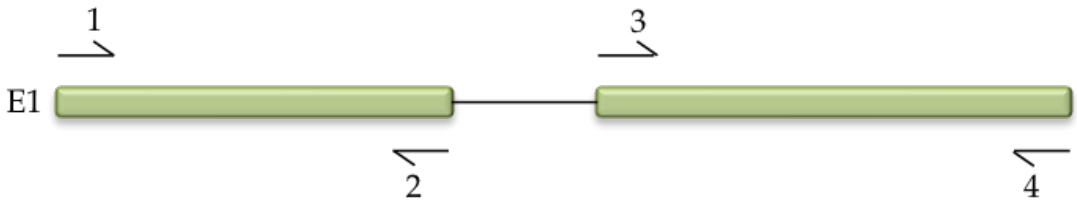


Figure 3.2 – A schematic of the design of the primers used in the production of the E & F series.

Primer No.	Primer Name	Sequence
1	Fw3600 (NdeI)	5' – GTGTGTCATATGCCTGTAAATAA AAAAAACATTGAAAAATGTATGAA - 3'
2	Rv3796 (-)	5' – GGAGGTCATGGGCACAACGCTATAGG – 3'
3	Fw3830 (-)	5' – AGTGATCCCAGGGCACCGCC – 3'
4	Rv4128 (NotI)	5' – AAAAAGCGGCCGCTTACATCCAGGG – 3'

Table 3.1 - A table showing the 4 primers used in this study to produce the E and F series of constructs. Primers 1 and 4 both contain restriction sites engineered into the primer, the identity of which is in parentheses. A dash in parentheses indicates no restriction site being used, for the purposes of blunt end ligation.

As mentioned previously, A1 was used as a template for the PCR reactions with *Pwo* polymerase to amplify both sides of the construct. Digestions on the 3600 and 4128 ends of the newly formed construct were performed with NdeI and NotI (New England Biolabs) respectively. The two blunt ends of the construct were ligated with T4 Ligase (Roche). This ligation then produced a single fragment that was subsequently ligated into the pTWO-E vector discussed in chapter 2. Making use of the different restriction sites allows for correct orientation within the vector. The product of this ligation was then used directly to transform chemically competent XL1-Blue cells. A colony PCR was then performed to test the successful introduction of the sequence into the vector. The positive results of which were confirmed by sequencing. The remaining E and F series of constructs were then produced in a standard manner using E1 as the template rather than A1.

3.2.1.2 – Overexpression Trials

Initially, all constructs were first transformed into competent BL21* cells. Any constructs that didn't produce viable cells after transformation were then discarded and only successful constructs were taken forward.

A starter culture was produced for each of the successful constructs and an experimental culture was produced. An induction time course was performed to determine the optimum induction period prior to harvest. If any of the protein products were toxic to the cells, this would be observed by seeing a drop in the OD_{600} post induction. The final constructs that were taken forward after these initial experiments were: A1, A3, B1, B3, C2, D1, D2, D3 and E1. Although this is a high rate of decline, it was deemed that there were suitable levels of non-toxic proteins being taken through to solubility studies to warrant time being spent optimising this step in the protocol. An interesting observation to note is that the three most successful constructs taken forward were A1, E1 and D1. The likelihood is that the three of these constructs were so successful due to where in the protein they began and ended, which was at designated regions throughout the construct that are *bona fide* domain

boundaries, rather than within domains, such is the case for most of the remaining constructs in the map in figure 3.1.

After this list of successful constructs was determined, further overexpression and solubility studies were performed, the results of which can be seen in figure 3.4. Based on the results seen in the figure, the A1, D1, E1 constructs were taken forward, as they showed an increased level of expression post-induction when compared to the other constructs tested.

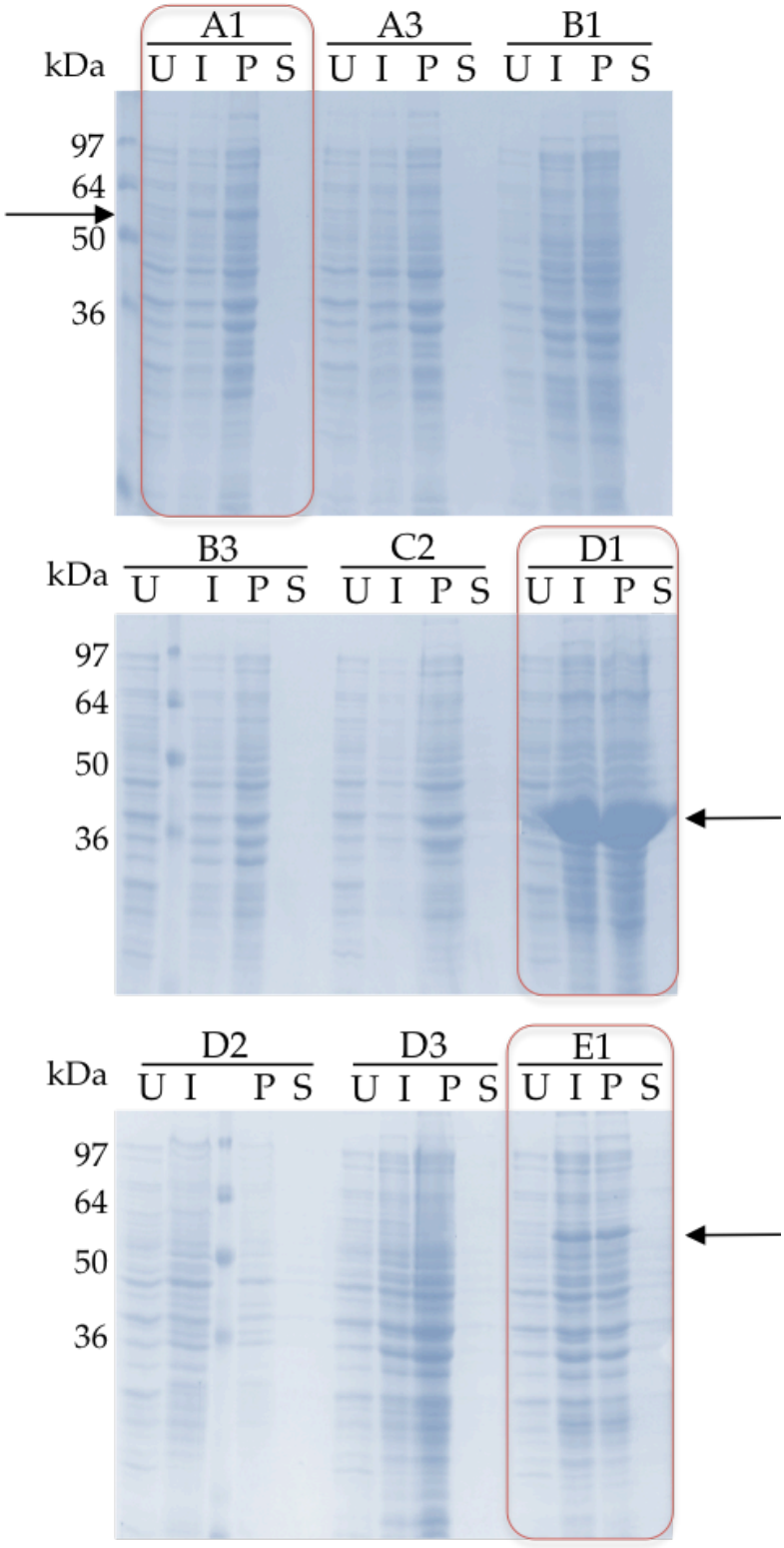


Figure 3.3 – 12 % SDS-PAGE gels showing the results of growth trials on the positive DNA-PKcs constructs. The arrows indicate bands present in the most successful constructs, A1, D1 and E1. U – Uninduced, I – Induced, P – Pellet, S – Supernatant.

The results in this figure indicate that there were varying levels of expression for these three constructs, however when using sonication as a means of cell lysis, the proteins were in fact insoluble. This could have been due to incomplete cell lysis, or possibly heat denaturation. To test these hypotheses, the lysis protocol was optimised.

The first of these variations was to use lysozyme rather than sonication. This treatment is a lot milder than sonication and there are no heat associated denaturation issues. This however had no beneficial effect on increasing the solubility of the proteins. Alternative overexpression approaches were then tested. The first of these approaches was a decrease in the induction temperature, from 37 °C to 18 °C. After several attempts with all three constructs, all that was seen was a decrease in overall yield within the whole cell extract, but the relative amount of soluble protein remained at previous levels.

Another variable tested was the final concentration of IPTG used to induce protein production. Initial experiments used a final concentration of 1 mM IPTG; this was reduced to final concentrations of both 0.2 mM and 0.1 mM. After testing this on all three constructs, no beneficial effects were seen in relation to solubility. This was then performed in combination with the reduced induction temperature mentioned previously, however this also failed to increase protein solubility.

Lysis buffer optimisation was then performed to attempt to increase yields of soluble protein. 6 different buffers were tested. Firstly the buffering compound itself was changed, using either HEPES or Tris. The concentration of NaCl was also varied, in a range from 20 mM to 500 mM. The pH was varied within the suitable buffering ranges of both the HEPES and Tris. Lastly the addition of 80 nM Dodecyl Maltoside (DM) as a detergent was tested.

The buffer changes mentioned above, both on their own and in combination with one another did very little to aid with the solubility of the constructs. It was at this point that alternative constructs were designed and implemented, with the aim of generating soluble DNA-PKcs.

3.2.1.3 – Codon Optimisation

The advantage of using *E. coli* as an expression system is the fact that it is cheap, quick and relatively straightforward to produce a protein recombinantly. A lot of commercially available vectors are also optimised for expression in *E. coli.* However, the frequency that the codons are translated in the ribosome by complementary tRNAs varies significantly between different organisms. This disparity in tRNA levels leads to a positive selective effect on constructs that result in faster and more accurate translation, thus leading to the aforementioned codon usage bias (116). A recombinant human DNA sequence will contain a degree of codons that are rarely used in *E. coli* and as such the translation levels will be sub optimal. Issues can then arise during protein synthesis. These issues can include: interrupted translation; frame shifting; mistranslational events; as well as inhibition of cell growth and protein synthesis (117).

One way to overcome this issue is to generate a 'codon optimised', synthetic construct. The rare codons are replaced with ones that are used in abundance in *E. coli* (118).

Codon optimised constructs for A1 and E1 were designed and purchased from Geneart, using their online software 'GeneOptimizer®' to modify codon usage from *Homo sapiens* to *E.coli* (119). These codon-optimised constructs were subsequently tested for overexpression and solubility. Different media were tested, using LB, TB and Auto-induction media. The cell strain was also varied, testing BL21*, BL21-AI (Arabinose Inducible), and Rosetta pLysS cell strains. Then a variety of N terminal tags were also tested, including untagged, 6xHis tag, GST tag and a SUMO tag. The tags were incorporated into the construct through cloning into the pTWO-E vector, shown in figure 2.1. SUMO (Small Ubiquitin like Modifier) is a small protein that is covalently linked to the target protein that behaves akin to a post-translational modification. It is also believed that the SUMO tag prevents degradation of the protein by translocating it to an intracellular region low in

proteases (120, 121). GST tags have the potential to protect against intracellular protease activity and stabilise recombinant proteins.(122)

Transformation of the cells was performed and an induction trial was performed. The results were viewed on several western blots, which can be seen in figure 3.5 and these results have been summarised into tabulated form, seen in table 3.2. The quality, measured by the levels of degradation products present in the sample, and quantity of protein, based on the expression levels quantified by both SDS-PAGE and the absorbance at 280 nm, were given a score and the better the conditions, the more ticks it was awarded.

This experiment helped to remove LB as a growth medium, as well as both untagged versions of the constructs, which didn't actually grow, a result which was seen after several repeated experiments. TB media was consistently better than any other media tested and was therefore used throughout the remaining experiments. Possible reasons for its superiority over other media could be due to the fact it is phosphate buffered preventing cell death due to pH changes, as well as the fact that it can reach much greater cell densities. Rosetta pLysS was chosen due to having consistent overexpression levels when compared to the other cell strains tested.

		Terrific Broth			Auto Induction medium			*Luria Bertani* medium		
		BL21*	BL21-AI	Rosetta pLysS	BL21*	BL21-AI	Rosetta pLysS	BL21*	BL21-AI	Rosetta pLysS
A1	No Tag	NO GROWTH								
	His	✔	✔ ✔	✔ ✔ ✔	✔	✔	✔	✗	✗	✗
	GST	✔ ✔ ✔	✔ ✔ ✔	✔ ✔ ✔	✔	✔	✔	✗	✗	✗
	SUMO	✔ ✔	✔ ✔	✔ ✔ ✔	✔	✔	✔	✗	✗	✗
E1	No Tag	NO GROWTH								
	His	NO GROWTH								
	GST	✔ ✔	✔ ✔	✔ ✔ ✔	✔	✔	✔	✗	✗	✗
	SUMO	✔ ✔	✔ ✔	✔ ✔	✔	✔	✔	✗	✗	✗

Table 3.2 - A list of the various conditions tested with the codon optimised constructs. Final conditions used A1 and E1 with a GST tag in Rosetta pLysS cells.

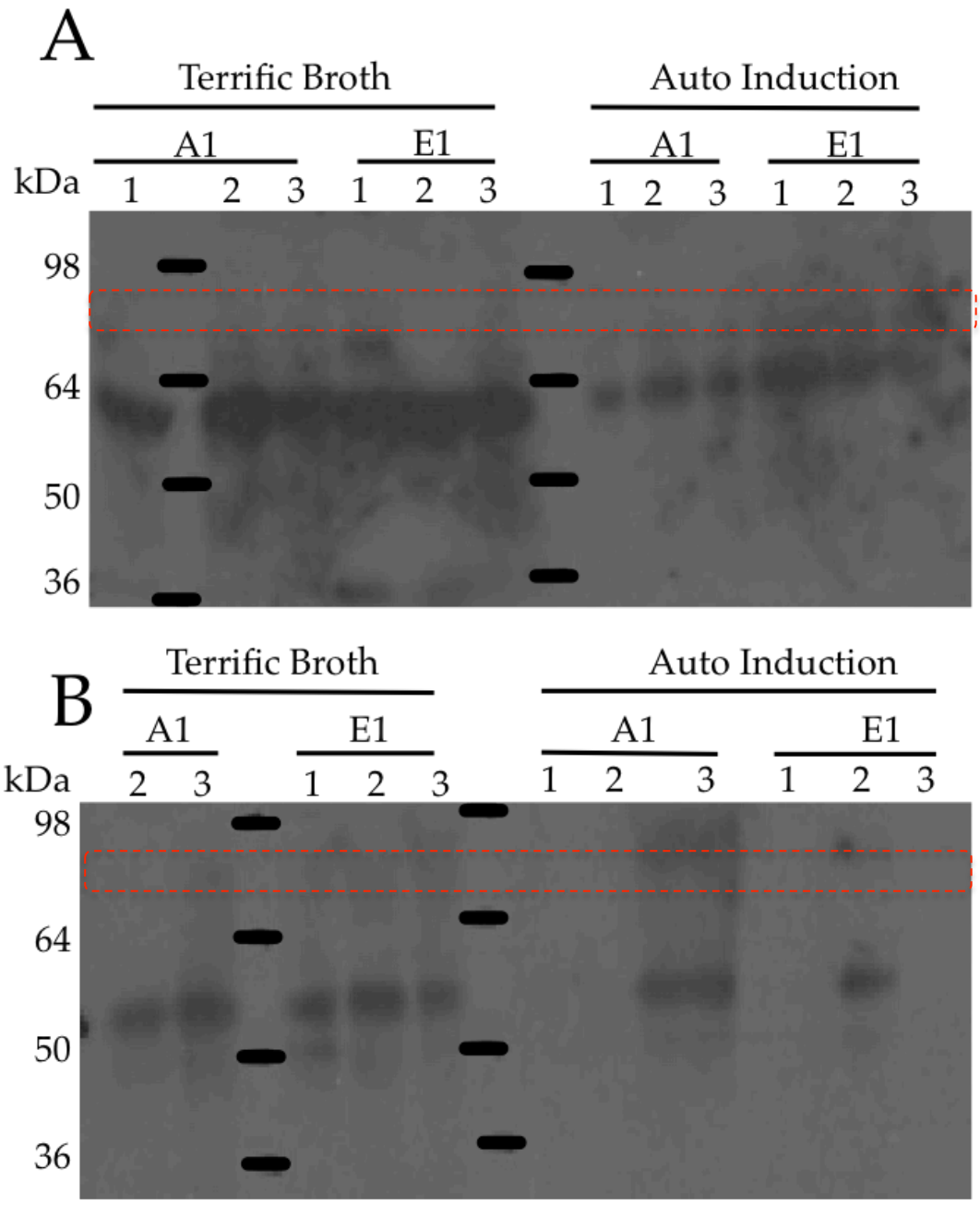


Figure 3.4 - (A) Western blot developed with α-DNA-PKcs antibody showing the results of over-expression tests using GST-A1. Condition 1 used the BL21* cell strain, condition 2 used the BL21-AI (Arabinose Inducible) cell strain, and condition 3 used the Rosetta pLysS cell strain. (B) Western blot developed with α-DNA-PKcs antibody showing the over-expression results for His-A1. LB broth was not included due to already being determined an unsuitable medium. All blots in both (A) and (B) were developed on the same film, and incubated with the membrane for 60 seconds. Bands in both (A) and (B) are both considerably lower (around 55 kDa) than the 85.9 kDa of GST-A1 highlighted with the dashed red box. .

3.3 Protein Purification

3.3.1 – Soluble GST-A1

After determining the optimal growth conditions, both the GST-A1 construct and His-A1 constructs were tested alongside one another but due to both the His-A1 and SUMO-A1 being His-tagged they were grouped and written about in the section proceeding this. To test the solubility status of the target protein, the cell pellet was lysed with lysozyme, in 20 mM HEPES pH 7.5, 150 mM NaCl, 10 % Glycerol, 5 mM $MgCl_2$, then subjected to further optimisation. The results of this lysis can be seen in figure 3.5.

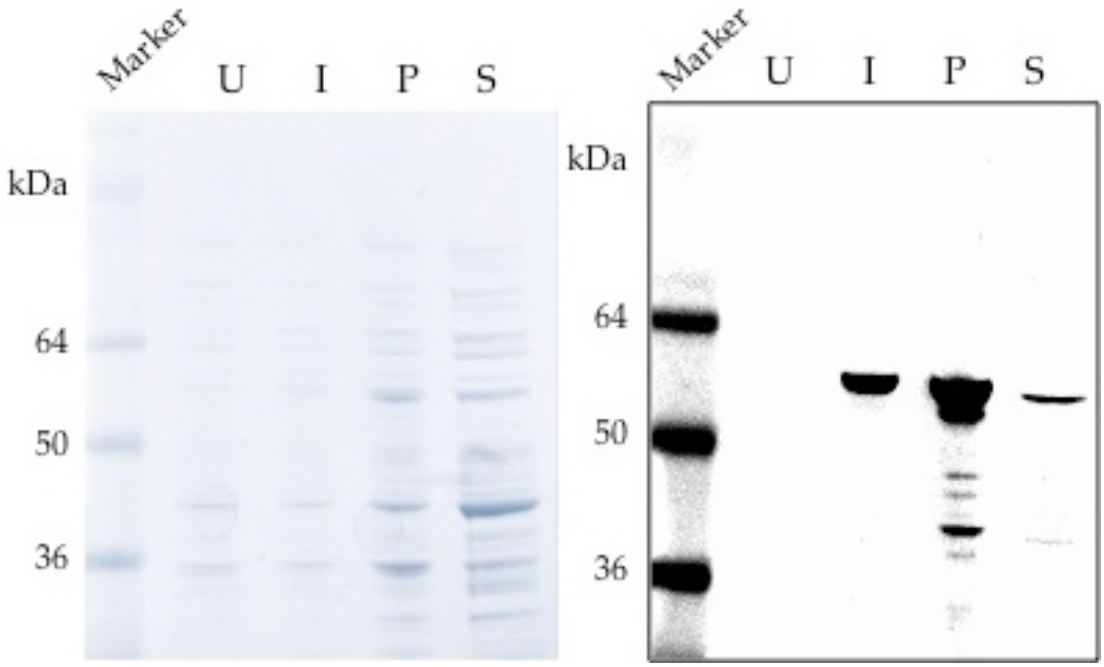


Figure 3.5 – 15% coomassie stained SDS-PAGE and western blot, developed with a-DNA-PKcs antibodies showing the uninduced (U), induced, (I), pellet (P) and supernatant (S) samples for GST-A1. Lysis was performed with lysozyme and incubated with rotation for 60 minutes at room temperature before clarification *via* centrifugation at 22,000 rpm for 60 minutes.

Figure 3.5 highlights the levels of expression observed for GST-A1. The SDS-PAGE shows that expression levels are low but the western blot does indicate a protein containing the epitope for an anti-DNA-PKcs antibody is being produced. The bands in this western

blot are resolving at approximately 55-60kDa, whereas the theoretical size for GST-A1 being 85.9 kDa. This could be as a result of non-specific antibody binding. Other reasons could include abnormally fast migration of bands through the SDS-PAGE gel or even an untagged construct, as this is close to the size of the band being observed (60.4 kDa)

In figure 3.6, the upper western blot (developed with α-DNA-PKcs antibody) is for the GST-A1 construct and the lower western blot (again developed with α-DNA-PKcs antibody) is the E1-GST construct. The one condition that had a good yield of soluble protein for both of the constructs was PIPES at pH 7. One interesting observation is the disparity between the band migration for both the A1 and E1 constructs. The GST-A1 construct seems to be present higher on the gel than it should be, greater than 98 kDa, even though the size of the protein is 85.9 kDa, whereas for GST-E1 the band is present ~64 kDa when its true size is 82.3 kDa. This disparity can also be seen when comparing the results in figure 3.7 with those in figure 3.6, which also has bands present ~64 kDa. One reason for this could be degradation of the tag from lysis to analysis, however this is unlikely due to the lysis and preparation techniques being consistent. Another possible explanation could be cross-reaction of the antibody against DNA-PKcs used for these western blots, although the epitope that these antibodies target is small and no *E.coli* proteins have sequence similarity to this region.

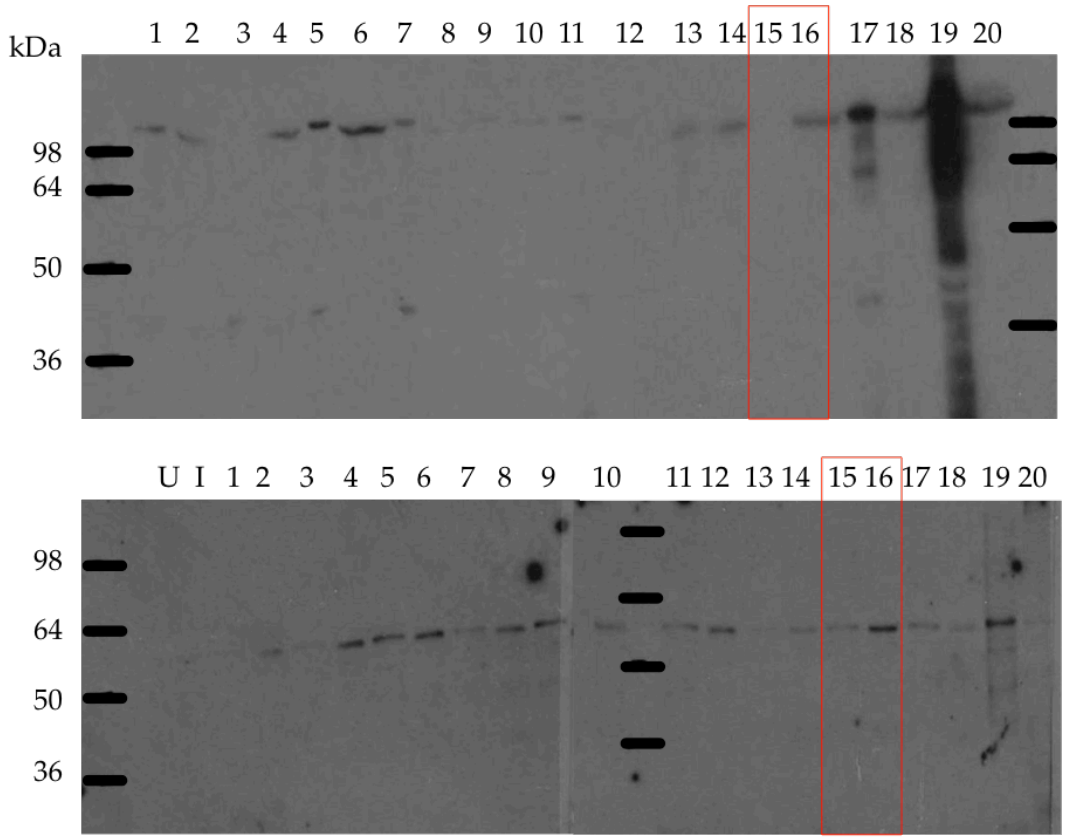


Figure 3.6 – Western blot developed with α-DNA-PKcs antibody showing the results of a buffer scouting experiment performed to increase solubility of GST-A1 (upper) and E1-GST (lower). U – uninduced, I – induced, odd numbers indicate a pellet fraction after lysis and even numbers represent the supernatant from that same lysis experiment. Highlighted in red indicate a pellet/supernatant pair for lysis performed in PIPES pH 7. Lane identities can be found in table 3.3.

Lane	Sample	Buffer	pH
1	Pellet	HEPES	7
2	Super		
3	Pellet		7.5
4	Super		
5	Pellet		8
6	Super		
7	Pellet	TRIS	7
8	Super		
9	Pellet		8
10	Super		
11	Pellet		9
12	Super		
13	Pellet	MES	6
14	Super		
15	Pellet	PIPES	7
16	Super		
17	Pellet	CAPS	10
18	Super		
19	Pellet		11
20	Super		

Table 3.3 - A table showing the identity of each of the lanes present in the western blots shown in figure 3.6. Highlighted in red is the condition that was optimal for both constructs, which was then taken forward.

Therefore PIPES at pH 7 was used for all subsequent experiments. A binding affinity assay was performed to determine the best resin to use for the purification of the construct. All subsequent results discussed in this chapter are for the GST-A1 construct.

3.3.2 – GST-A1 Protein Purification

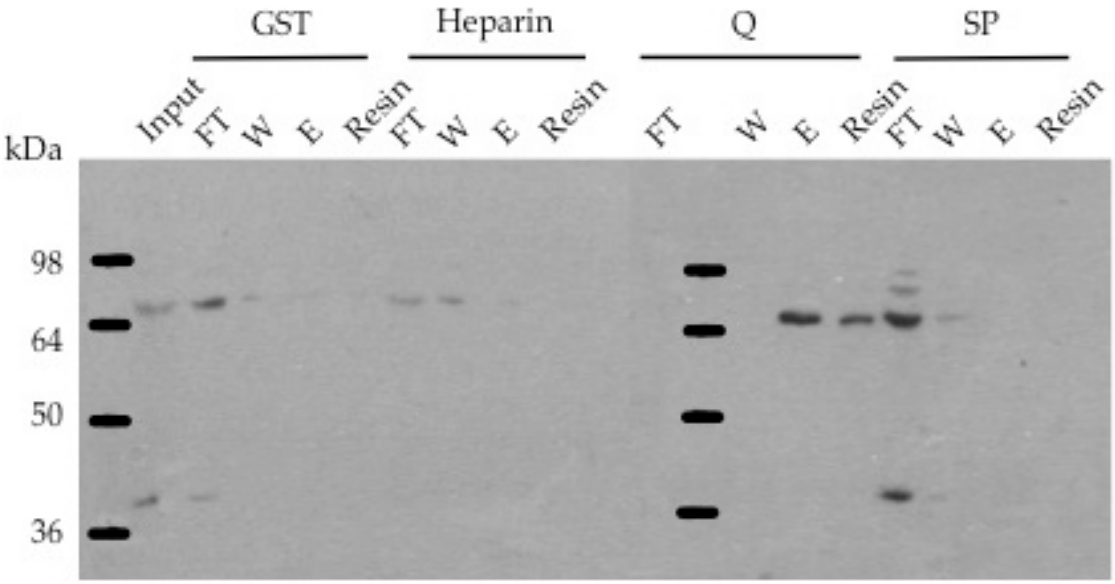


Figure 3.7 - Western blot, developed with α-DNA-PKcs antibody, showing the results of the binding affinity experiment with GST-A1. FT – flowthrough, W – wash, E – elution.

500 µL of supernatant in PIPES pH 7 was added to 100 µL of pre-equilibrated resin, and allowed to incubate for one hour at room temperature, at which point the resin was washed and bound proteins were then eluted with the appropriate eluent for that particular resin. The samples were then loaded onto an SDS-PAGE gel for the purposes of a western blot. This blot can be seen in figure 3.7.

Based on this experiment, the most effective resin to use was the strong anion exchange resin, Q, as it was the only sample with protein present in the eluate rather than the flow-through. As seen in the figure, the protein did not bind to its appropriate affinity resin. This could be due to the concentration of reduced glutathione being too high in the buffer used to equilibrate the column, as well as the sample buffer. However it is also migrating on the gel lower than expected, which strongly indicates a lack of the required N terminal GST tag. It could also indicate antibody cross-reactivity leading to the purification of an alternative, non-target protein. This point will be discussed further later in this chapter. Due to finding a resin that the protein bound to, the scale was then increased and subsequent

purifications were performed on a 1 mL Q Hitrap (GE Healthcare) column. The theoretical pI of GST-A1, as calculated by Protparam (123) is 8.77, therefore at pH 7 it should bind to an anion exchanger, concurring with the result observed.

The overexpression scale was then increased. The pellet harvested from 1 L of culture was resuspended in 30 mL of PIPES buffer. After loading and washing the 1 mL Q Hitrap column, the protein was eluted with a salt gradient, from 200 mM NaCl to 1 M NaCl over 40 column volumes. Results are shown in figure 3.9, the fractions that contained the peak highlighted were then analysed by SDS-PAGE and western blot, the results of which can also be seen in the figure.

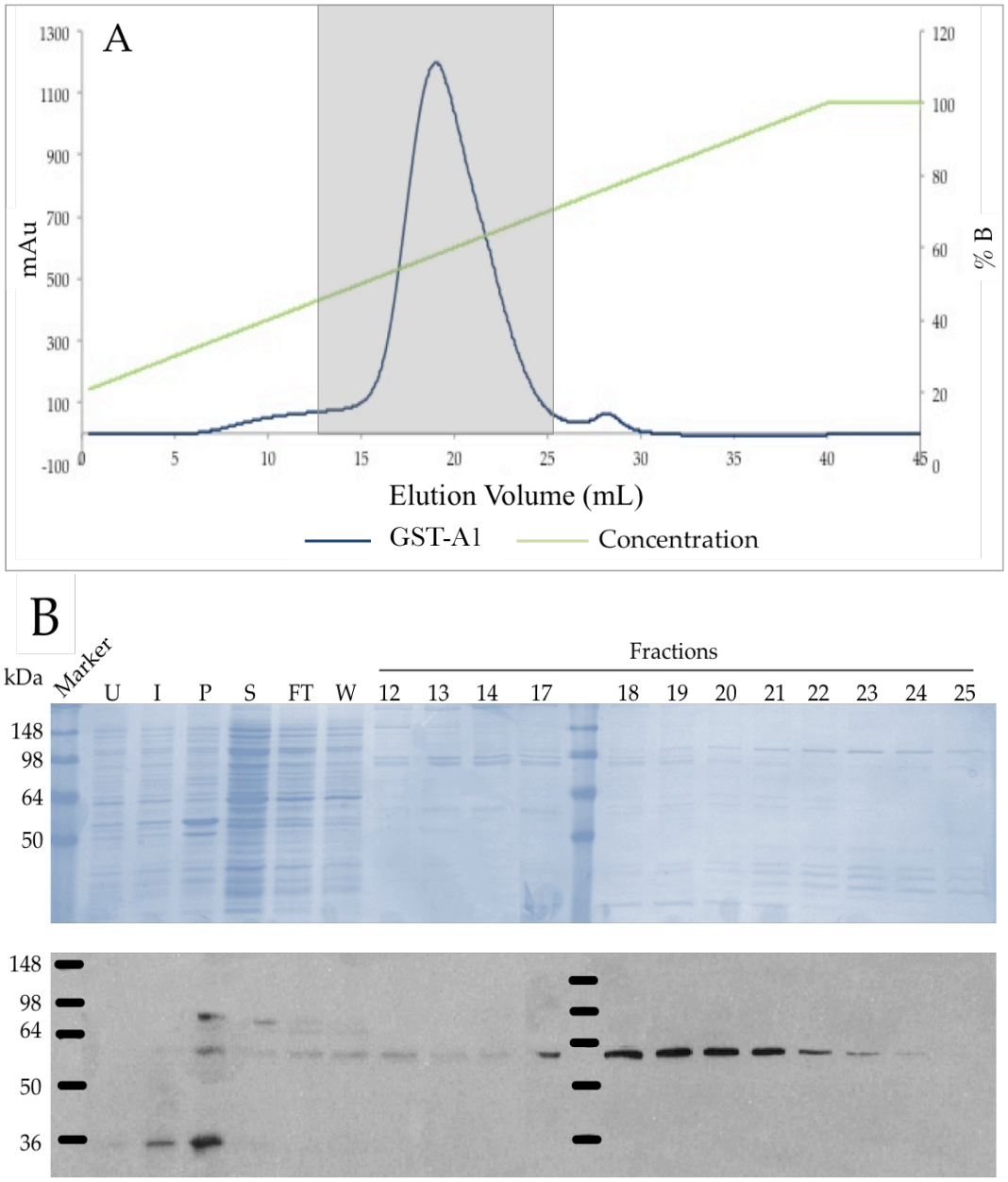


Figure 3.8 - (A) Elution trace from 1 mL Q column. This particular trace is showing only the elution portion of the purification, with elution being performed by a salt gradient. The highlighted grey area indicates the fractions collected (12-25). The green line in the graph indicates the concentration of elution buffer (B) on a percentage scale, where 100% equates to 1M NaCl. (B) 15 % SDS-PAGE coomassie stained gel and western blot developed with α-DNA-PKcs antibody showing these peak fractions.

Fractions 17 – 24 were collected, although the bands were running lower than they should be for the A1 construct with a GST tag. The expected size of the GST fusion protein

was 85.9 kDa. When the tag is removed however the size of the protein is 60.4 kDa, which was the approximate size of the band. This could be either due to tag removal during the purification process, proteolytic degradation or an alternative ribosomal binding site. Although unavailable at the time of experimentation, an anti-GST antibody would have been useful to detect presence/absence of the GST tag throughout the process from induction to lysis and purification. After collection, the fractions were pooled and concentrated to a final volume of 500 µL. This sample was then loaded onto an s200 10/300 GL size exclusion chromatography column (GE Healthcare). The results of this experiment can be seen in figure 3.10.

The blue trace in the figure represents the GST-A1 sample. As a reference, the red trace represents the calibration of the column with Blue Dextran, a 2 MDa branched polysaccharide that is too large to access any of the pores present on the Sepharose within the column. The volume that this is eluted from the column is termed the 'void volume'.

As mentioned previously the GST-A1 is around 85.9 kDa, which is not large enough to elute in the void volume. GST tags however are known to dimerise but this resultant ~172 kDa protein would still be small enough to pass through the column without being present in the void volume, as the largest protein size that can be accommodated by this column is 600 kDa therefore tag dimerisation causing void volume elution cannot be the cause.

The second GST-A1 peak shows no protein, based on the SDS-PAGE, and this is most likely free nucleic acid due to the 260nm / 280 nm ratio for samples 15, 16 and 17 being around 1.95, 1.98 and 1.99 respectively. A protein sample that contains no nucleic acid contamination will have a 260nm / 280nm ratio of 0.5.

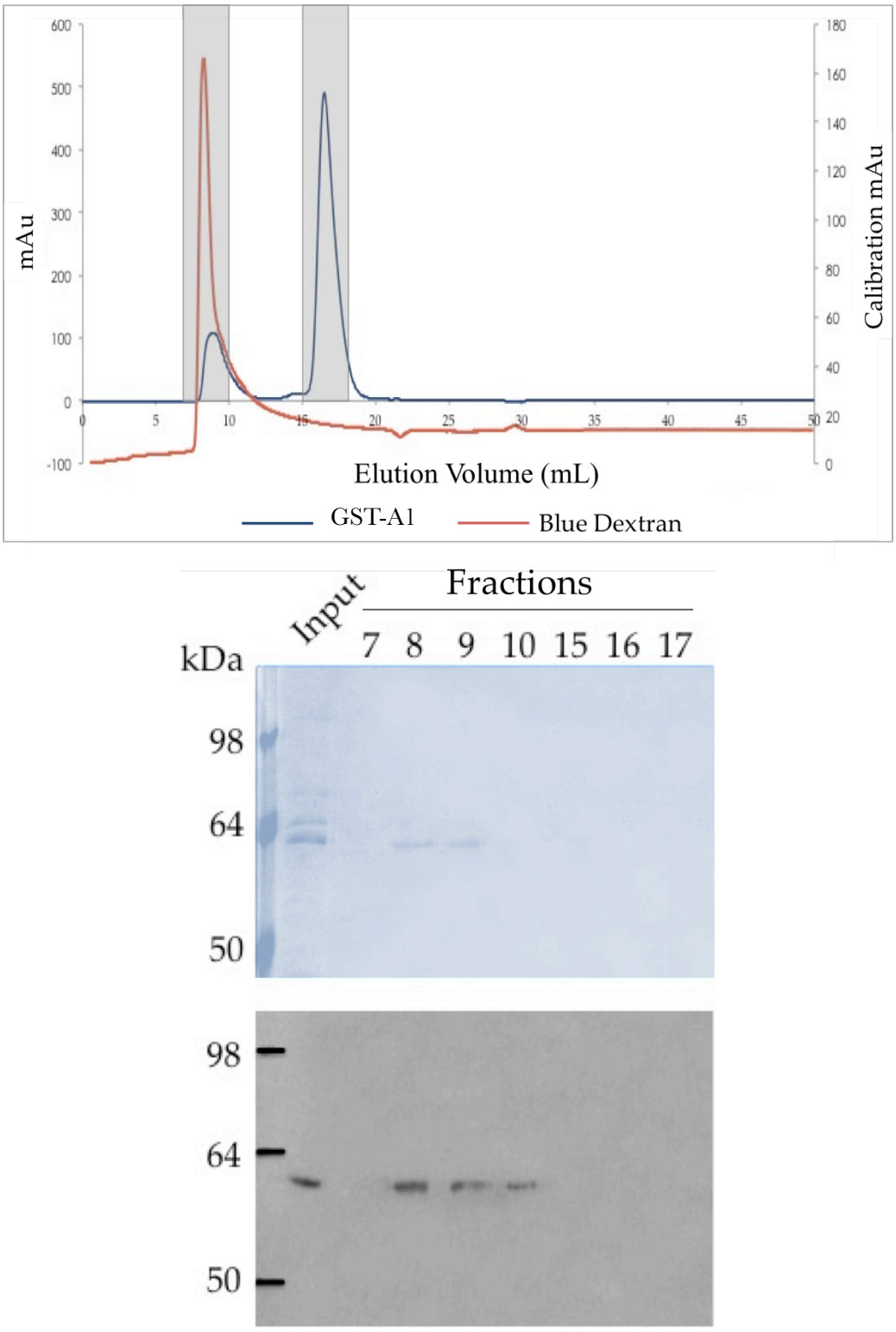


Figure 3.9 - Size exclusion chromatographic trace, using an S200 10/300 GL column (GE Healthcare) with a column volume of 24 mL. Below is the associated 15% coomassie stained SDS-PAGE and western blot developed with α-DNA-PKcs antibody for the purification of GST-A1. The grey areas indicate the fractions collected for analysis (7-10 & 15-17).

The 260 nm / 280 nm trace for fractions 8, 9 and 10 were all above 1.2, indicative of moderate levels of nucleic acid contamination. Therefore the formation of a protein / DNA complex could potentially be the reason for the protein being eluted in the void volume. One way to test if the tag dimerisation is in fact the reason for void volume elution was to perform a TEV cleavage assay, to remove the tag and then analyse the samples.

The TEV cleavage assay was performed in 50 mM Tris-HCl pH 8, 0.5 mM EDTA, 1mM DTT. 22 µL of the GST-A1, along with 1 µL of acTEV™ (Invitrogen), and 177 µL of reaction buffer were added into an eppendorf tube. The entirety of the reaction was performed at room temperature and after every hour 30 µL of the reaction mixture was removed and immediately denatured with SDS prior to running on a 12% SDS-PAGE gel. The results of this assay can be seen in figure 3.11.

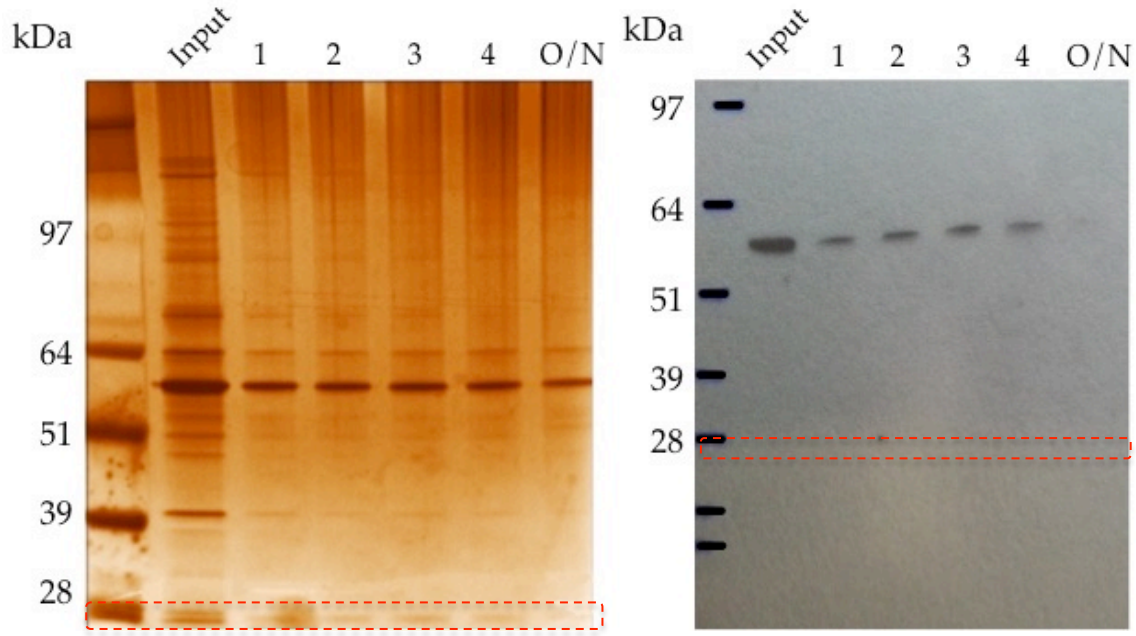


Figure 3.10 – Silver stained 15% SDS-PAGE and western blot developed with α-DNA-PKcs showing the TEV cleavage assay performed on GST-A1. The acTEV™ protease is 27 kDa and its migration region is highlighted in the figure ith the red dashed box (124).

Based on this figure, there has been no cleavage at the TEV site, which would normally be indicated by the band migrating faster due to a decrease in size, as well as the lack of appearance of a new band around 25.4 kDa. The experiment was varied several times in an attempt to improve the assay, however cleavage was never observed. No control has been included in this figure, however control experiments were performed, and indicated that the acTEV ™ was in fact a functioning protease. This further points to the possibility that the protein being taken through the purification does not contain either a GST tag or a TEV protease site, and is therefore not the target protein. The molecular weight of the protein, based on its migration on the gel is correct for the protein without the tag. One interesting point to note is the disappearance of the signal corresponding to the protein in the overnight (ON) sample, compared with the 4 hour sample. This could be due to the protein being kept at room temperature overnight, leading to subsequent degradation. Several repeats of this experiment yielded similar results with no appearance of a lower band that would indicate successful cleavage.

Dynamic light scattering (DLS) was then performed in order to determine an approximate size of the species, which could help to identify if it is a monomer, a dimer, with or without a GST tag or in fact in complex with DNA.

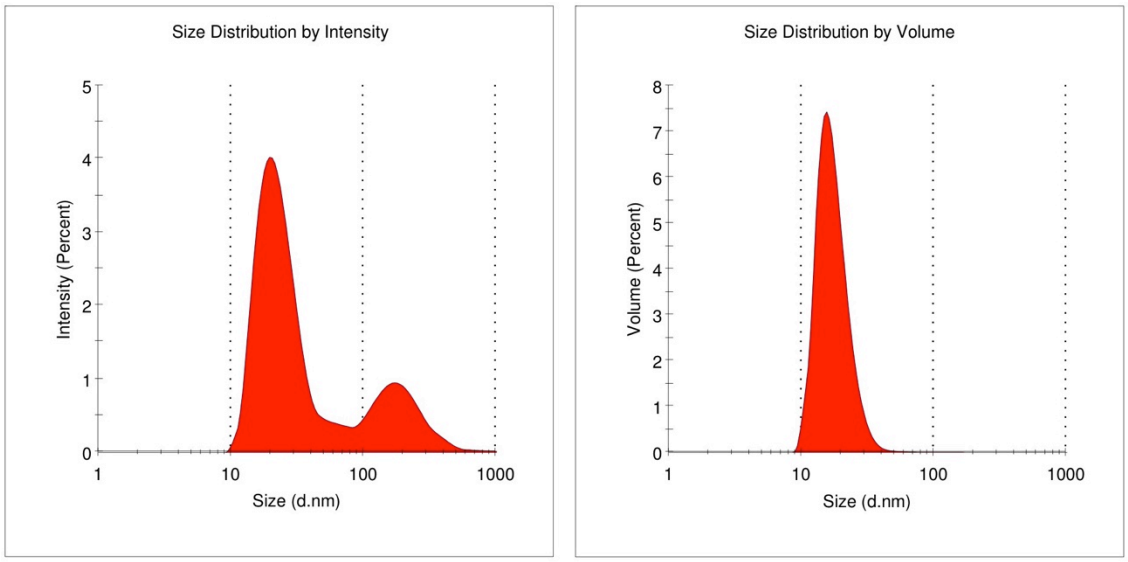


Figure 3.11 - DLS data showing the size distribution against both the intensity of the scattering signal and volume in per cent for GST-A1. The peaks are an average of four repeats that were performed at two concentrations; 0.25 mg/mL and 0.125 mg/mL.

The main peak, visible in both graphs, had a mode diameter of 15.51 nm (s.d. 7.05 nm), which produced a MW of 552 kDa, calculated by using the empirical mass against a size calibration curve automatically by the software (Malvern Instruments). This value is most likely due to the large and varied levels of aggregation, visible in the intensity peak, and not a true representation of the size of the protein in solution, or the protein in complex with DNA

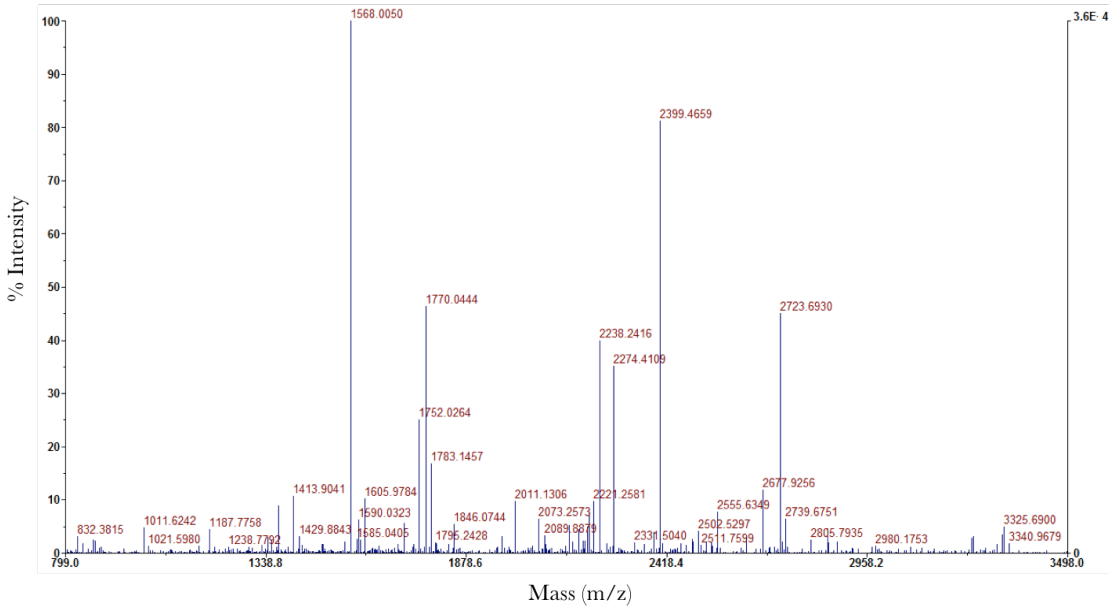
Peptide mass fingerprinting was then carried out using the protocol discussed in chapter 2 to determine a more accurate answer to the identity of the band. Using the same sample seen in the silver stained 'input' band seen in figure 3.11, a new SDS-PAGE experiment was performed and the prevalent band was excised and digested with trypsin, which cleaves after K or R residues within the sequence. The results of the fingerprinting can be seen in figure 3.13.

The excised band was identified as a 60 kDa *E. coli* chaperonin. With 65% coverage, 23 peptide matches and a protein score of 183, this information can be considered accurate.

A definition of the protein score is -10^{LogP}, where P is the probability that the observed match is a random event. Any protein score above 70 has a statistical significance, with a $p < 0.05$.

The fact that the main protein being purified is a chaperone indicates that the codon optimised GST-A1 required significant attention from the host chaperones in order to fold correctly and maintain this structure. Although no direct evidence for the protein ever being correctly folded is available, the presence of a chaperone throughout the procedure indicates attempts to fold the protein. The interesting aspect of this outcome however, is the fact that even though it was a chaperone it produced a signal when used in a western blot developed with α-DNA-PKcs. This could be due to the fact that they are co-migrating, seeing as they are similar sizes. The epitope for the antibody is for a portion of the FATC domain of DNA-PKcs, which no homology to any portion of chaperonin. If they are maintaining a complex (85.9 kDa + 60 kDa with the GST tag or 60.4 kDa + 60 kDa without the tag) then this could in fact be the reason that the protein is being eluted in the void volume, but would not explain why only one was being seen in the SDS-PAGE gels or western blots. Another factor against this theory is the lack of peptides being observed that would map to the DNA-PKcs sequence, which, if they were co-migrating, would be detectable after the tryptic digest.

A solution to this has not been determined and this will be discussed in more depth in the future work chapter of this thesis. In parallel to the GST-A1 work, studies were also performed on the two other constructs that performed well in Rosetta pLysS cells, grown in TB. These constructs were His-A1 and SUMO-A1.




```
        10         20         30         40         50         60
MAAKDVKFGN DARVKMLRGV NVLADAVKVT LGPKGRNVVL DKSFGAPTIT KDGVSVAREI

        70         80         90        100        110        120
ELEDKFENMG AQMVKEVASK ANDAAGDGTT TATVLAQAII TEGLKAVAAG MNPMDLKRGI

       130        140        150        160        170        180
DKAVAAAVEE LKTLSVPCSD SKAIAQVGTI SANSDETVGK LIAEAMDKVG KEGVITVEDG

       190        200        210        220        230        240
TGLQDELDVV EGMQFDRGYL SPYFINKPET GAVELESPFI LLADKKISNI REMLPVLEAV

       250        260        270        280        290        300
AKAGKPLLII AEDVEGEALA TLVVNTMRGI VKVAAVKAPG FGDRRKAMLQ DIATLTGGTV

       310        320        330        340        350        360
ISEEIGMELE KATLEDLGQA KRVVINKDTT TIIDGVGEEA AIQGRVAQIR QQIEEATSDY

       370        380        390        400        410        420
DREKLQERVA KLAGGVAVIK VGAATEVEMK EKKARVEDAL HATRAAVEEG VVAGGGVALI

       430        440        450        460        470        480
RVASKLADLR GQNEDQNVGI KVALRAMEAP LRQIVLNCGE EPSVVANTVK GGDGNYGYNA

       490        500        510        520        530        540
ATEEYGNMID MGILDPTKVT RSALQYAASV AGLMITTECM VTDLPKNDAA DLGAAGGMGG

MGGMGGMM
```

Figure 3.12 – Graph highlighting the mass peaks and their associated intensity (%). The mass of these peaks relates to the peptides highlighted in the sequence coverage map. This map shows the sequence of a 60 kDa *E. coli* chaperonin being purified instead of the target protein.

3.3.2 – Soluble His-A1 & SUMO-A1

The His-A1 and SUMO-A1 seen in table 3.1 had similar growth and expression levels to the GST-A1 and were taken forward simultaneously. Both tags were present on the N-terminus of the protein and the SUMO tag also contained a 6xHis tag for identification and purification purposes, however this was downstream of the SUMO tag itself. The solubility was tested in PIPES buffer by using both sonication and lysozyme as lysis techniques, with the results being directly compared on an SDS-PAGE gel seen in figure 3.14.

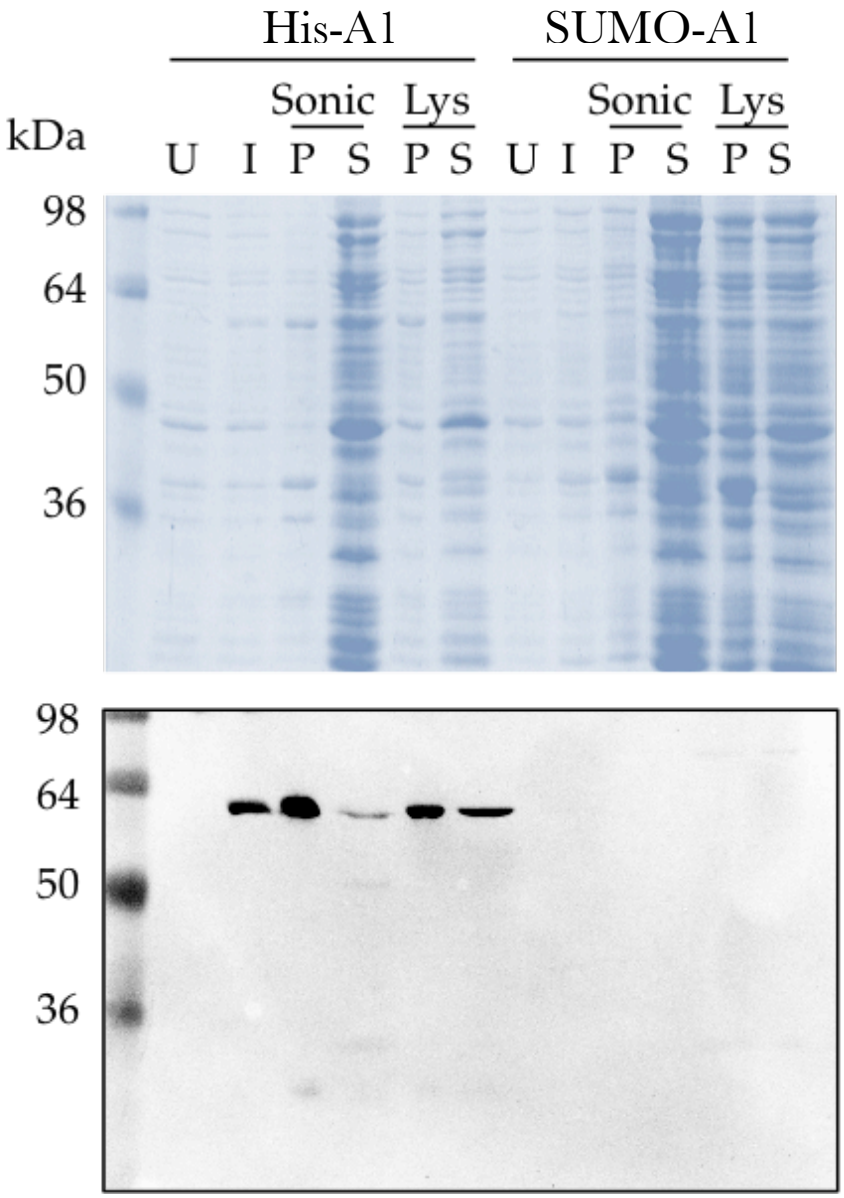


Figure 3.13 – Coomassie stained 15% SDS-PAGE and western blot developed with α-6xHis antibody showing the expression and lysis experimentation results for both His-A1 and SUMO-A1. This antibody was chosen due to the tag being present on both constructs. α -DNA-PKcs western blots were also performed with identical results, however these results are not shown. This experiment shows a direct comparison between sonication (Sonic) and lysozyme (Lys) as a means of cellular lysis. U – Uninduced, I – Induced, P – Pellet, S – Supernatant. The calculated molecular weight of His-A1 is 61.2 kDa and of SUMO-A1 is 72.7 kDa.

Directly comparing the results from His-A1 and SUMO-A1, only the 6xHis tagged construct actually produced protein. This result was verified through repeat experiments as

well as using α-DNA-PKcs rather than the α-His antibody results shown in the figure. The other observable, and repeatable, result from this experiment is the difference between the lysis techniques, with lysozyme treatment generating more soluble protein when compared to the sonicated sample.

At this point, due to the construct containing an N-terminal 6xHis tag, a binding affinity experiment was performed using Ni-Sepharose not bound to a column. After an interaction had been determined, a larger scale experiment was performed, whilst maintaining the use of the resin.

In the experiment described in figure 3.14, 500 µL of resin with 50 mL supernatant from 1 L of cell culture was used. The western blot, developed with α-DNA-PKcs, shows the His-A1 construct present in the supernatant, as well as coming off slightly at 100 mM imidazole, but with the majority of the protein being eluted at 500 mM imidazole.

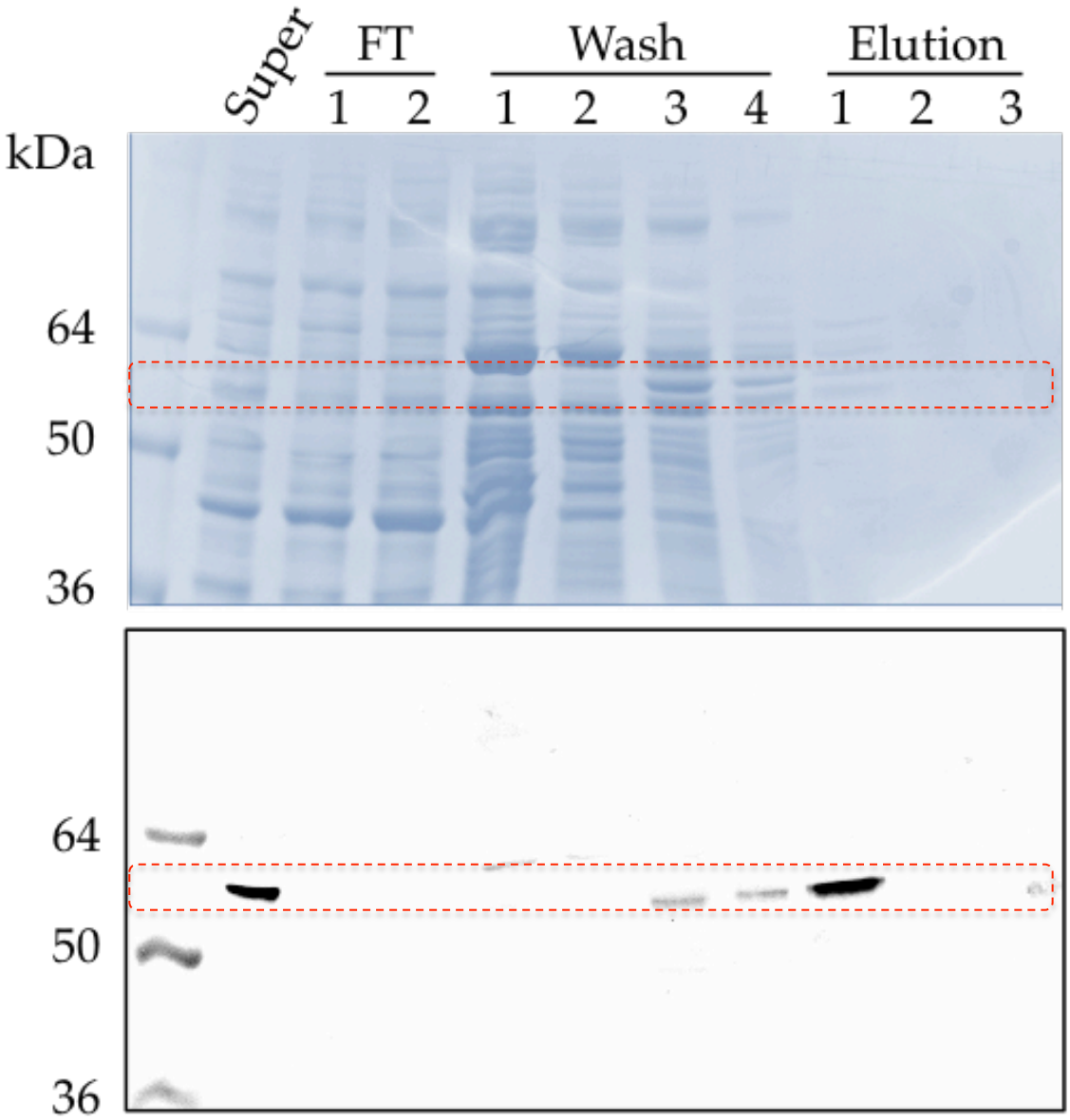
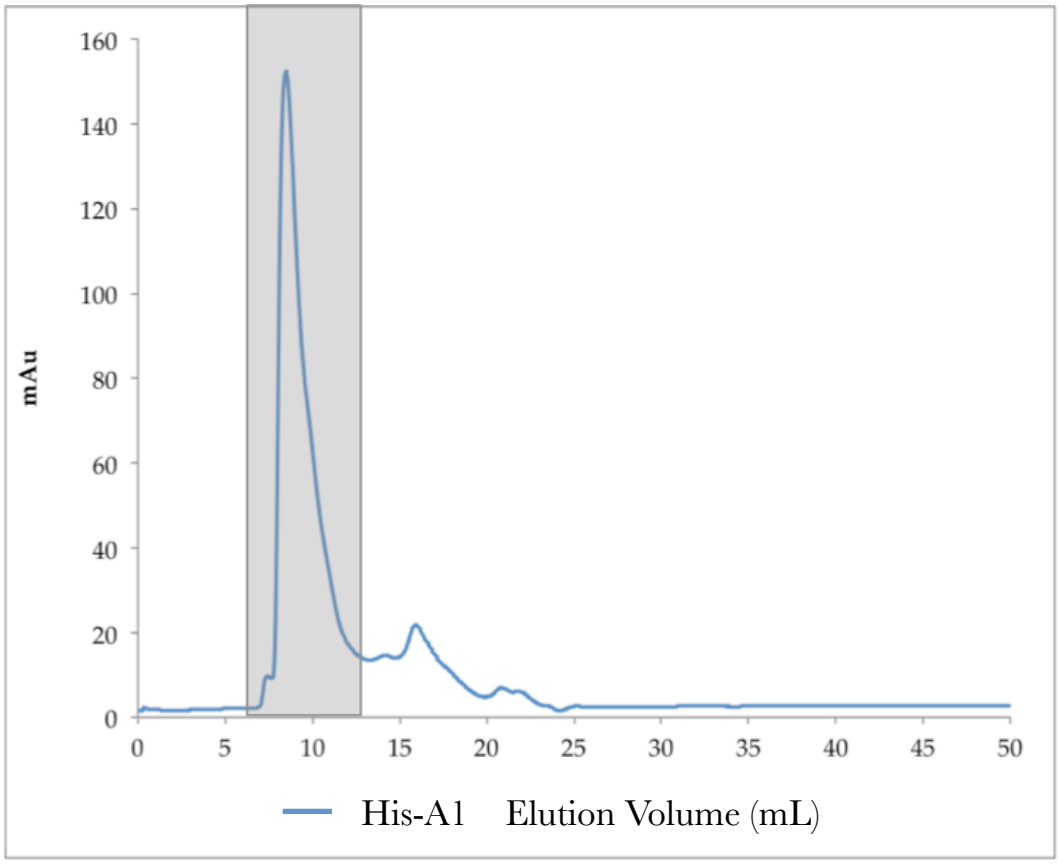


Figure 3.14 – 15% coomassie stained SDS-PAGE gel and western blot showing the results of the batch purification of His-A1. Western blot developed with α-DNA-PKcs. Washes 1 & 2 were performed with 20 mM imidazole. Washes 3 & 4 were performed with 100 mM imidazole. Elution was performed at 500 mM imidazole. The 'Elution 1' sample was concentrated to a final volume of 500 µL and loaded onto an s200 10/300 GL column (GE Healthcare). Red dashed box indicates predicted molecular weight.



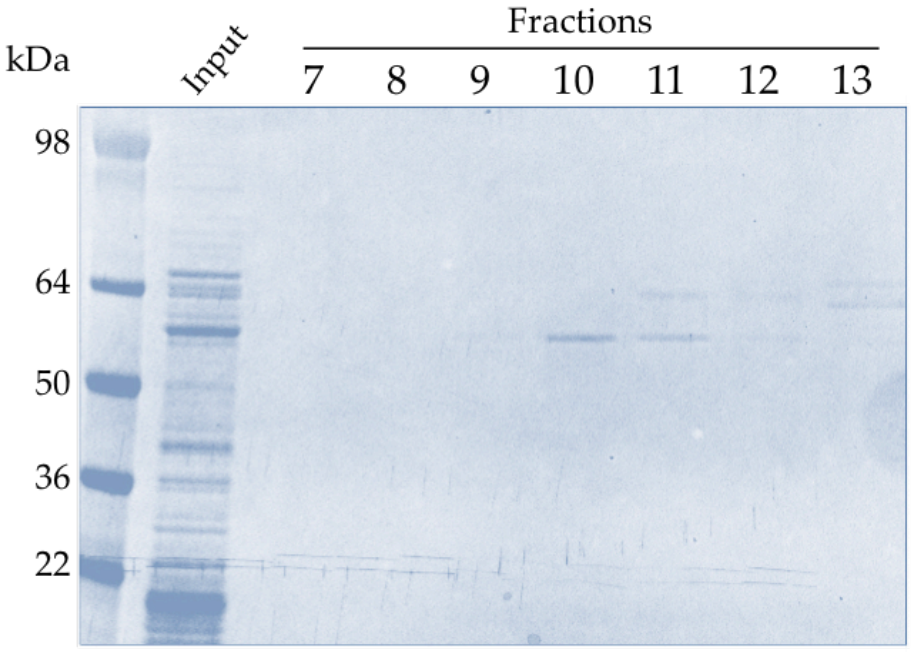


Figure 3.15 - His-A1 size exclusion trace and 15% coomassie stained SDS-PAGE gel showing highlighted (grey) fractions 7 – 13.

The void volume for this column is ~ 7.8 mL, and this is the point at which the peak is being eluted from the column. This indicates that the protein is much larger than predicted. It could be due to soluble aggregates forming, greatly increasing the hydrodynamic radius. It could also be oligomerisation of the protein, however this is less likely with a 6xHis tag rather than the GST tag mentioned previously. What is more likely is in fact nucleic acid contamination because the 260 nm / 280 nm ratio for sample 10 is 1.91.

In order to determine if these data were obtained as a result of nucleic acid contamination, experiments were performed with the aim of removing the nucleic acids from the protein.

The first of these experiments was to incorporate a high salt wash into existing protocols. The resin was replaced with 1 mL HisTrap columns, which allowed for greater control and variation throughout the wash stages due to the fact it is used in combination with an ÅKTA purification system and subtle changes to the concentration of elution buffer can be made more accurately. The imidazole concentration in the lysis buffer was increased to 60 mM, and the concentration of NaCl in the wash step was increased to 1.5 M. The target protein was then eluted using a gradient of imidazole. Prior to running on a size exclusion column the 260 nm / 280 nm ratio was calculated and the values for the combined fractions was 1.97, implying that a salt wash did not remove any of the proposed nucleic acid contamination.

A second purification step was then included to try and achieve this removal. Using the output of this last purification, a binding affinity assay was performed using ion exchange chromatography resin. Including strong and weak anion exchanging resin (Q and DEAE respectively), strong cation exchanging resin (SP) and Heparin. Firstly the salt concentration was reduced from 350 mM to 20 mM, for the binding step. An intermediate wash step at 100 mM NaCl was performed and then the protein was eluted at 1 M NaCl. The results of this assay can be seen in figure 3.16.

The theoretical pI of this protein, according to ProtParam (123), is 9.16. Based on this the protein should bind to an anion exchanger, which again is the result observed. There is clear binding for both the strong (Q) and weak (DEAE) anion exchange resins. The interesting difference between the two is that the DEAE actually separates the two bands that are migrating very close to one another, however the Q does not, most likely due to its strength as an anion exchanging resin. Strong anion exchange resins are charged over a large pH range whereas weak anion exchange resins remain charged over a much narrower pH range. This therefore allows for the two bands present in the input lane of figure 3.17, to be separated as the difference in their pI will be smaller than the difference of the range of charge between Q and DEAE resins.

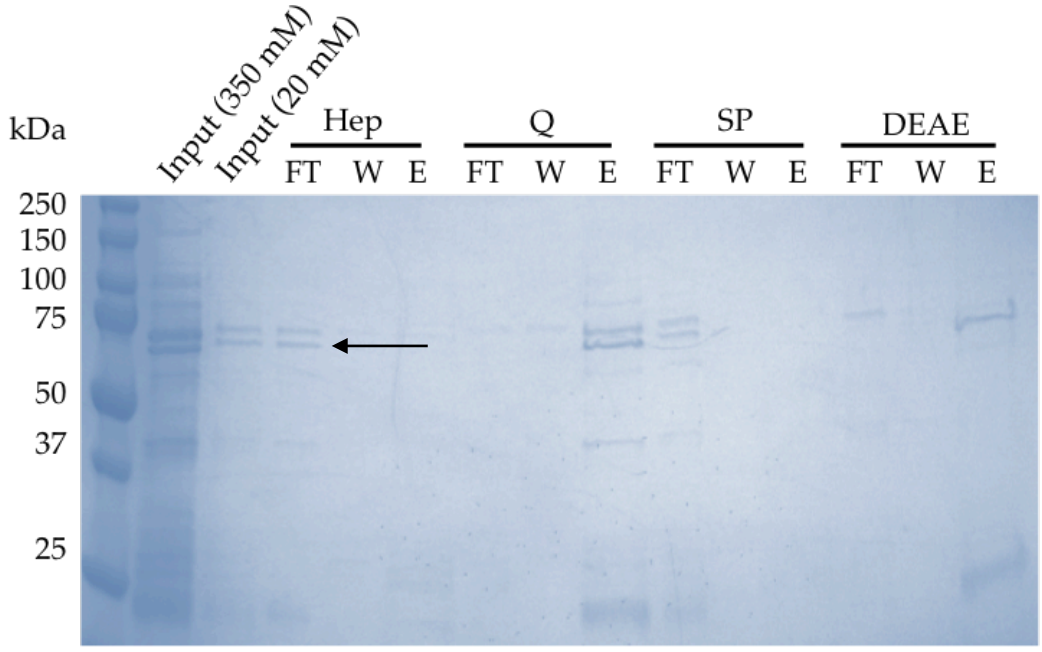


Figure 3.16 – SDS-PAGE gel, stained with coomassie, showing the results of an ion exchange binding affinity assay for His-A1. The input had a salt concentration diluted from 350 mM to 20 mM. Lanes show the flow-through (FT), wash (W) and elution (E). The lower band of the two in the couplet (highlighted by the arrow in the Heparin flow through lane) was deemed to be the correct band and therefore DEAE resin was incorporated into subsequent purification steps.

An intermediate ion exchange step using DEAE resin was then incorporated into a larger prep after the existing affinity chromatography step. The output from the DEAE column was then concentrated to a final volume of 500 µL and loaded onto a pre-equilibrated s200 10/300 GL column in order to determine if the ion exchange step served its purpose.

The result in figure 3.17 shows the chromatographic trace for a size exclusion step introduced after the DEAE resin to determine the effectiveness of nucleic acid removal. The first peak was eluted at around 8 mL, which is the void volume for this column. The second peak begins being eluted at around 12 mL. Fractions were collected for both peaks and were analysed on an SDS-PAGE and western blot to determine their contents. These results can be seen in figure 3.19

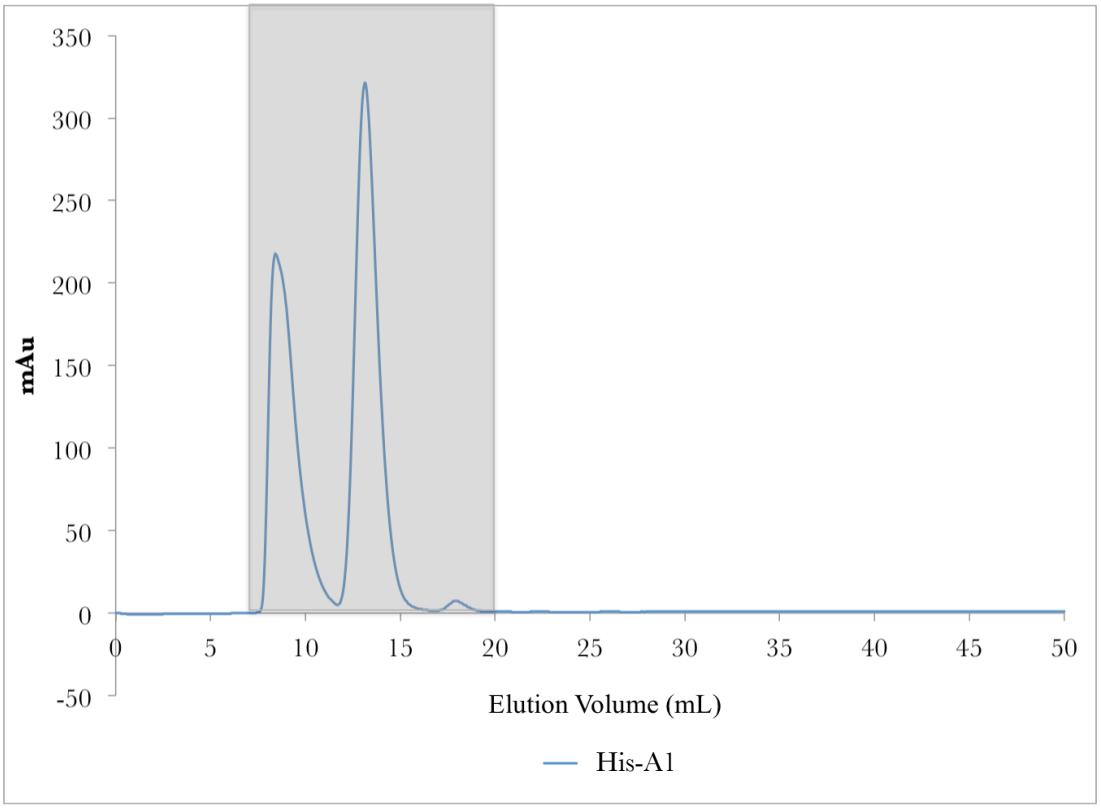


Figure 3.17 - Size exclusion chromatographic trace for the purification of His-A1 after previous anion exchange chromatography, shown in figure 3.16.

Fractions 8 and 9 contain the target protein and both relate to the void volume peak. Their 260 nm / 280 nm ratios are 1.97 and 1.99 respectively, indicating a failure of DEAE to remove the nucleic acid contamination. This could be due to the fact that at this pH, both the protein and the nucleic acid will have net negative charges and will therefore both bind to the resin. Future work could involve a repeat of this experiment at around pH 5, in order to generate a net positive charge on the surface of the protein, which in turn will serve as a separation technique between protein and nucleic acid.

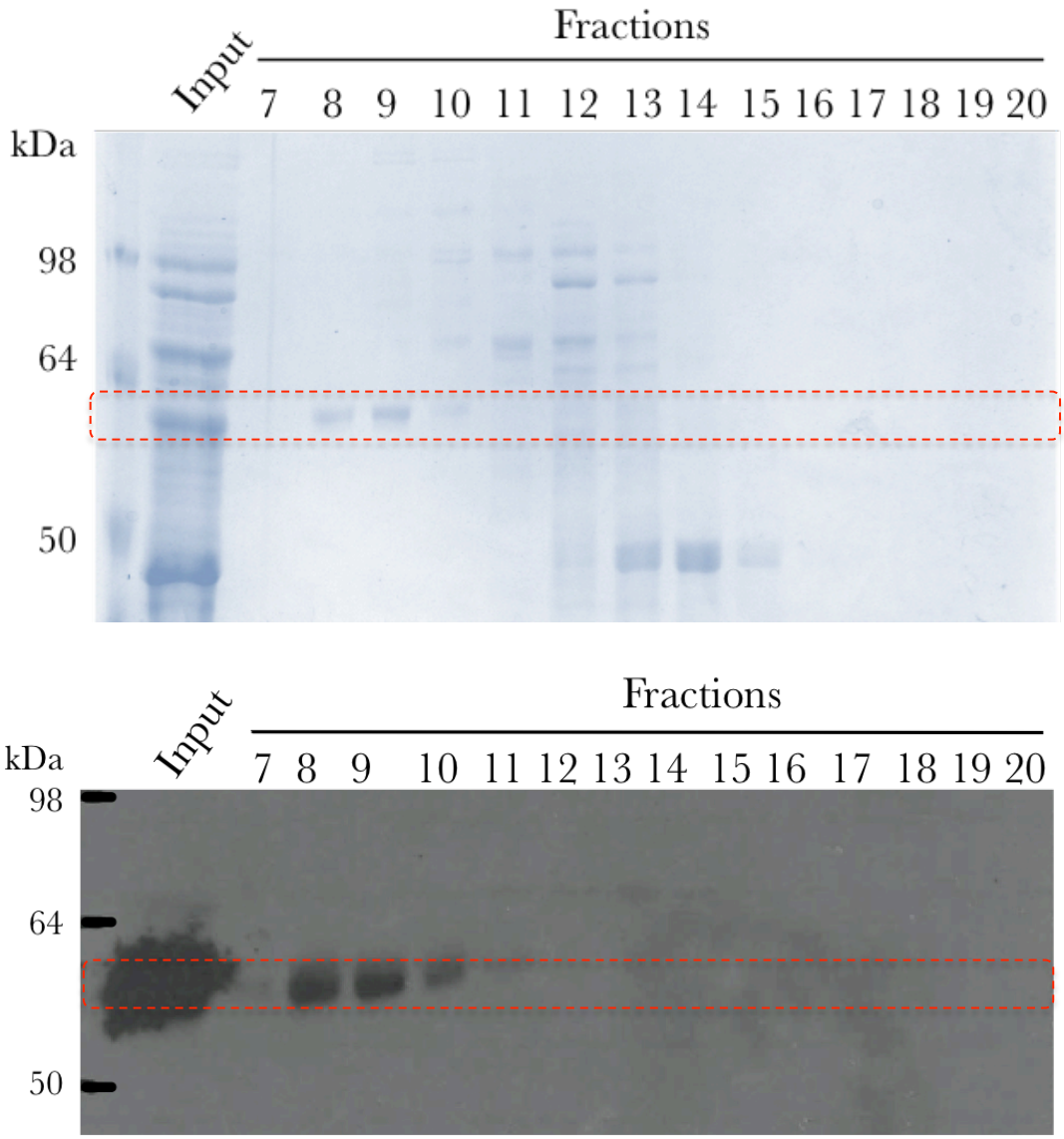


Figure 3.18 – 15% coomassie stained SDS-PAGE and western blot developed with α-DNA-PKcs antibody, showing the fractions obtained during size exclusion chromatography step, shown in figure 3.18, after the protein had already passed over a weak anion exchange chromatography column (DEAE). Red dashed box indicates predicted molecular weight.

The next technique employed to remove this apparent nucleic acid contamination was precipitation with polyethyleneimine (PEI). PEI is a cationic electrolyte at this pH and serves as an intermediate purification step that work by separating molecules based on their charge. In this instance the concentration of the PEI will be high enough to precipitate the nucleic acid whilst keeping the protein in the solution (125).

The final concentration of PEI used, after initial small-scale experimentation, was 0.025 %. The PEI was combined with benzonase during the lysis of the cells and then the remaining protocol made use of the high salt wash mentioned previously. After the PEI treatment, the protein was subjected to ammonium sulphate (AMS) in order to precipitate the protein, removing any trace of PEI. Tests were performed to determine the optimal concentration of AMS relative to the protein by starting off at 0 M until it was saturated (3.5 M). The optimal concentration determined and then used for this experiment was 3.2 M. Once the protein was precipitated, it was separated from the supernatant and resuspended in fresh lysis buffer for further treatment.

This involved the sample being loaded onto an s200 10/300 GL column, where it eluted in the void volume. Due to the protein being precipitated from the lysate, this acted as an alternative to affinity chromatography and, once resuspended, the protein was loaded directly onto the S200 size exclusion column. The fractions collected from this purification can be seen in figure 3.17. The 260 nm / 280 nm ratio after the PEI treatment in the most concentrated fraction (8), was 1.74. This is an improvement on previous results of ~ 1.97 however the problem is persistent enough to cause the protein to elute in the void volume. The western blot shown in this figure was developed with α-6xHis antibodies.

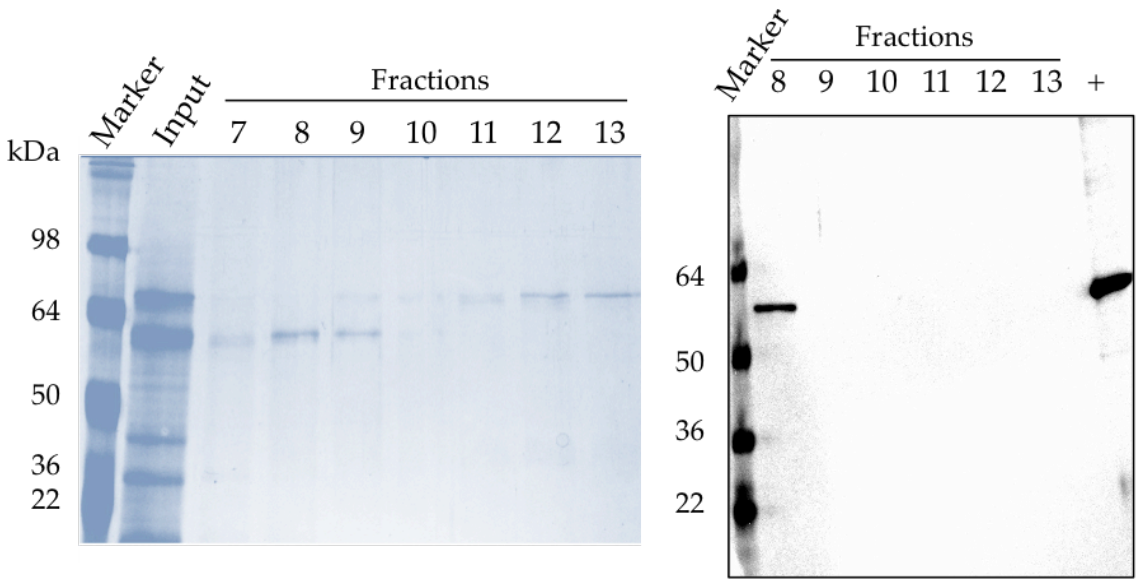


Figure 3.19 – 15 % SDS-PAGE and western blot, developed with α-6xHis antibodies, showing the results of size exclusion chromatography after treatment with AMS and PEI. + refers to a positive control for the western blot, which was a purified 6xHis tagged protein.

Due to this step being unsuccessful, an alternative protocol was attempted using hydrophobic interaction chromatography (HIC). The rationale behind using this technique being that if increased hydrophobicity is induced artificially, through the introduction of a high ionic strength buffer, any nucleic acid contamination then present on the protein will not bind to the resin and flow through, thus allowing purified protein without the contamination to be eluted. The chemical used to induce hydrophobicity by exposing the hydrophobic amino acids that would normally be buried within the structure, was AMS.

The concentration at which the protein precipitates using AMS was determined when performing the PEI treatment. The optimal precipitation concentration was 3.2 M, but the highest concentration possible prior to any precipitation was 1 M. Once the AMS was added to the sample, it was then divided up into three aliquots and passed over different HIC resins. These were Butyl-S, Octyl and Phenyl. The difference between the resins is the length of the backbone, where the length of the backbone affects their efficacy, with longer chains binding to increasingly hydrophobic proteins. Shorter length chains attract less hydrophobic proteins and therefore may not succesfully purify the protein, however longer chains induce stronger

interaction of the protein to the resin and yields may decrease through protein not being eluted. The results of this experiment can be seen in figure 3.19.

The SDS-PAGE gel shows that the target protein binds to all three resins. However none of them have served to purify the target protein away from the contaminant proteins present in the solution, nor have they managed to significantly remove the nucleic acid contamination. The input 260 nm /280 nm ratio was 1.97. The Butyl-S resin had a ratio of 1.35, however the protein had a concentration of 0.08 mg/mL. The Octyl resin reduced the ratio to 1.91 and the Phenyl resin reduced it to 1.84.

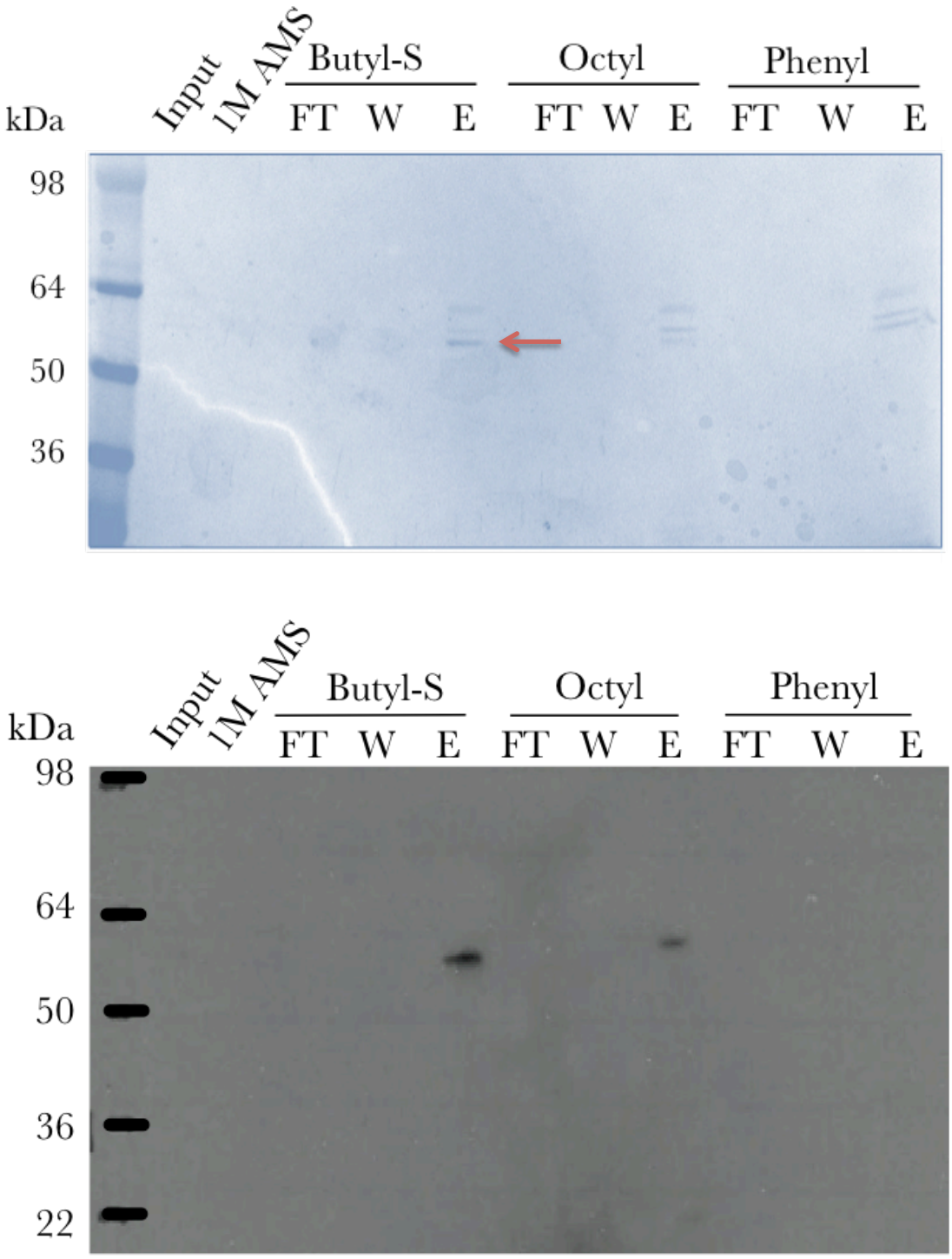
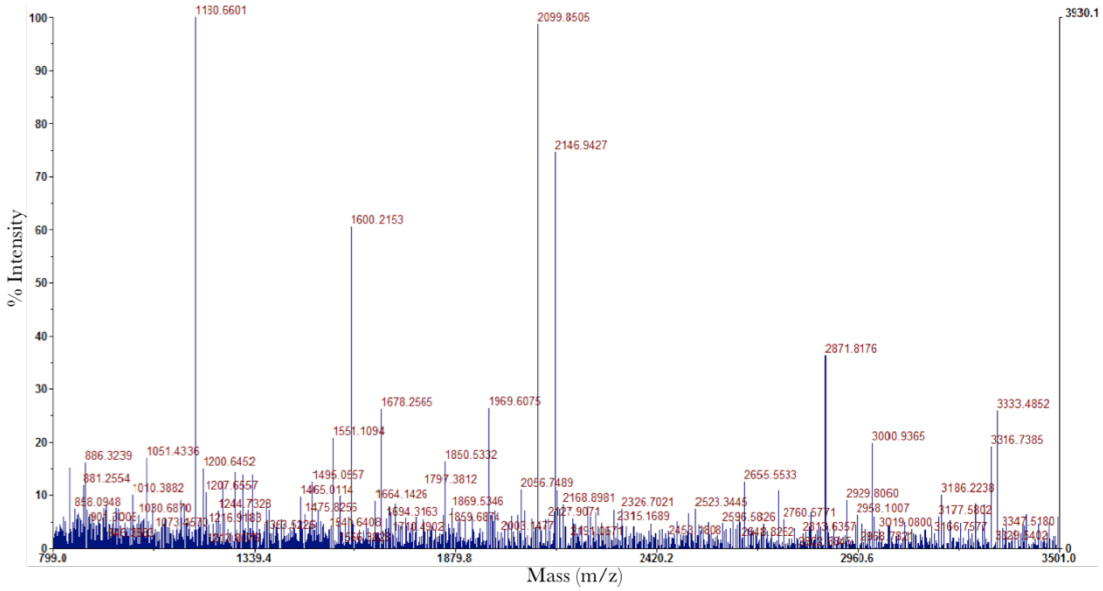


Figure 3.20 – 15 % coomassie stained SDS-PAGE gel and western blot developed with a-DNA-PKcs antibodies, showing the results of an HIC binding affinity experiment. The results in the gel show the protein binding to all three resins with equal affinities whereas based on the western blot there is a strong affinity for the Butyl-S and a slightly weaker interaction with the Octyl resin, with no binding interaction with the phenyl resin. The 1M AMS lane refers to the input after treatment with 1M AMS in order to induce hydrophobicity. The arrow indicates which band was excised for subsequent mass spectrometry analysis.

The sample highlighted in figure 3.20 was analysed using another SDS-PAGE, with 15 µL of sample loaded to allow for band excision and digestion with trypsin for peptide mass fingerprinting to be performed. Following the protocol discussed earlier, the experiment was performed and the results of this can be seen in figure 3.22.

The sequence seen in the figure is for a chaperone called GroEL from *E. coli*. The protein score was 2078 with 157 matched peptides and sequence coverage of 73 %. The strength of this evidence is convincing and leads to the conclusion that when using both GST-A1 and His-A1 that are codon optimised, they are not stable enough to exist without the presence of a chaperone. This interaction is strong enough to exist far from the intracellular environment and is strong enough to withstand some very stringent purification conditions. These results will be further discussed in chapter 5.




```
        10         20         30         40         50         60
MAAKDVKFGN DARVKMLRGV NVLADAVKVT LGPKGRNVVL DKSFGAPTIT KDGVSVAREI

        70         80         90        100        110        120
ELEDKFENMG AQMVKEVASK ANDAAGDGTT TATVLAQAII TEGLKAVAAG MNPMDLKRGI

       130        140        150        160        170        180
DKAVTAAVEE LKALSVPCSD SKAIAQVGTI SANSDETVGK LIAEAMDKVG KEGVITVEDG

       190        200        210        220        230        240
TGLQDELDVV EGMQFDRGYL SPYFINKPET GAVELESPFI LLADKKISNI REMLPVLEAV

       250        260        270        280        290        300
AKAGKPLLII AEDVEGEALA TLVVNTMRGI VKVAAVKAPG FGDRRKAMLQ DIATLTGGTV

       310        320        330        340        350        360
ISEEIGMELE KATLEDLGQA KRVVINKDTT TIIDGVGEEA AIQGRVAQIR QQIEEATSDY

       370        380        390        400        410        420
DREKLQERVA KLAGGVAVIK VGAATEVEMK EKKARVEDAL HATRAAVEEG VVAGGGVALI

       430        440        450        460        470        480
RVASKLADLR GQNEDQNVGI KVALRAMEAP LRQIVLNCGE EPSVVANTVK GGDGNYGYNA

       490        500        510        520        530        540
ATEEYGNMID MGILDPTKVT RSALQYAASV AGLMITTECM VTDLPKNDAA DLGAAGGMGG

MGGMGGMM
```

Figure 3.21 - Graph highlighting the mass peaks and their associated intensity (%). The mass of these peaks relates to the peptides highlighted in the sequence coverage map. This map shows the sequence of an *E. coli* chaperone called GroEL being purified instead of the target protein. Sequence results show 73% coverage, with 157 matched peptides and a protein score of 2078.

3.3.3 – Alternative Constructs

3.3.3.1 – Construct N (R3746-M4128)

Due to the failure of a codon optimised approach, the next logical step to generate the desired soluble protein was to test proteins with alternative domain boundaries. Sajish *et* al (126) recently published work performed on a soluble DNA-PKcs construct with alternative domain boundaries to the ones used in this study. The soluble construct, Construct N, began 30 amino acids downstream of the minimum kinase domain mentioned previously, at R3746 and continued to M4128 in the FATC domain of DNA-PKcs.

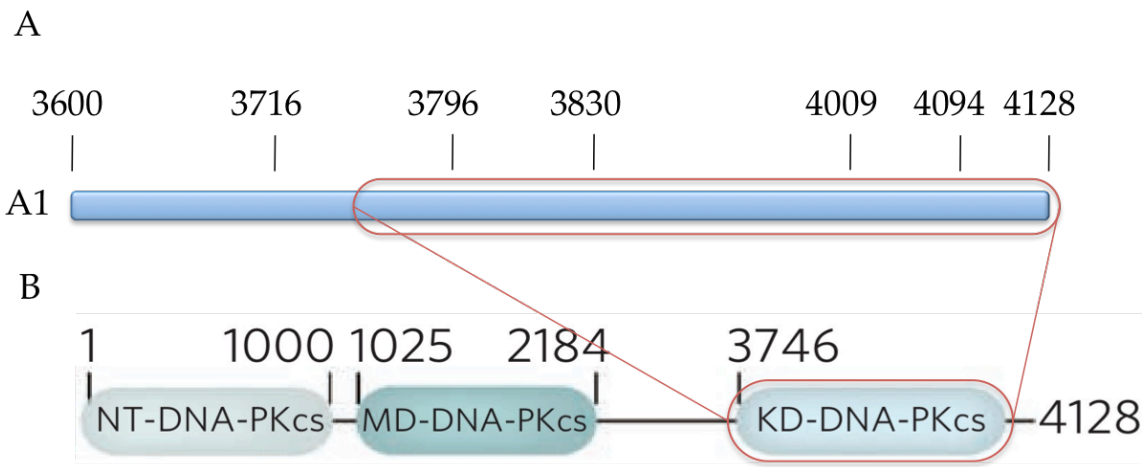


Figure 3.22 - (A) Construct schematic for A1, the construct used throughout these studies. (B) Construct schematic from Sajish *et al* showing the domain boundaries for their soluble DNA-PKcs (126). NT – N-Terminal, MD – Middle Domain, KD – Kinase Domain.

Using this publication as a reference, three new constructs were designed and purchased from Geneart. The first construct was designed without a tag, the second construct contained an N-terminal V5 tag, a 14 amino acid long peptide derived from a small epitope, present on paramyxovirus of simian virus 5. The third construct contained a C-terminal 6xHis tag.

Initial over-expression tests were performed, using Rosetta pLysS cells, grown in TB medium and the results of this can be seen in figure 3.23. This cell strain and media combination was used due to high growth yields in previous experiments.

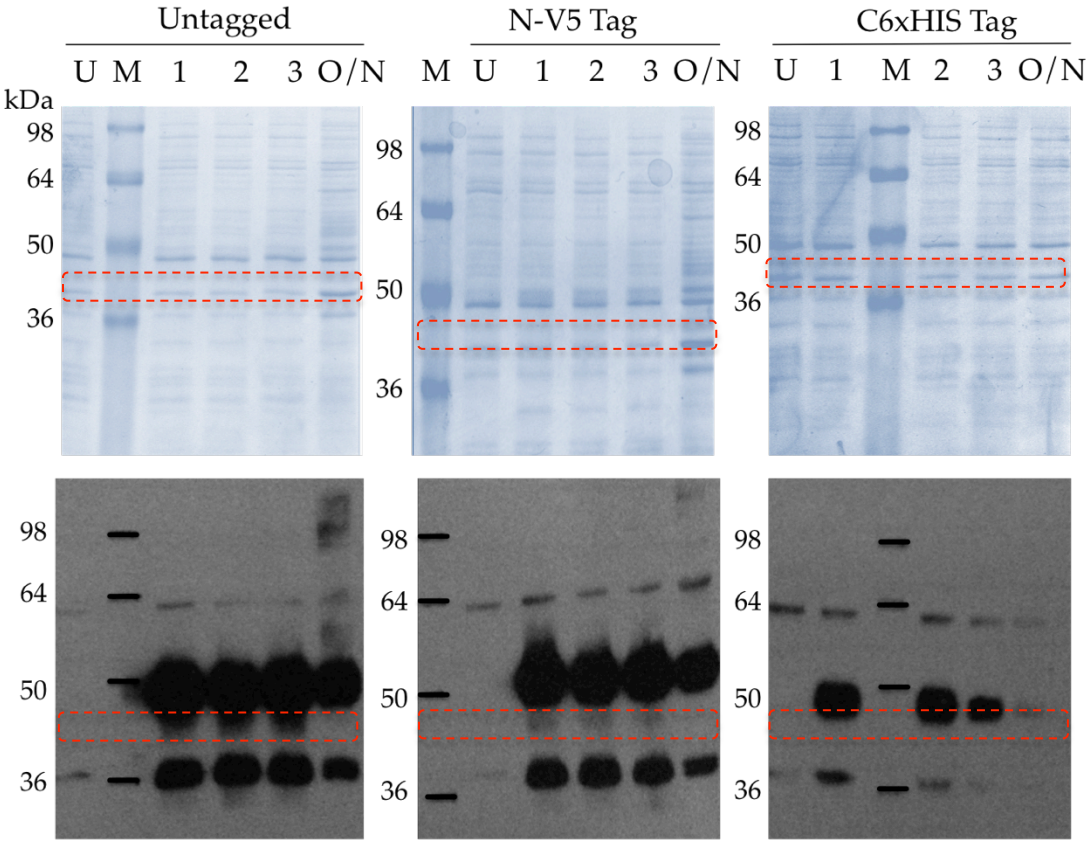


Figure 3.23 – SDS-PAGE gels and western blots developed using α-DNA-PKcs antibody showing Induction trials using three constructs derived from the soluble construct shown by Sajish *et al* (126). The untagged construct has an expected MW of 43.5 kDa. The V5-tagged protein has an expected MW of 44.9 kDa and the 6xHis tagged protein has an expected MW of 44.4 kDa. Red dashed box indicates predicted molecular weight.

The SDS-PAGE gels show low levels of over-expression but based on the western blots, developed with an α-DNA-PKcs antibody, there appears to be more significant over-expression. That being said, the chemiluminescent detection method can be affected by exposure time, with longer times producing greater signal, possibly exaggerating absolute levels of expression. Future work would include a variety of exposure times to reduce this as a variable. The western blots show both the untagged and V5-tagged proteins with significant

levels of either a smaller contaminant product or breakdown product. Future work would include limited proteolysis experiments to determine definitive breakdown products as well as stable sub domains present within the protein.

The 6xHis tagged protein was taken forward for further experimentation, due to lower levels of this smaller 'contaminant' band, and 2 hours was chosen as the optimum induction time. Initial solubility experiments showed distinct lack of solubility, concurrent with previous construct experiments. This was also true for a variety of lysis techniques

All cell pellets were lysed using lysozyme at a final concentration of 1 mg/mL for 1 hour at room temperature, at which point they were centrifuged at 22,000 g for 1. Based on the gel and western seen in figure 3.24, the lysis worked well with the vast majority of proteins present in the supernatant, however the target protein for each of the various conditions was always found in the pellet.

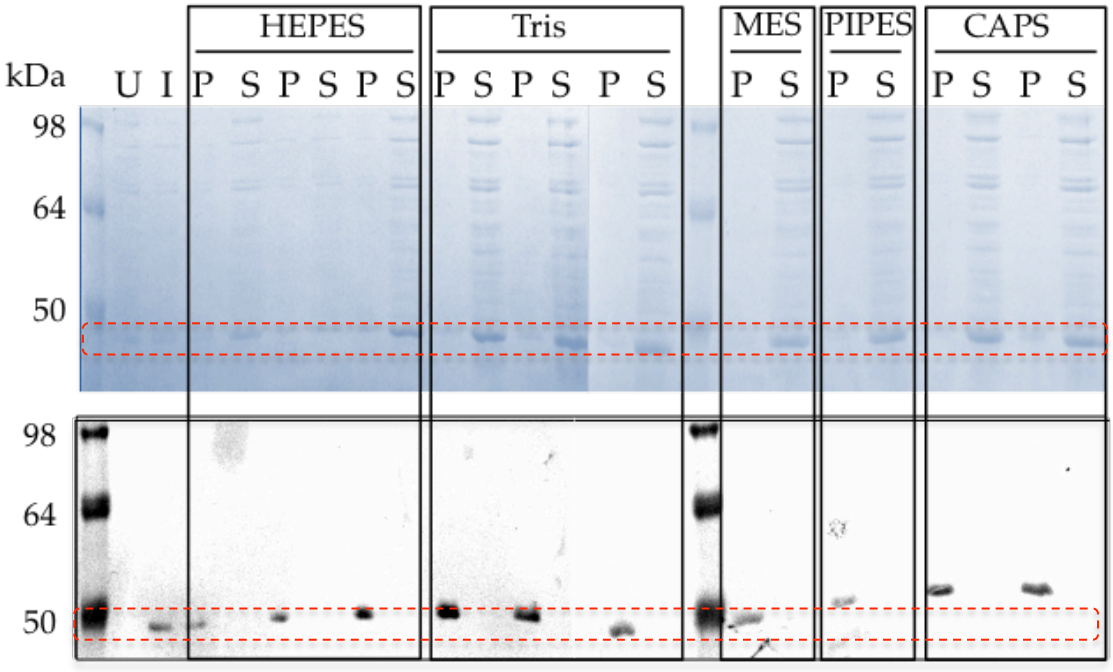


Figure 3.24 – 15 % SDS-PAGE and western blot, developed with α-6xHis antibodies showing the results of a buffer scout attempting to generate soluble target protein. Red dashed box indicates predicted molecular weight.

These figures indicate that these construct boundaries have done little to improve the solubility.

3.3.3.2 – pelB Leader

The pelB leader sequence refers to a 22 amino acid sequence of pectatelyase B of *Erwinia carotovora* CE (127) that, when attached to a protein, encodes for the transport of the protein from the cytoplasm, to the periplasm between the inner and outer membranes (128). Once there, the leader sequence is cleaved by signal peptidases present in the periplasm.

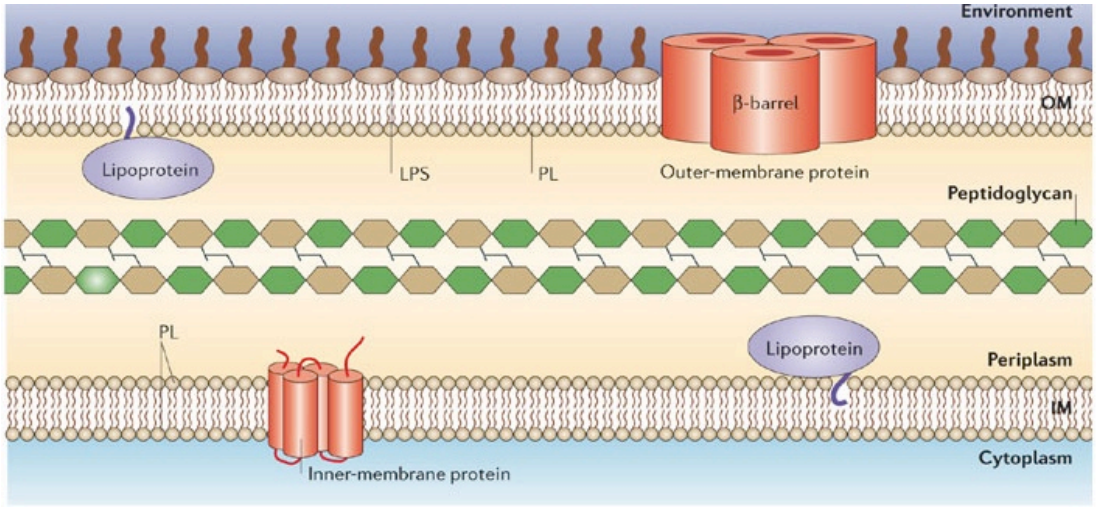


Figure 3.25 - Cartoon detailing the cell envelope, between the cell wall and cell membrane in gram negative bacteria such as *E. coli* (129).

One advantage of having a protein contained within the periplasm rather than the cytoplasm is that there are fewer proteases present that could potentially lead to proteolytic degradation. Another advantage is the fact that there are fewer chaperones present when compared to the cytoplasm. Due to the fact that this was such an issue for both 6xHis tagged and GST tagged codon optimised constructs discussed in the previous chapter, having a protein fold with fewer chaperones present could be beneficial. Yet another advantage is the fact that there are significantly fewer proteins present in the periplasm, and disruption of the outer membrane, whilst keeping the inner membrane intact, is a good means of keeping the

protein relatively pure before any further purification. This outer membrane lysis step was performed using osmotic shock using Tris-Sucrose-EDTA (TSE) buffer.

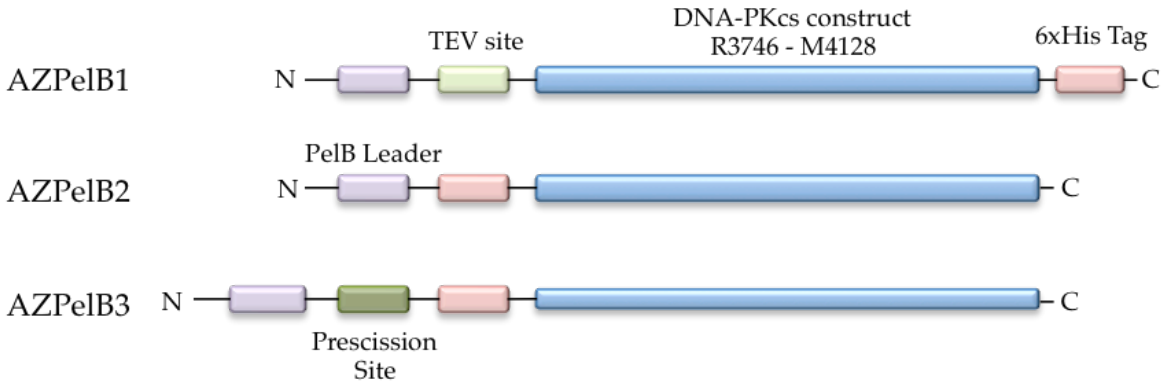


Figure 3.26 - Schematic of the three pelB leader constructs that were designed: Purple – PelB leader sequence; Light green – TEV cleavage site; Dark green – PreScission cleavage site; Red – 6xHis tag; Blue – construct with boundaries R3746 – M4128. All three constructs were cloned into the pTWO-E vector seen in figure 2.1.

In each of the constructs the pelB leader was at the N-terminus of the protein with either a TEV site upstream of the target protein (AZPelB1), a 6xHis tag upstream of the protein (AZPelB2), or a PreScission site as well as a 6xHis tag upstream of the target protein (AZPelB3). The presence of these cleavage sites should theoretically not be necessary due to the automatic leader cleavage by a signal peptidase but the efficiency of this cleavage is unknown and is negated by the presence of these cleavage sites.

The first experiment performed was an over-expression test, to determine if the presence of the PelB leader affected the levels of protein production. The results can be seen in figure 3.27. Each of the proteins is expressed well, with 3 hours being the optimal time before the cells are harvested.

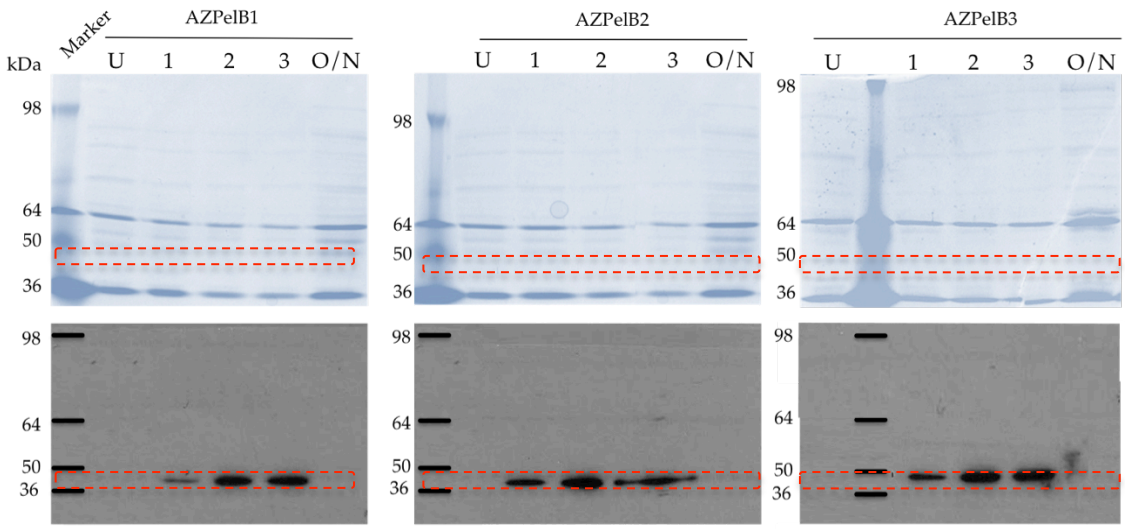


Figure 3.27 – 15% SDS-PAGE gels and western blots developed with α-DNA-PKcs antibody showing the expression experiments of constructs containing a PelB leader. AZPelB1 has an expected MW of 47.4 kDa, AZPelB2 has an expected MW of 46.6 kDa and AZPelB3 has an expected MW of 47.4 kDa. Red dashed box indicates predicted molecular weight.

At this point the osmotic shock experiment was performed in order to disrupt the outer membrane of the cell, whilst keeping the inner membrane intact. This was performed using TSE buffer (200 mM Tris-HCl pH 8.0, 500 mM Sucrose, 1mM EDTA). Cell pellets were resuspended in 1 mL of this TSE buffer and kept on ice for 30 minutes at which point they were centrifuged at 16,000 g for 30 minutes at 4 °C. This supernatant then contained the periplasmic proteins. The results of this experiment can be seen in figure 3.29. The media sample was included in the gel due to the fact that some pelB leader sequences can be extremely efficient at the export of the protein, and occasionally export the protein through the periplasm and through to the media in which the cells are grown.

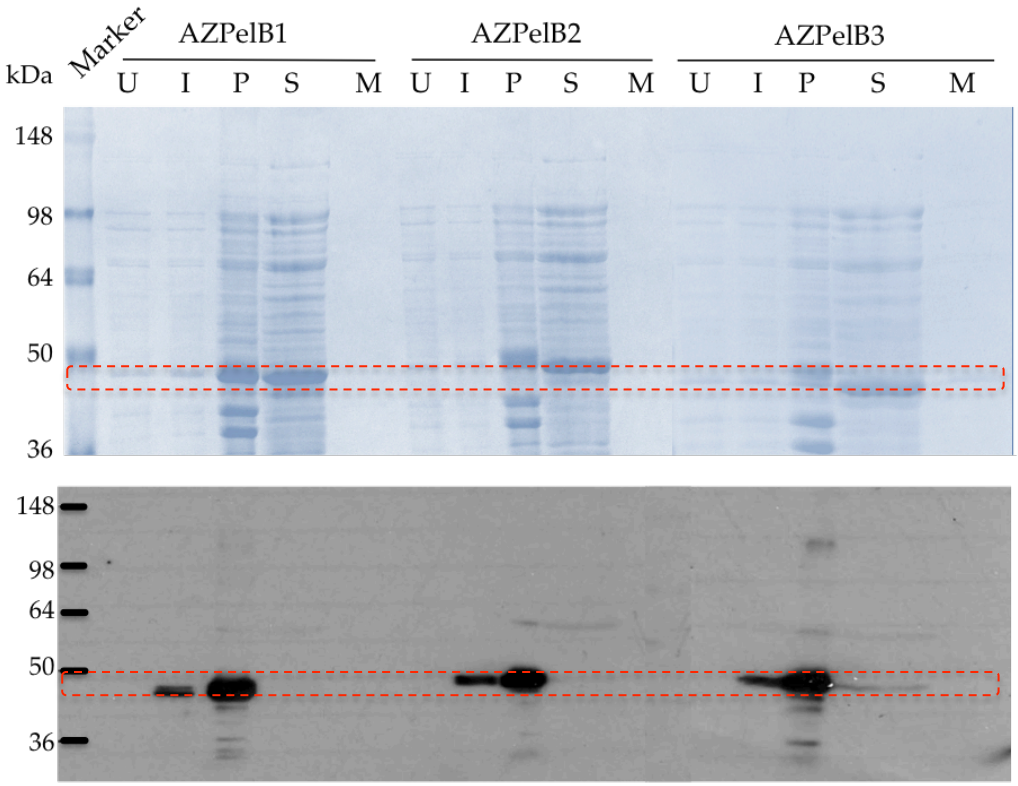


Figure 3.28 – 15% SDS-PAGE and western blot developed with α-DNA-PKcs antibody showing the results of testing the periplasmic expression of the three proteins. The lanes show uninduced (U), Induced (I), Pellet (P), Supernatant (S) and Medium (M). Red dashed box indicates predicted molecular weight.

The figure shows that the pelB leader sequence has been ineffective in exporting the protein to the periplasm. Although there is a strong band visible in the SDS-PAGE at the correct apparent Mw that does not produce a western blot signal. This could potentially be the desired product with the C-terminal region, the epitope for the a-DNA-PKcs antibody, degraded. This would require further experimentation to validate. At this stage however it is unknown whether it is in a soluble form in the cytoplasm or whether it has also formed inclusion bodies. The next experiment therefore, was to test if the presence of the pelB leader had increased solubility, even though it failed to carry out its primary function.

The pellet from this experiment was then taken forward and lysed *via* sonication with three 30 second bursts, each with 30 second intervals and then centrifuged at 20,000g for 60

minutes at 6°C. As well as the previous uninduced and induced samples, the fresh pellet and supernatant samples were analysed on an SDS-PAGE gel and western blot to determine if the pelB leader had any effect on solubility.

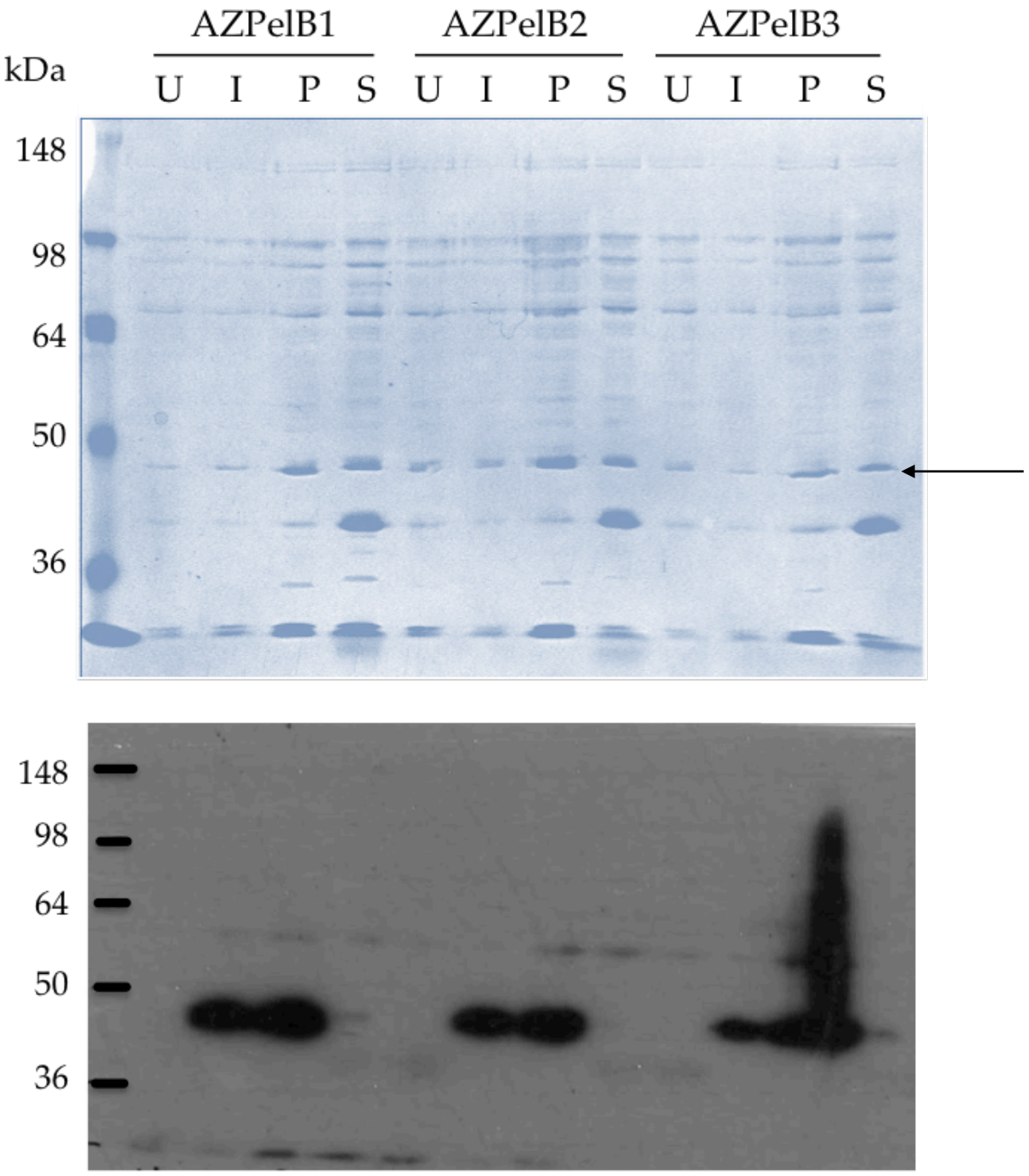


Figure 3.29 – 15 % SDS-PAGE and western blot, developed with α-6xHis antibodies, testing the solubility of the pelB constructs post-sonication. The arrow indicates the protein band, based on the information seen in the western blot below.

What is apparent from the figure is the fact that although there were moderate levels of expression, and with the lysis being effective, the target protein is still insoluble, and is present only in the pellet fractions.

This therefore means that soluble DNA-PKcs catalytic domain could not be generated using the alternative domain boundaries presented by Sajish *et al* (126). The authors of this paper make no reference to the yields of protein obtained through this protocol, although they imply that it is suitable for their cellular assays. After extensive attempts at improving the solubility, varying the vector, cell strain, buffer system, and making use of codon optimisation, all attempts at generating soluble protein failed. These results will be discussed further in Chapter 5.

The next route to take in an attempt to produce soluble DNA-PKcs kinase domain was denaturing and refolding the expressed protein from inclusion bodies, which is normally performed through the addition of a chaotrope such as guanidinium hydrochloride or urea.

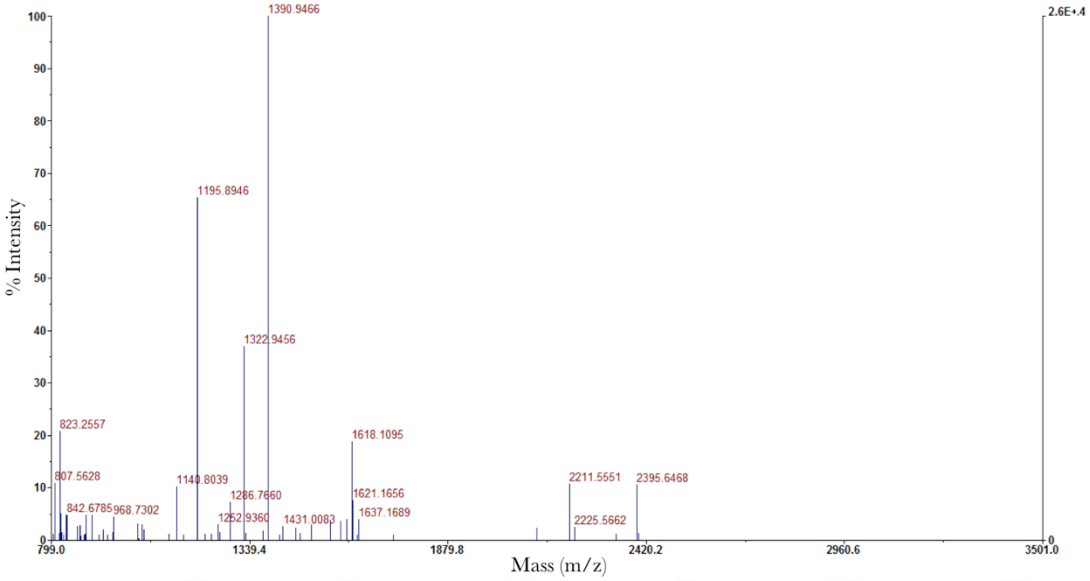
3.4 Protein Denaturation

Under native conditions a protein will fold into a conformation that is most thermodynamically stable in that environment. The conformation that it adopts is affected purely by the amino acid sequence, more specifically, the side chains they possess and how they interact with one another (130). The process of protein folding is a very rapid one, and it occurs during protein synthesis. Secondary structure dynamics studies have shown that for single alpha helices and short polypeptides the folding to form entire helices can be completed in hundreds of nanoseconds (131). High concentrations of the new protein localised close to one another can lead to a protein misfolding, which can very easily lead to an insoluble product such as an inclusion body (132).

If the protein is forming inclusion bodies, by denaturing and refolding in suitable conditions, bio-active protein can be recovered (132). This therefore was the next course of action.

When the proteins present in the cell pellet (target and host proteins alike) are resuspended in a chaotropic solution like urea, the chaotrope disrupts the hydrogen bonding network between water molecules and thus reducing the stability of the native state of a protein (133). The secondary structure is disrupted, with α helices and β strands no longer being ordered, but in fact adopting a randomly coiled conformation (134). The primary structure however is made up of covalent peptide bonds linking the amino acids, which cannot be broken down by the presence of such a chaotrope.

Prior to any significant devotion of time to the denaturing and refolding of protein, and based on the discovery of chaperone purification in the previous chapter, the identity of the overexpressed protein that was due to be denatured needed to be determined. This was done with peptide mass fingerprinting. After induction, HIS-A1 samples were analysed by SDS-PAGE and stained with SimplyBlue Safestain (Invitrogen). At this point the band that correlated with the western blot signal was excised. This was then run on the MALDI-TOF and the results can be seen in figure 3.31.




```
          10         20         30         40         50         60
MHHHHHHGGG ENLYFQGSVN KKNIEKMYER MYAALGDPKA PGLGAFRRKF IQTFGKEFDK

          70         80         90        100        110        120
HFGKGGSKLL RMKLSDFNDI TNMLLLKMNK DSKPPGNLKE CSPWMSDFKV EFLRNELEIP

         130        140        150        160        170        180
GQYDGRGKPL PEYHVRIAGF DERVTVMASL RRPKRIIIRG HDEREHPFLV KGGEDLRQDQ

         190        200        210        220        230        240
RVEQLFQVMN GILAQDSACS QRALQLRTYS VVPMTSRLGL IEWLENTVTL KDLLLNTMSQ

         250        260        270        280        290        300
EEKAAYLSDP RAPPCEYKDW LTKMSGKHDV GAYMLMYKGA NRTETVTSFR KRESKVPADL

         310        320        330        340        350        360
LKRAFVRMST SPEAFLALRS HFASSHALIC ISHWILGIGD RHLNNFMVAM ETGGVIGIDF

         370        380        390        400        410        420
GHAFGSATQF LPVPELMPFR LTRQFINLML PMKETGLMYS IMVHALRAFR SDPGLLTNTM

         430        440        450        460        470        480
DVFVKEPSFD WKNFEQKMLK KGGSWIQEIN VAEKNWYPRQ KICYAKRKLA GANPAVITCD

         490        500        510        520        530        540
ELLLGHEKAP AFRDYVAVAR GSKDHNIRAQ EPESGLSEET QVKCLMDQAT DPNILGRTWE


GWEPWM
```

Figure 3.30 - Graph highlighting the mass peaks and their associated intensity (%). The mass of these peaks relates to the peptides highlighted in the sequence coverage map. This map shows a manually entered sequence for DNA-PKcs, more specifically His-A1. Sequence results show 75% coverage, with 226 matched peptides and a protein score of 8138.

What can be seen in the figure is the sequence for human DNA-PKcs – more specifically the manually entered sequence for the entirety of the A1 construct. It shows 75 % sequence coverage, with 226 matched peptides and a protein score of 8138, which confirms that the overexpressed protein is the target protein. Although coverage indicates the first 30 residues are missing, this is most likely not the case due to the fact that the N-terminal 6xHis tag is being detected with immunoblotting and can therefore be assumed to be present. At this point further work can be performed with the knowledge that it is being done on the target protein rather than a contaminant or chaperone.

3.4.1 – Urea

Throughout this study both urea and guanidinium hydrochloride (GuHCl) were tested, but a decision was made early on that urea would be the chaotrope of choice and only the results of this will be discussed in the chapter. This decision rested on the fact that refolding with GuHCl was less reproducible than with urea and the fact that a lot of the analysis was performed with SDS-PAGE. When GuHCl comes into contact with SDS it precipitates, which causes the gel to run very irregularly making the analysis a lot more difficult, although this could have been avoided through precipitating any GuHCl present prior to performing any SDS-PAGE analysis.

The levels of expression of the A1 construct were very high when using IPTG at a final concentration of 1 mM for an overnight induction, using TB as the growth media. This therefore meant that there was a lot of starting material to take forward, which would prove to be essential due to the high attrition levels commonly seen with denature/renature protocols.

The cell pellets that were stored at -80 °C were taken forward and resuspended in a solution of 20 mM HEPES pH 7.5, 1 M NaCl, 10% glycerol, 5 mM $MgCl_2$. Based on the mass spectrometry data shown in figure 3.31 the pellet was known to contain target protein

and was subsequently processed by being washed 5 times in lysis buffer either on its own or in combination with one of two detergents, Brij-35 and Triton x-100. This wash step was extremely useful as it served as a purification step by partially solubilising a series of host proteins and increasing the ratio of target protein to contaminant host proteins. The results of this pellet washing protocol can be seen in figure 3.31.

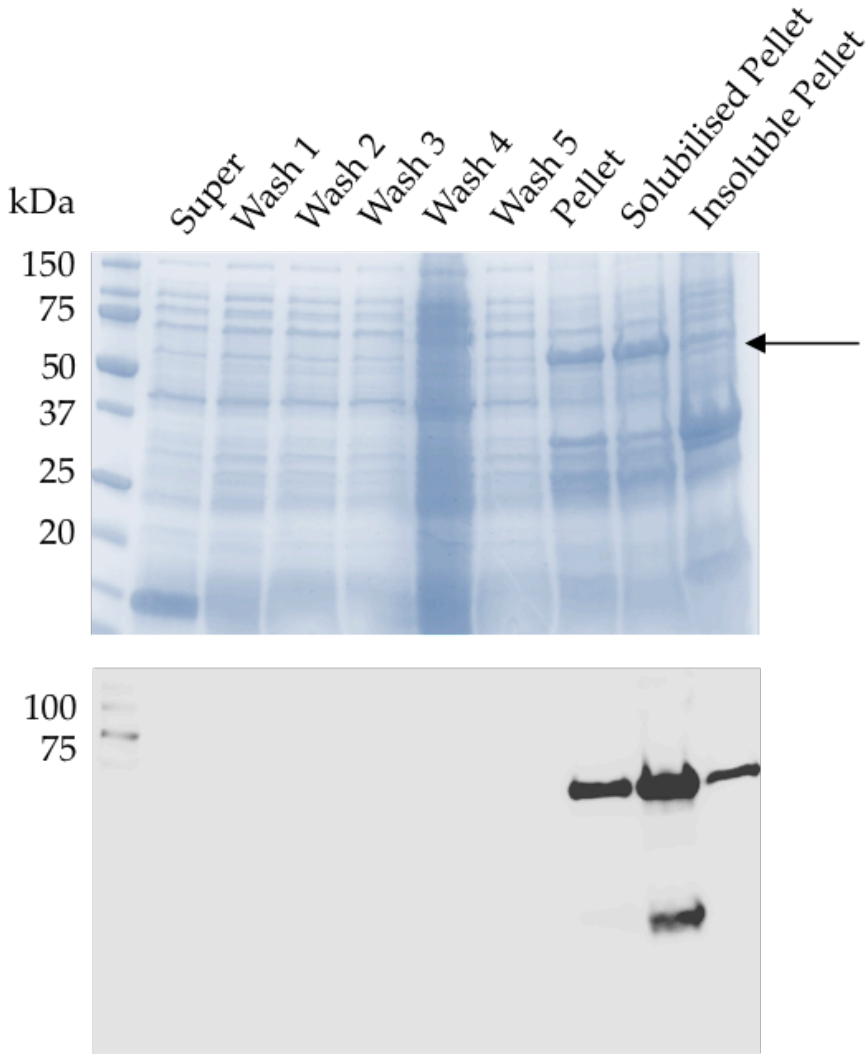


Figure 3.31 – 15 %SDS-PAGE and western blot showing the result of pellet washing, as well as the denaturing of the pellet. Wash 1 consisted of buffer only. Washes 2 & 3 consisted of this buffer supplemented with 0.05 % Brij-35. Wash 4 contained the buffer supplemented with 0.1 % Triton X-100 and wash 5 was the buffer only.

After being washed 5 times in a the combination of solutions shown in the figure, the pellet was then taken and suspended in the wash buffer, supplemented with 8 M urea and incubated with rotation overnight. This solution was then centrifuged at 25,000 g for 2 hours. The input pellet, solubilised pellet and remaining treated pellet can also be seen in figure 3.31. What is apparent is the appearance of a breakdown product significantly lower than the target protein after contact with urea. This was a frequent presence but was removed through purification. The other aspect of this result to be discussed is the presence of a small amount of DNA-PKcs still present in the pellet even after urea treatment. Even after significant rounds of refinement and optimisation this could not be improved upon.

To purify this newly solubilised protein the N-terminal 6xHis tag was employed and bound, whilst in a denatured state, to Ni^{2+} IMAC resin. Several rounds of optimisation were performed, firstly using 1 mL HisTrap columns and then 5 mL HisTrap columns (GE Healthcare). Both however had a significant amount of target protein present in the flow-through. Final optimisation involved the use of 10 mL of Ni-Sepharose resin (GE Healthcare) loaded into an XK-16 column and equilibrated. This too had some degree of flow-through issues, which were negated by a second pass of the flow-through sample. The capacity for this resin is 40mg of His-tagged protein per ml of resin. This 10 ml column therefore could potentially allow for 400mg of protein per run, however this is calculated based on a folded, globular protein. A protein in a denatured state may occupy significantly more space whilst adsorbed to the resin, reducing its effective capacity. Further experiments at greater column volume could be tested in future, but were not included in this study due to the success of re-loading the flow through fraction from the first purification. The protein was purified in a denatured state in order to increase the likelihood of getting the protein to refold in the correct manner without any contaminants present inhibiting this process.

The results of the optimised purification protocol can be seen in figure 3.33. The flow through issue is apparent from the FT sample from purification one. The flow through from

purification two shows a smaller amount of target protein not binding. All positive fractions from both runs were combined and concentrated to a final concentration of 5 mg/mL and stored in 100 µL aliquots at -80 °C ready for refolding experimentation.

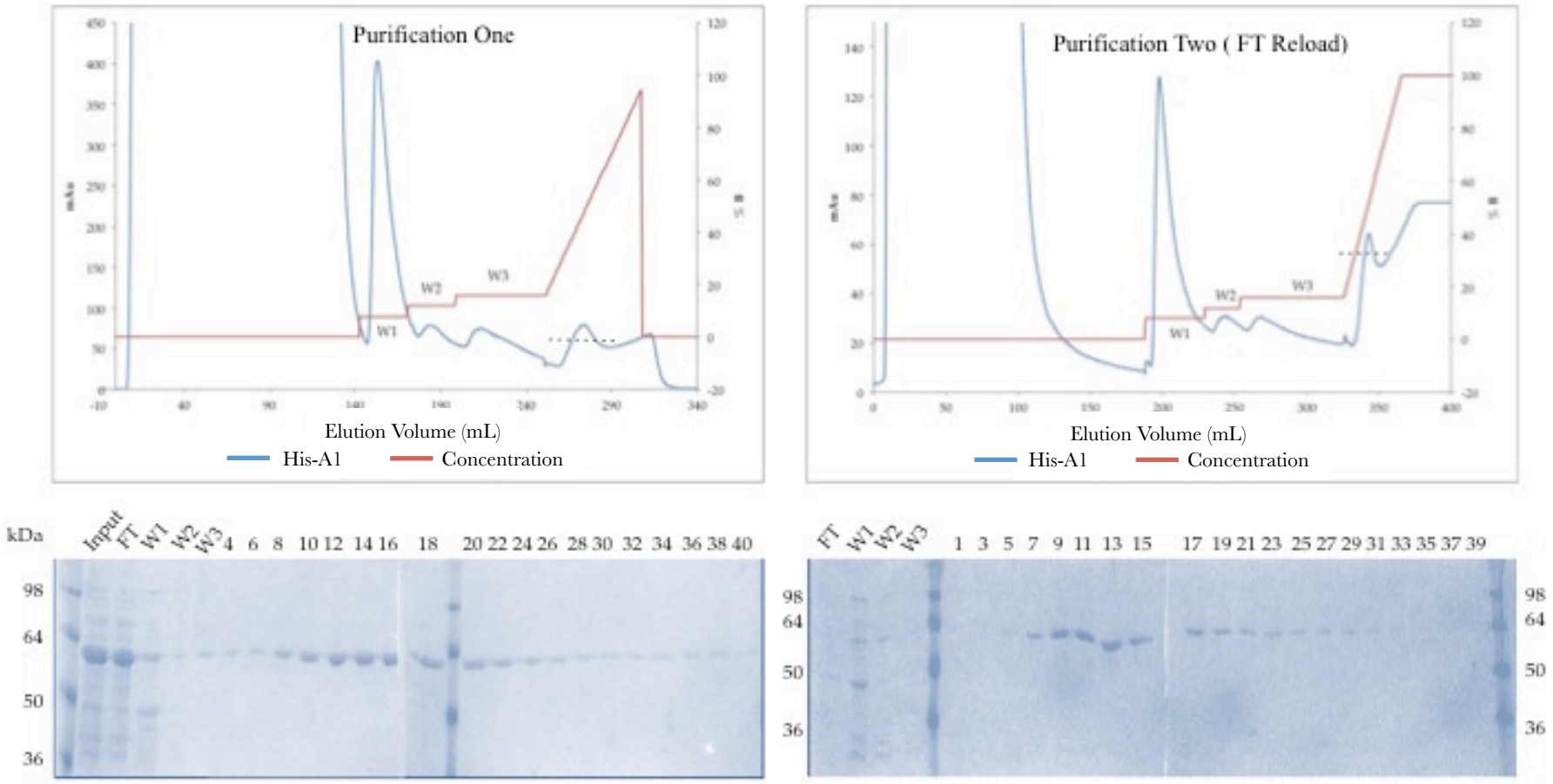


Figure 3.32 - optimised purification protocol for His-A1 , in the presence of Urea. The red line in the graph indicates the concentration of elution buffer (B) on a percentage scale, where 100% equates to 1M NaCl. FT – Flow through, W1 – Wash 1, W2 – Wash 2, W3 – Wash 3.

3.5 – Refolding

The list of variables with a refolding experiment can be extensive, and can include the speed at which the denaturant is removed, whether an intermediate buffer is used, whether the refolding is done in solution or with the protein bound to a fixed structure such as IMAC resin, and whether or not an external ligand is added to the solution to act as a refolding scaffold.

These variables all needed to be tested in order to improve the chances of achieving a correctly folded, stable, catalytically active protein. In order to test as many conditions as possible, increasing the likelihood of finding optimal refolding conditions, an in-house AZ screen was used. This was a compilation of 96 conditions including a variety of buffers, such as Tris, salts such as KCl, compounds such as L-Arginine, oxidants and reductants such as cystamine and cysteamine and the presence of chaotropes such as GuHCl. This screen was developed by AstraZeneca through both experimental discovery and literature searches for conditions that favoured refolding.

3.5.1 – Batch

The protein was thawed and a solution of either Adenosine 5′-(β,γ-imido)triphosphate (AMP-PNP) or AZ135XXX (a DNA-PKcs ATP analogue inhibitor developed within AZ that is still subject to confidentiality) was added to the protein prior to any refolding. This was done with the theory that a protein will refold more accurately around a known ligand than without any sort of scaffold. A sample of protein was also put into the screen without any ligand.

The refolding experiment was performed on a micro-scale, involving a rapid 50-fold dilution with mixing at the nL scale. 24 nL of protein at 5 mg/mL in 8 M urea was added to 1.2 µL of refold buffer containing the folding additive. This was performed three times (*apo*,

AMP-PNP, AZ135XXX). After the refold buffer was added to the 96 well plate a liquid handling robot added the protein to the drop, at which point the plate was sealed, agitated and left overnight to equilibrate. The plates were then imaged to determine the state of the solution after the overnight incubation. Only one observation was performed with the rationale being that most refolding will occur almost immediately due to the rapid 1 in 500 dilution. The overnight period was also included prior to observation to ensure equilibration will have occurred. A condition that was deemed favourable was one that had a clear drop. If any precipitation was visible this condition was excluded from further analysis. This methodology does not preclude the possibility of clear precipitate, which is a distinct possibility. However it served as a starting point that could be up-scaled, and this could then determine the likelihood of a clear precipitate forming rather than correctly folded protein.

Of the 96 conditions tested there were 3 clear optimal conditions that were taken forward, all of which required the use of a folding additive, either the AMP-PNP or AZ135XXX. When the protein was left to refold *apo* none of the conditions yielded clear drops. The three conditions chosen were:

- B1 (100 mM MES pH 5.8, 2 M Guanidinium Hydrochloride)
- B2 (100 mM MES pH 5.8, 1 M Guanidinium Hydrochloride)
- D9 (500 mM Tris pH 8.0, 2 M Guanidinium Hydrochloride)

All three conditions contain the chaotrope Guanidinium hydrochloride at either 1 or 2 M concentration. It is theorised, based on the work of Rashid *et al* (135), that between 0.5 and 1.5M guanidinium hydrochloride the protein actually exhibits a non-native intermediate state and is not fully denatured until approximately 4M. Therefore at these concentrations it can be supposed that the drop from 8M urea to 1-2M guanidine hydrochloride has brought the protein from a fully denatured state into an intermediate state, which is in a more stable position prior to dialysis, completing the refold into native structure. This will vary from

protein to protein and is something that will be tested in future work. This future work could involve taking the refolded protein and subjecting it to functional kinase assays in increasing concentrations of guanidinium hydrochloride to determine at which point the protein loses all functionality.

The drops for each of these conditions can be seen in figure 3.34 (A). Shown in figure 3.33 (B) is a drop for the condition G2 (100 mM MES pH 5.8, 900 mM Arginine). In this drop there is visible precipitation and this image has been included in the thesis as an example of a condition that was excluded due to not being suitable for induction of protein refolding.

Due to the fact that these three conditions were successful for both the AMP-PNP and AZ135XXX, the decision was made to use only the AMP-PNP for further experiments. This was due to confidentiality issues, and the fact that if a crystal structure were solved with this ligand in the binding site, publication of these results could be prohibited, whereas with a commercially available compound like AMP-PNP, this would not be an issue.

These experiments were then performed on a larger scale. The intermediate refolding buffer volume was increased to 5 mL and to maintain the same ratio of buffer : protein, an entire 100 µL aliquot of denatured A1 at 5 mg/mL with AMP-PNP was added, drop-wise, to the solution, whilst maintaining a rapid stirring. Upon addition of the entire 100 µL, the solution was rolled at 4 °C overnight in order to equilibrate. At this point the solution was centrifuged at 25,000 g for 1 hour and visible inspection was performed to detect the presence of any precipitate. The results of this experiment can be seen in figure 3.33 (C). This shows that these conditions remain favourable, even at a larger scale. At this point the 5 mL of intermediate protein solution was dialysed into a buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl, 5% Glycerol, 2 mM TCEP.

During the dialysis experiment, the protein that was refolding under the ‘B1’ conditions (100 mM MES pH 5.8, 2 M guanidinium hydrochloride) precipitated, which was

a result that was repeated several times, each time with the same outcome. This condition was therefore excluded from future experiments and the remaining 2 conditions were used. The next step was to determine the stability of the protein within the 2M guanidinium hydrochloride solution for storage purposes. HIS-A1 in both conditions was taken and stored overnight at either 4 °C or -20 °C. Determining optimal storage conditions allows for large scale synthesis and storage to ensure reproducibility in terms of subsequent experimentation. This storage however, is something that can only be utilised when the preceding steps have been optimised and also shown to be reproducible. Based on the results seen in figure 3.33 it is visible that storage of either intermediate at either condition is just as favourable/detrimental to the overall yield of protein.

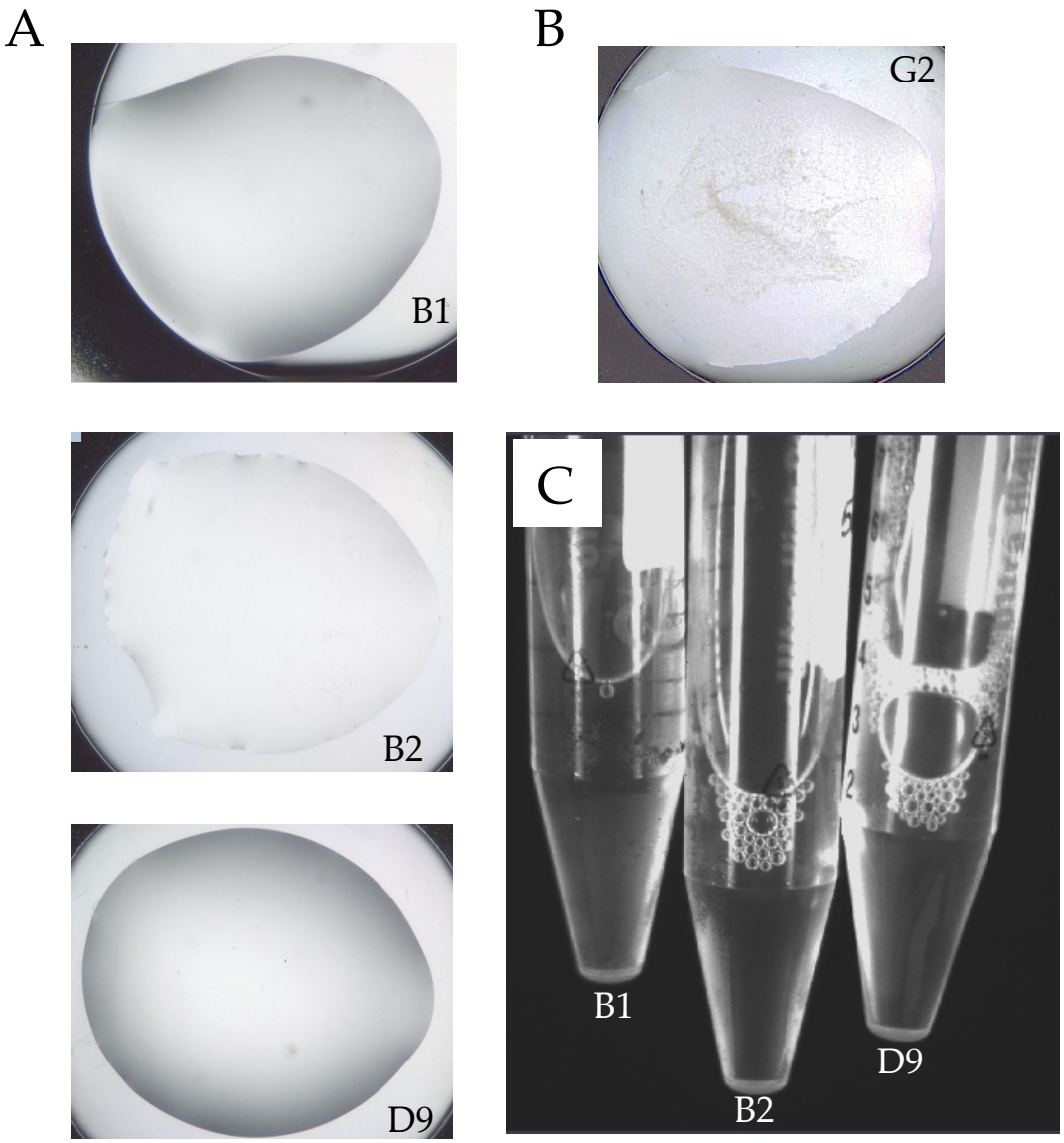


Figure 3.33 - (A) Small scale drops containing the refold buffer for the successful conditions. B1 contained 100 mM MES, pH 5.8, 2 M guanidinium hydrochloride. B2 contained 100 mM MES, pH 5.8, 1 M guanidinium hydrochloride and D9 contained 500 mM Tris, pH 8.0, 2 M guanidinium hydrochloride. (B) An example of a condition that leads to precipitation of the protein. G2 contained 100 mM MES, pH 5.8, 90 mM Arginine. (C) Larger scale refolding with the three successful conditions.

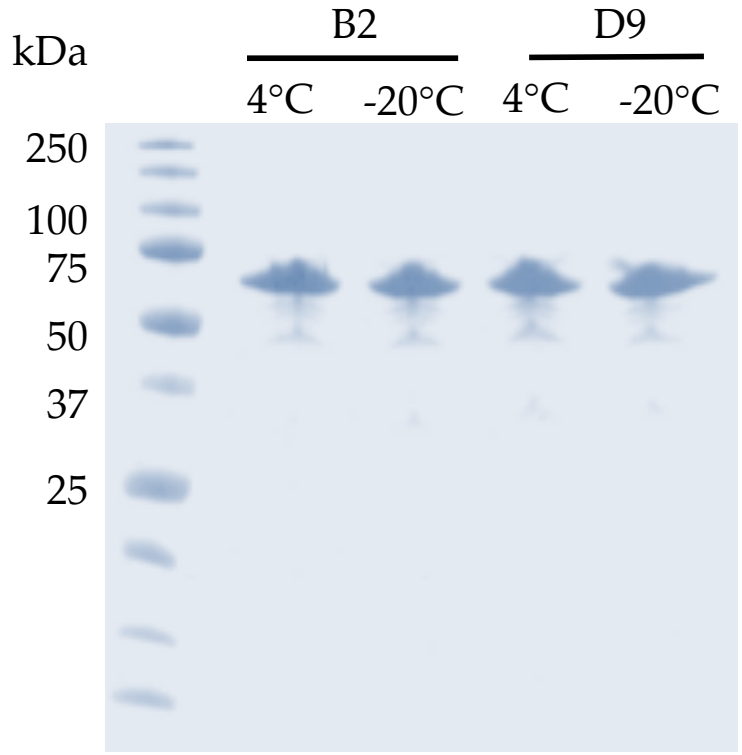


Figure 3.34 – 15% coomassie stained SDS-PAGE showing the results of testing both 4 °C and -20 °C as storage over a period of 24 hours. B2 contained 100 mM MES, pH 5.8, 1 M guanidinium hydrochloride and D9 contained 500 mM Tris, pH 8.0, 2 M guanidinium hydrochloride.

All four of these samples were then dialysed, in parallel, into the same buffer mentioned previously. For the B2 sample the final concentration of GuHCl after dialysis was ~0.1 pM and for the D9 sample it was ~ 0.2 pM. The contents of each of the membranes were then concentrated as much as possible. The results of this concentration can be seen in figure 3.36.

What can be seen is the first instance of a difference between the two conditions. The B2 conditions have precipitated considerably more than the D9 conditions, however both have dropped their yields relative to their inputs (seen in figure 3.35). A significant drop in yield is to be expected during a refolding experiment due to the protein precipitating or incorrectly folding but variability on this scale does show that further optimisation can be performed. One immediate experiment to test would be a slower dialysis once in the B2 and

D9 conditions as a more controlled removal of the chaotrope may improve protein folding thus increasing the yield.

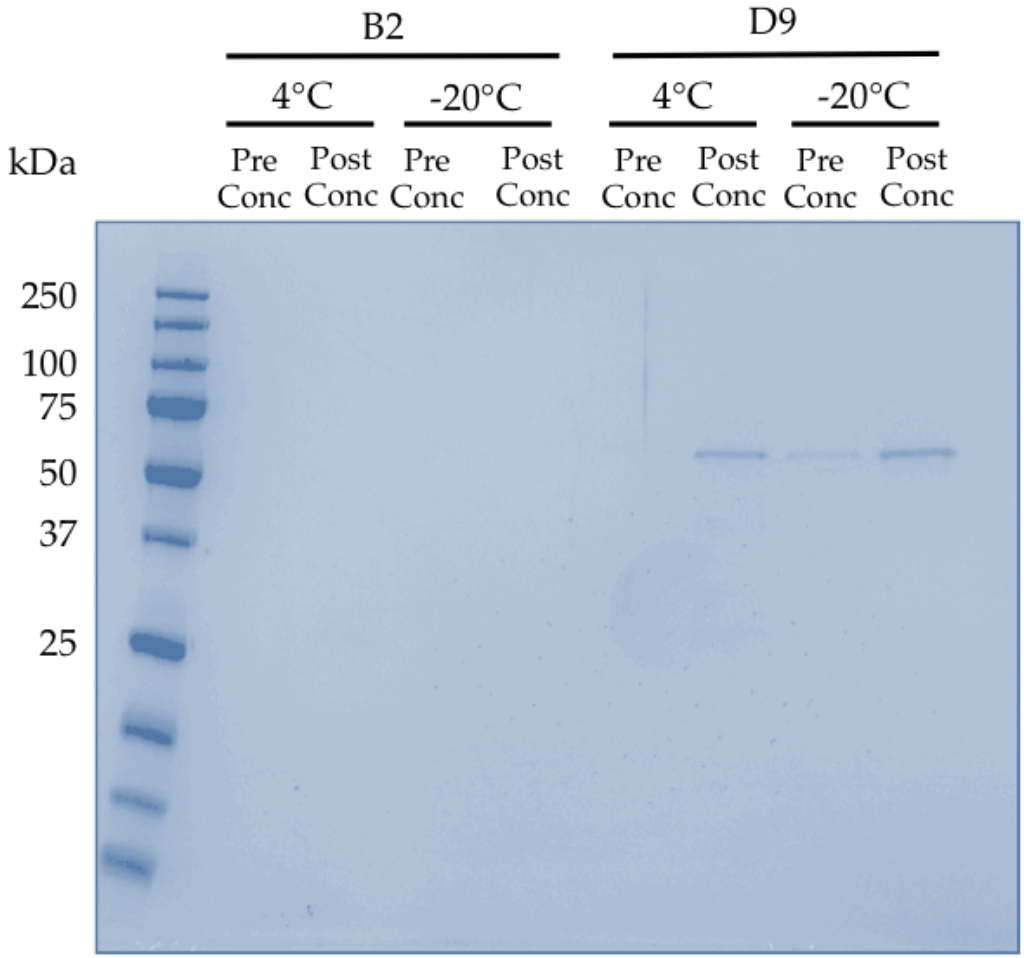


Figure 3.35 – SDS-PAGE showing the results after dialysing from the intermediate refold buffer into a stable buffer for further testing.

As a result of this experiment the protein refolded in the D9 conditions was passed over an equilibrated s200 10/300 GL column that had previously been calibrated using a combination of proteins found in the high and low molecular weight calibration kits (GE Healthcare).

The purpose of the calibration curve is to determine the elution volume for a globular protein of a particular molecular weight. Once this has been determined for each of the four proteins used in the calibration a graph can be plotted, showing the correlation between the

log of the molecular weight (logMW) of the protein and the partition coefficient (Kav). This then produces a calibration trace (green) seen the graph in figure 3.37 and by plotting a trendline between each of the points, the logMW of the unidentified protein peak, and therefore the MW, can be calculated.

$$Kav = \frac{(V_e - V_o)}{(V_c - V_o)}$$

Where V_e = the elution volume of the particular peak, V_o = the void volume of the column being used, and V_c = the volume of the column being used. The void volume for this column was determined experimentally by running a sample of blue dextran (2 MDa) down the column. It is too large to access any pores present on the matrix and therefore passed straight through the packed bed of the column. The value obtained was 7.8 mL (results not shown).

The volume of the column can be calculated as follows:

$$V_c = \pi \times r^2 \times h$$

$$V_c = \pi \times 5^2 \times 0.3$$

$$V_c = \underline{23.6 \text{ mL}}$$

As an example of the calculations performed, Conalbumin had an elution volume of 15.60 mL:

$$\text{Conalbumin Kav} = (15.6 - 7.8) / (23.6 - 7.8)$$

$$\text{Conalbumin Kav} = 7.80 / 15.8$$

$$\text{Conalbumin Kav} = \underline{0.493}$$

This value was then plotted against the logMW of Conalbumin (75 kDa) = 1.875

These calculations were performed for each of the proteins and the results of this, as well as the calibration curve can be seen in figure 3.36.

This was done as a control in order to determine if the product of the intermediate refold and dialysis folded into a globular protein of approximately the correct size. Using an s200 column such as this, the hydrodynamic radius can be determined and extrapolated, based on a calibration curve, to see if it behaves like a globular protein of the same size as is theorised.

The computed molecular weight of the protein is 60.4 kDa, and it is running around this height on the SDS-PAGE gel seen in figure 3.35.

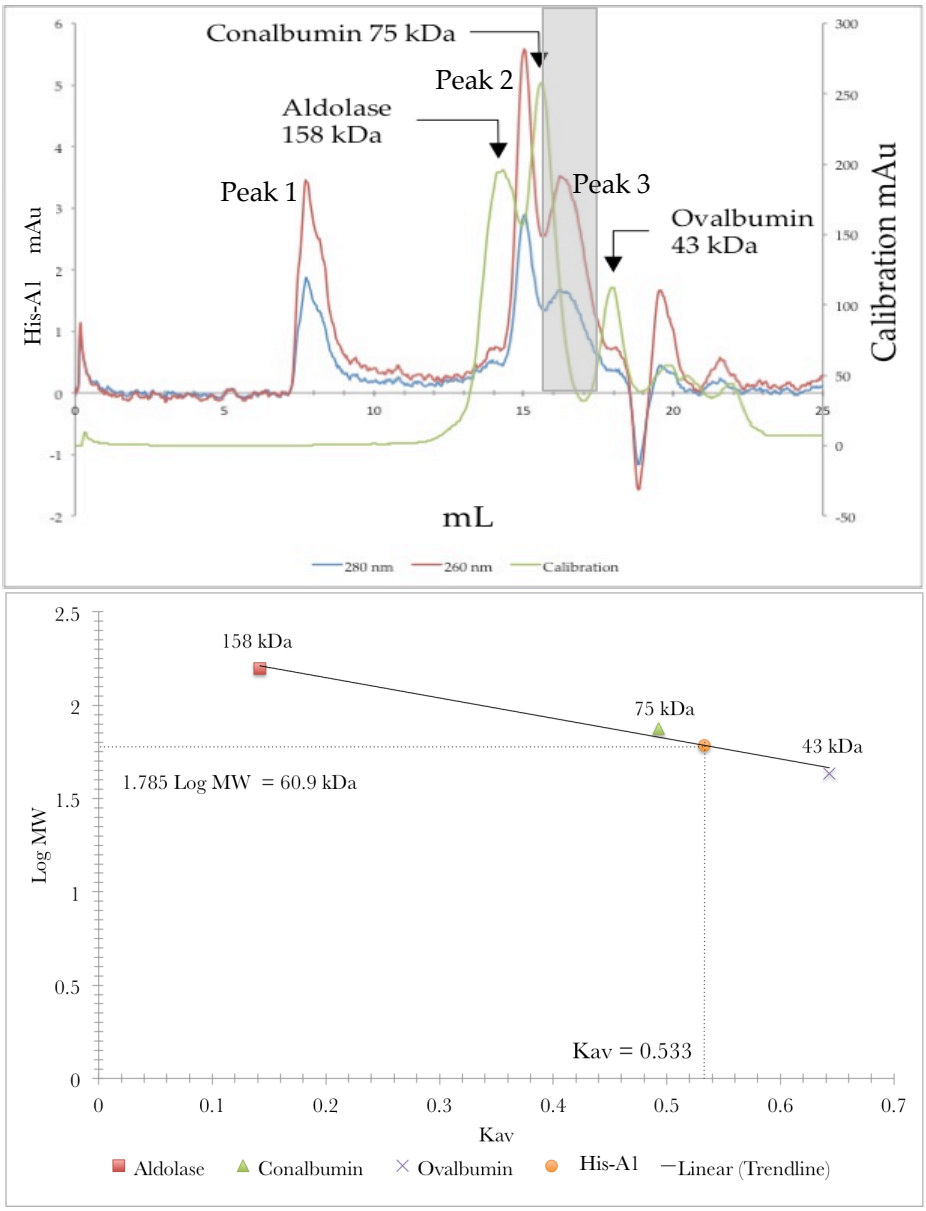


Figure 3.36- Size exclusion chromatography and associated analysis of the protein refolded in the D9 conditions. The green curve indicates the peaks for each of the calibration proteins used: Aldolase (158 kDa) at 14.3 mL; Conalbumin (75 kDa) at 15.6 mL; and Ovalbumin (43 kDa) at 17.9 mL. The grey box indicates the peak that was analysed to determine its approximate molecular weight. The line of best fit has an R value of 0.961.

Peak 1 seen in the trace is eluted at 7.8 mL, which has previously been determined to be the void volume of this column. This is most likely soluble aggregates that have formed as a result of the refolding process. There are also two peaks, 2 and 3, that merge with one another. Peak 2 at 15 mL, and peak 3 at 16.2 mL. If peak 3 is converted to a Kav value, it

can be plotted on a trend line generated from the calibration curve to determine the approximate MW of the protein responsible for that peak. In this instance the Kav for the peak 3 of HIS-A1, assumed to be the correct peak for the protein, was 0.533. Based on the trend line this generated an exponential factor of 1.785, which equates to a MW of 60.9 kDa. Due to the calculated MW being 60.4 kDa, it is safe to say that this is in fact folding in the correct orientation and has not been subject to degradation. The small amounts of protein eluted from the column however meant that further SDS-PAGE analysis could not be performed and therefore this is merely supposition. In future work, when the scale is increased this can be validated through SDS-PAGE analysis as well as western blotting.

The SEC trace shows the absorbance at 260 nm and 280 nm and what can be seen is that there still remains a 260 nm/280 nm ratio of ~1.8. This would indicate that there is nucleic acid contamination, which is unlikely, due to the stringent conditions the protein has passed through to get to this stage. An explanation to this result has not been experimentally determined, but one theory could be the presence of ATP, or more likely AMP-PNP which has a peak absorbance at 254nm, as this was required as a scaffold for the refolding of the protein.

The biggest factor however is the yield, in that the peak size for the 280 nm trace is around 1.8 mAu. Dilution when running a size exclusion column is inevitable, but starting with 5 mg/mL, prior to refolding, has lead to an almost non-existent peak of approximately 0.1 mg/mL. Although theoretically this amount of protein is enough for some small-scale experimentation to be performed, there is not enough for extensive screens as well as the necessary biophysical characterisation required.

These values are representative of several attempts at this refolding experiment. Due to this the next logical step was to optimise the protocol and attempt to increase the final yield of solubilised, correctly folded protein. Initial optimisation involved replacing the dialysis step as it was during this step that the greatest drop in yield occurred and there could

theoretically be more effective ways to remove the intermediate denaturant. The first alteration was to directly replace this step of the refold with an on-column refold step. Whereby the drop-wise dilution from 8 M urea to the D9 solution was performed. At this point the ~50 mL of D9 containing the target protein was loaded onto a 5 mL Histrap (GE Healthcare) column. Once loaded the concentration of GuHCl was reduced from 2 M to 0 over 20 column volumes. At this point the protein was eluted with the dialysis buffer mentioned previously, supplemented with 500 mM imidazole. Although theoretically this method would improve the likelihood of generating higher yields, in actual terms it was not reproducible, and after a large number of attempts changing a multitude of variables this particular method was deemed unsuccessful.

Time constraints stopped any alternative attempts at optimising this refolding step to attempt to increase the final yields of protein after the refold. Although this is the case, the work performed to get to this stage sets up a platform for future work to enable either larger scale experiments or optimisation to generate refolded, soluble, catalytically active DNA-PKcs kinase domain for the purposes of biophysical characterisation and X-ray crystallographic studies.

Chapter 4

Flap Endonuclease 1

This chapter of the thesis will go into detail on the determination of the X-ray crystal structure for FEN 1 from *Pyrococcus abyssi*, and will make comparisons with various other existing crystal structures. Results will also be discussed from initial high-pressure crystallography experiments performed at the ESRF in Grenoble, France.

4.1 FEN 1 Structure at Ambient Pressure

4.1.1 – Protein Production & Purification

4.1.1.1 - Construct information

The construct used throughout this study was for full-length *Pab* FEN 1 spanning from M1 to P343 with no tags present on either termini. A schematic of the construct, and its domain organisation is in figure 4.1.

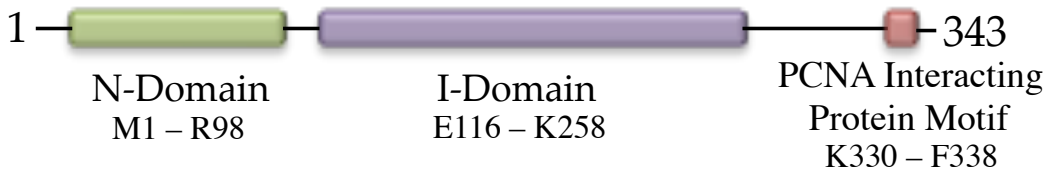


Figure 4.1 – Domain organisation of *Pab* FEN 1. The N-terminal domain (N-Domain) spans M1 to R98. The internal domain (I-Domain) spans from E116 to K258. The PCNA interacting protein motif (PIP), spans from K330 to F338.

The N terminal domain (N-domain) and the internal domain (I-domain) are both globular domains, the latter of which contains a conserved region that harbours the active site. It is also the region that forms a helical archway that is thought to play an important role in DNA binding (136, 137). The construct used was cloned by Rashed-Al-Naber into the pET3aTr overexpression vector (138)

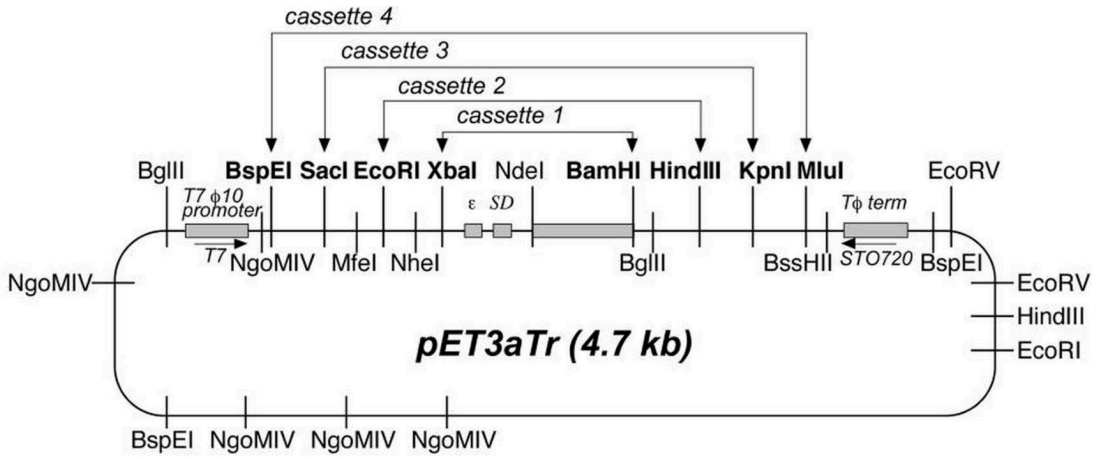


Figure 4.2 - pET3aTr vector map (138). Cassettes 1-4 allow the coexpression of up to 4 genes in the one expression plasmid, which relates to projects other than those mentioned in this thesis. Cassette 1 was the only one used in this study for the expression of FEN 1 from *Pab*.

4.1.1.2 - Expression conditions

Athanasios Adamopoulos (AA) and Rashed-Al-Naber (RAN) optimised the over-expression conditions. The conditions involved using transformed BL21* cells, induced and grown overnight. The SDS-PAGE gel shown in figure 4.3 shows both the results of this induction as well as the subsequent lysis.

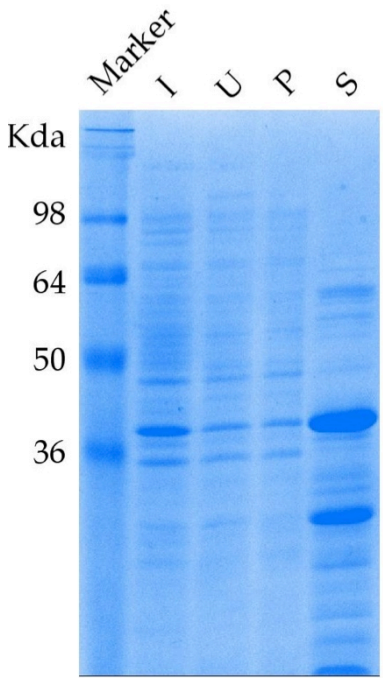
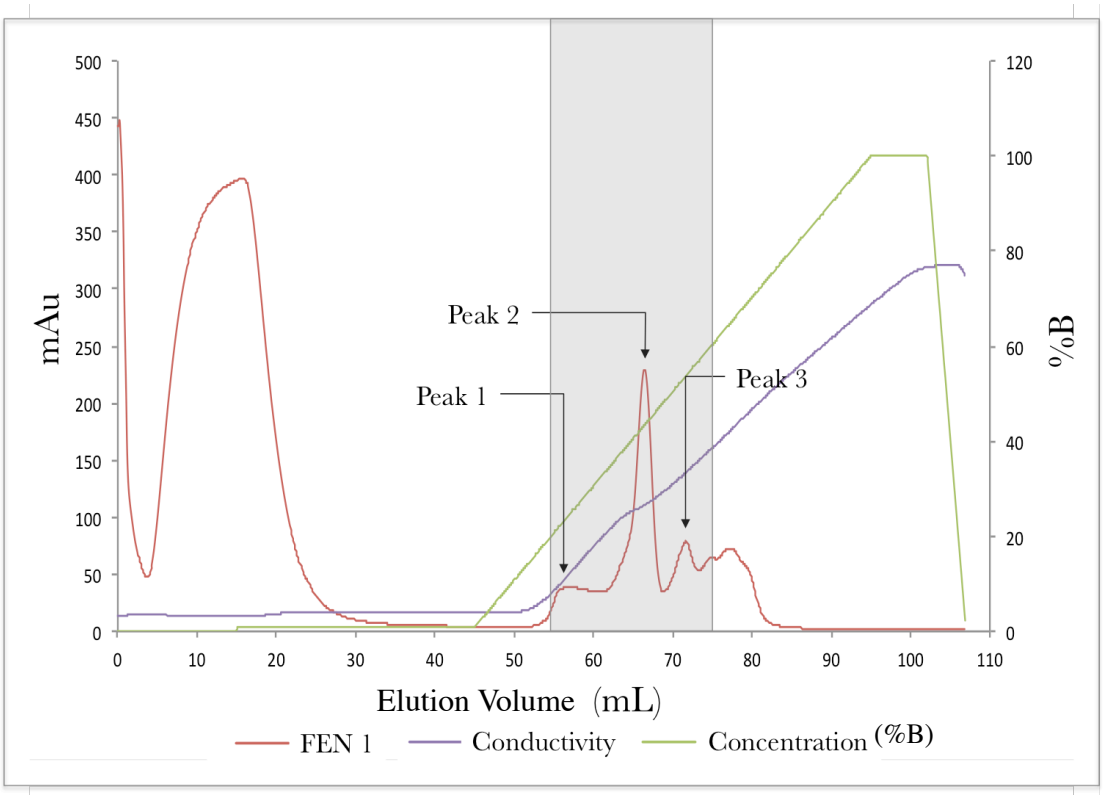


Figure 4.3 – 15 % SDS-PAGE gel showing successful induction and lysis of FEN 1

4.1.1.3 - Protein isolation

An initial purification protocol was set up by AA, but was subjected to optimisation. A cell pellet was removed from storage at -80°C and fully resuspended in 20 mL lysis buffer (20 mM HEPES pH 7.5, 300 mM NaCl) with protease inhibitors (Roche). It was then sonicated in 3 x 30-second bursts with 30-second intervals. After sonication the lysate was incubated for 20 mins in a water bath at 70°C in order to denature the host *E.coli* proteins whilst keeping the target FEN 1 protein intact.

The heated lysate was then centrifuged at 25,000g for 1 h separating the pellet and supernatant. The supernatant containing the soluble FEN 1 was then diluted with 180 mL 20 mM HEPES pH 7.5, reducing the salt concentration from 300 mM to 30 mM. At this point the solution was loaded onto a 1 mL HiTrap Heparin column (GE Healthcare). The results of this can be seen in figure 4.4:



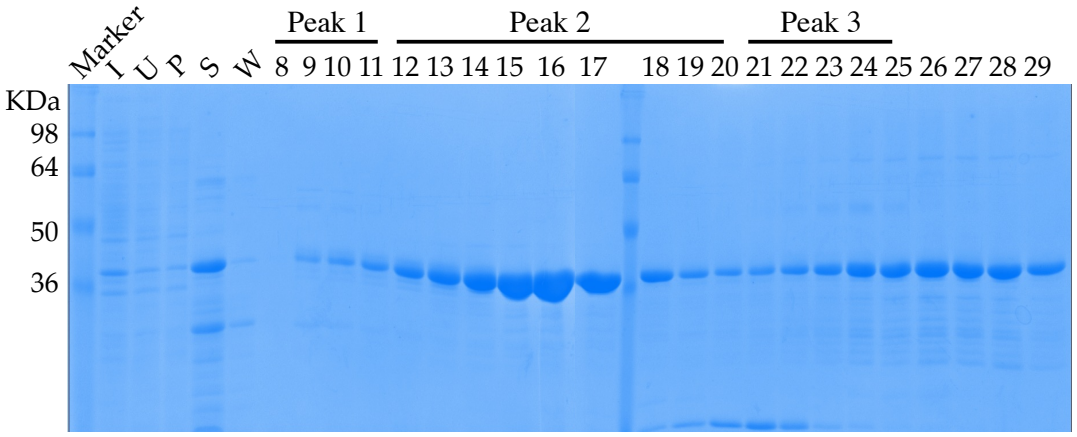


Figure 4.4 - Chromatographic trace of the FEN 1 purification using a 1 mL Heparin column as well as the accompanying gel showing the contents of fractions 8 to 29, highlighted in grey in the trace. The green line in the graph indicates the concentration of elution buffer (B) on a percentage scale, where 100% equates to 1M NaCl. Fractions 9 – 21 contained pure FEN 1. They were pooled and concentrated to a final volume of 500 µL.

Fractions 21 - 29 also contained the same protein, however the levels of contaminants are higher therefore these fractions were discarded. The concentrated sample was loaded onto a pre-equilibrated size exclusion chromatography column Superdex 200 10/300 GL (GE Healthcare) (Figure 4.5).

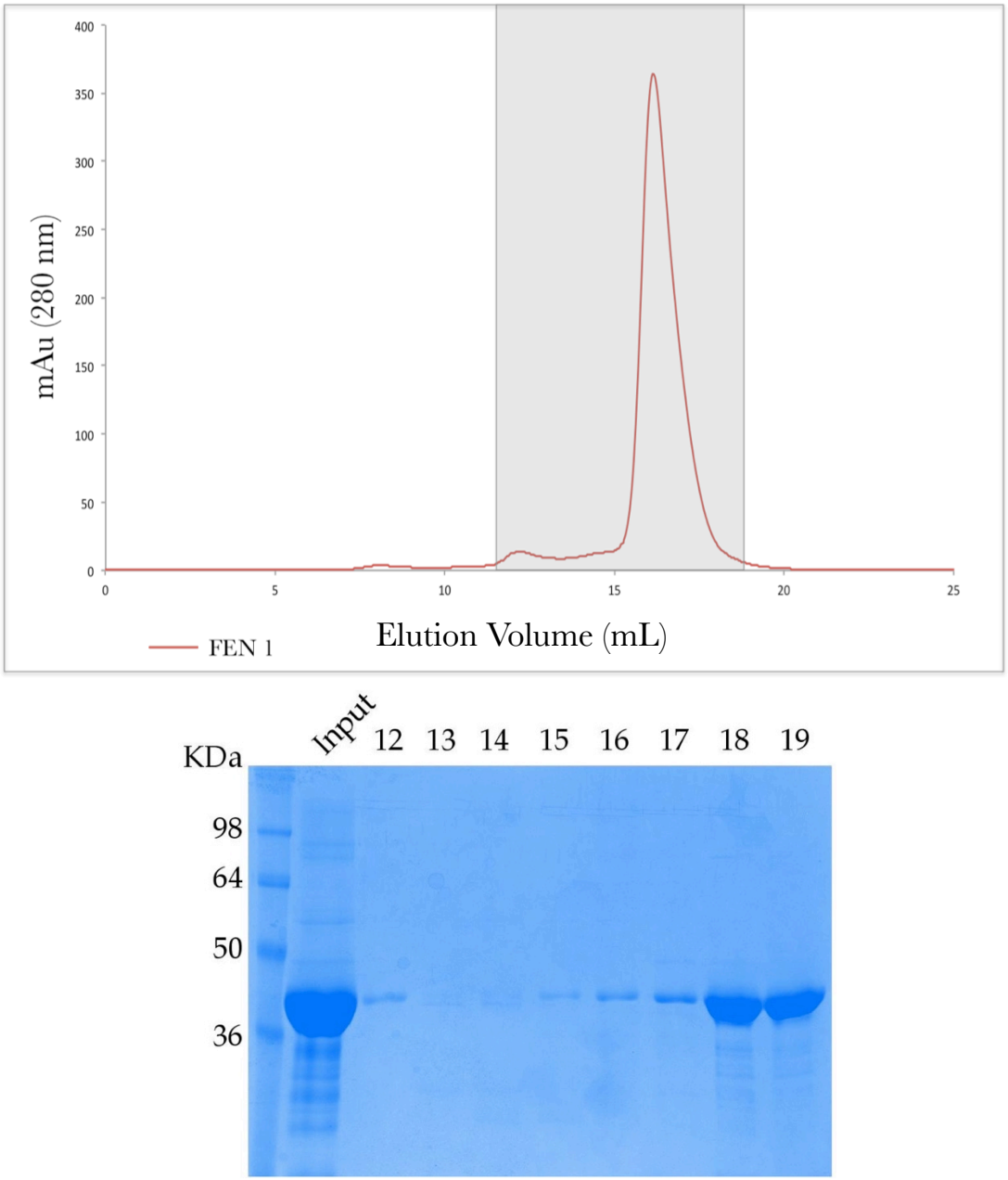


Figure 4.5 - Chromatographic trace and associated 15% SDS-PAGE gel for the size exclusion chromatography experiment purifying FEN 1. The size exclusion step was performed in 20 mM HEPES pH 7.5, 400 mM NaCl. Fractions 15 – 19 were pooled.

The NaCl concentration was set at 400 mM based on the concentration of salt during elution of the main peak collected in the heparin purification in figure 4.4. Pooled fractions were concentrated and the final concentration of the sample was tested using a NanoDrop 2000 UV-Vis spectrophotometer (Thermo Scientific) measuring the absorbance at 280 nm as well as the 260 nm/280 nm ratio detecting potential nucleic acid contamination.

Using the online software ProtParam from the Expasy Suite (123), the extinction coefficient of the protein was calculated as 35870. This was then used to calculate the concentration of the protein, which after the size exclusion step was 18 mg/mL. The 260 nm/280 nm was measured to be 0.56. This indicates that the sample was free from nucleic acid contamination. After determining the concentration and purity, the sample was aliquoted and flash frozen by being dropped into liquid N_2, and then stored in the -80°C freezer until further use.

4.1.2 – Biophysical Characterisation

4.1.2.1 – Size exclusion chromatography elution analysis

The Superdex s200 10/300 GL column was calibrated according to the manufacturers guidelines in 20 mM HEPES pH 7.5, 400 mM NaCl, using four proteins from the low molecular weight calibration kit (GE Healthcare). This calibration curve was then used to determine the partition coefficient for FEN 1 in order to determine the molecular weight of the protein being eluted. The method for this is discussed in chapter 3.

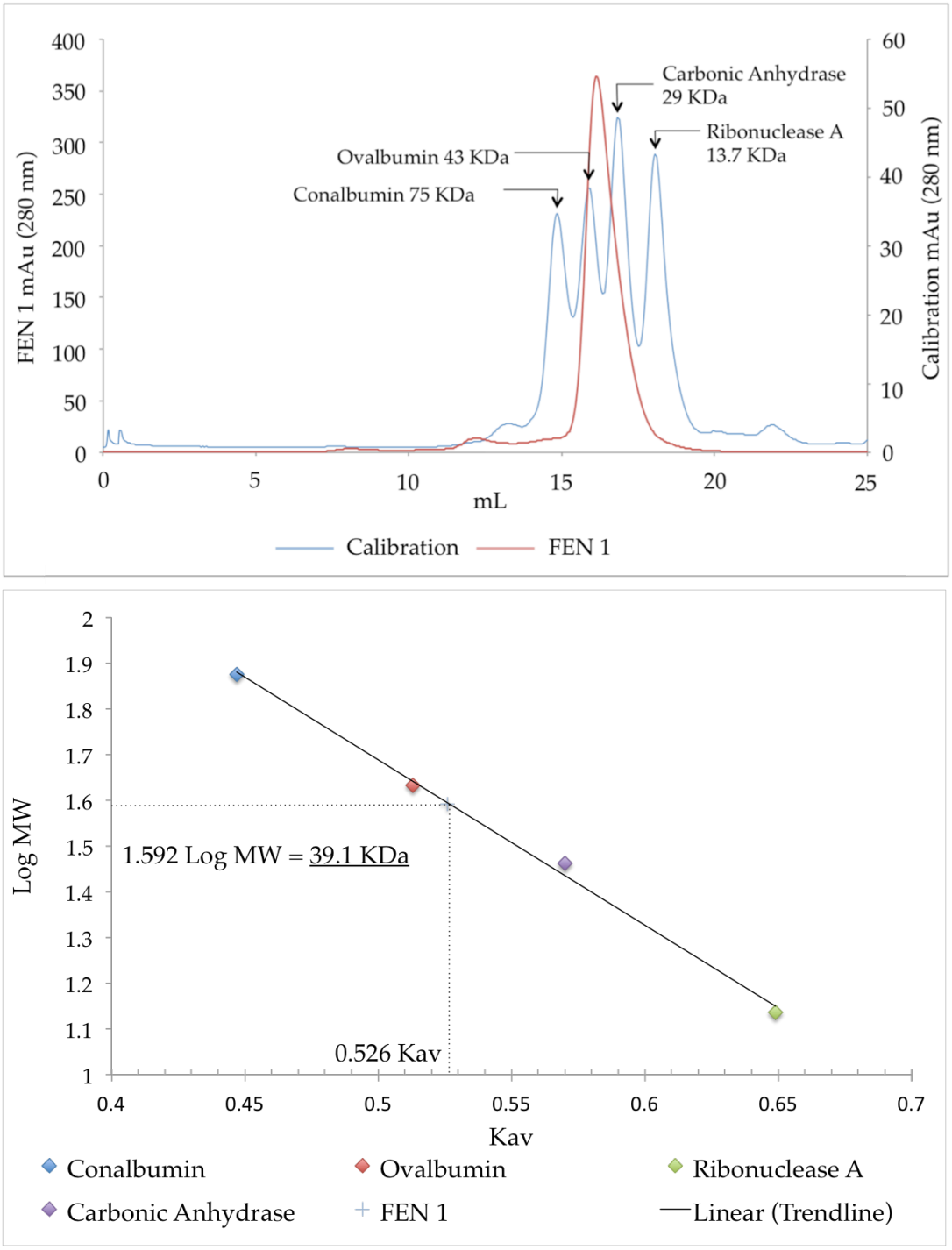


Figure 4.6 - Calibration curve containing Conalbumin, Ovalbumin, Carbonic Anhydrase and Ribonuclease A, overlaid with the SEC trace shown in figure 4.5. Below is a graph showing LogMW against Kav. The line of best fit has an R value of 0.910

Using an elution volume for the FEN 1 sample of 16.125 mL, and the equation mentioned in chapter 3 previously, a Kav value of 0.526 is obtained. Plotting this on the trendline seen in figure 4.6 shows a LogMW value of 1.592.

$$10^{1.592} = 39.1 \text{ kDa.}$$

This shows that the protein purified during the size exclusion chromatography, behaves in the same manner as a globular protein approximately 39.1 kDa in size.

Due to the fact that the calculated MW of FEN 1 is 38.9 kDa this seems to indicate that the FEN 1 protein being purified is both globular and monomeric at this concentration. A sample of the protein was then subjected to further characterisation. If however the true oligomeric state of FEN 1 is in fact a dimer, the 400mM salt could potentially disrupt any potential electrostatic interactions. This could therefore artificially suggest a monomeric, rather than dimeric state.

4.1.2.2 – Dynamic Light Scattering.

Results discussed in the previous section indicate that the protein being purified is in fact a monomer. To strengthen the evidence of the oligomeric state of the protein dynamic light scattering (DLS) was performed in order to determine both the size of the molecule as well as the size distribution of the molecules within the solution.

The technique makes use of the phenomenon whereby particles in solution will experience random motion known as Brownian motion. It is caused by the interaction of the protein with surrounding molecules affecting its trajectory. It is possible to model this motion mathematically using the Stokes-Einstein equation:

$$D_h = \frac{k_B T}{3\pi\eta D_t}$$

This equation shows the relationship between the particle size (hydrodynamic diameter (D_h)) and the values obtained for the diffusion coefficient (D_t). The k_B is the Boltzmann constant ($1.3806 \times 10^{-23} JK^{-1}$) which is a constant indicating the relationship between energy (J) and temperature in kelvin (K). T is the temperature that is set manually. In the case of these experiments it was set at 25°C (298.15 °K). η relates to the dynamic viscosity of the solution. The buffer constituents were loaded into Malvern Instruments software and it determined the viscosity based on theoretical values.

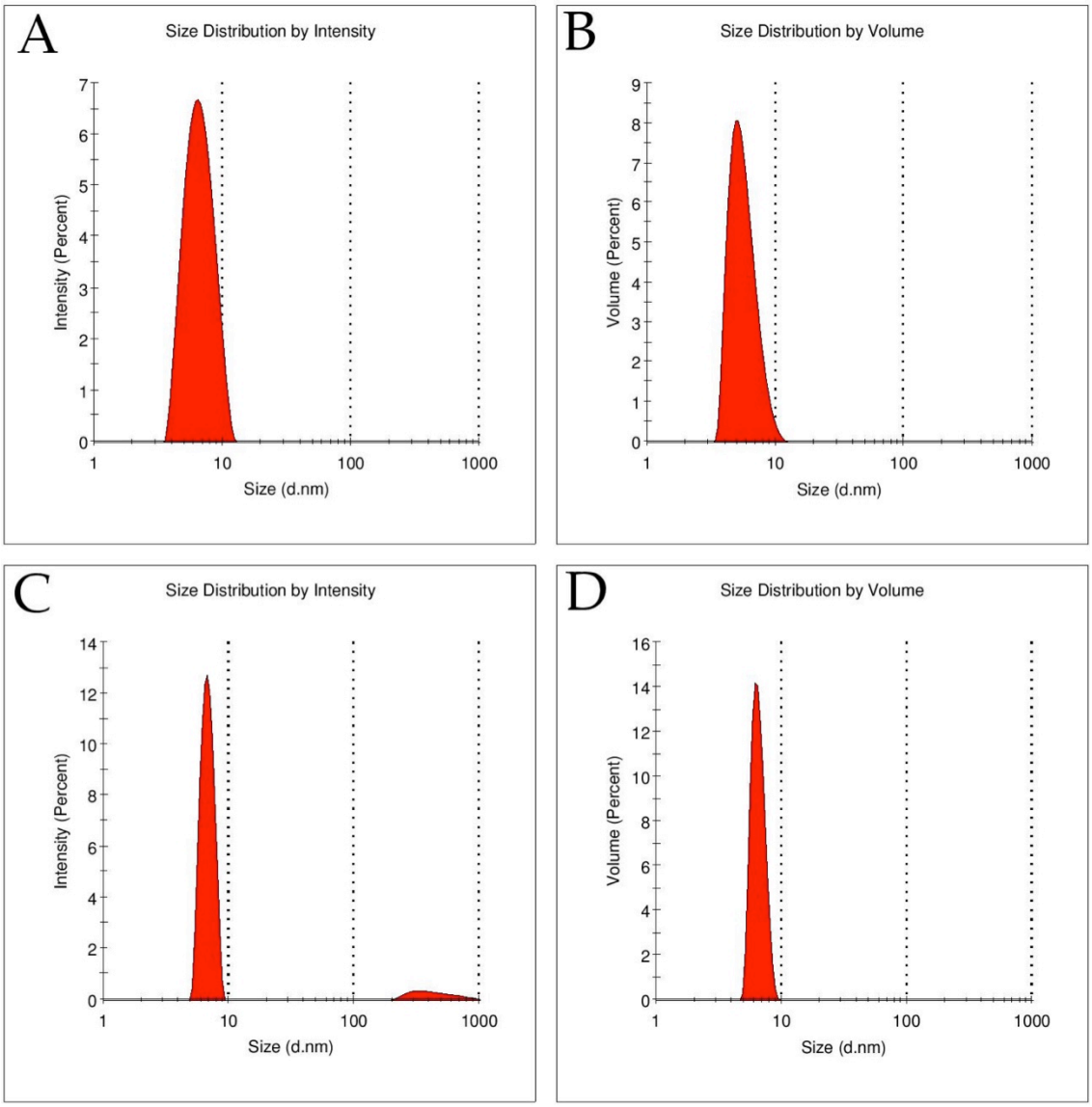


Figure 4.7 – Averaged DLS data for FEN 1 at two concentrations. (A) Shows size by intensity and (B) shows size by volume of FEN 1 at 1 mg/mL. (C) Shows the size by intensity and (D) shows size by volume of FEN 1 at 18 mg/mL.

When analysing these results, (B) shows a smaller hydrodynamic diameter than any of the other three graphs, even comparing it directly to the FEN 1 at 1 mg/mL showing size against intensity (A). This shows a shift to the right indicating a larger hydrodynamic diameter. According to the mass distribution results the mode hydrodynamic diameter is 5.12 nm (StD 1.35 nm) with an estimated molecular weight of 38.1 kDa. The mode is chosen as a means of averaging rather than the mean simply due to the fact that it is not affected as

much by larger molecules. The intensity of light scattering does not increase linearly with an increase in mass, a larger molecule will scatter considerably more than a smaller one (i.e. a dimer will have > 2x intensity than a monomer).

According to the intensity distribution the mode diameter is 6.45 nm (StD 1.67 nm) with an estimated molecular weight of 57 kDa. This suggests that there is a small amount of dimer present but due to the mass results indicating solely monomer, the levels of dimer must be very small.

The data represented in panels (C) and (D) correlate to one another a lot more than (A) and (B). This is most likely due to the increased prevalence of a higher order structure, most likely a dimer, at 18 mg/mL than at 1 mg/mL. The mode hydrodynamic diameter values for intensity was 6.75 nm, for mass it was 6.16 nm (StD 0.81 nm and 0.79 nm respectively) showing again that these results correlate with one another more than (A) and (B). The estimated molecular weight by intensity was 58.8 kDa and by mass was 53.4 kDa. This could suggest that within the solution, at this concentration, there is a greater presence of dimer relative to the monomer, and higher than the values seen at 1 mg/mL which could indicate a concentration dependent oligomerisation.

4.1.3 – Crystallisation

4.1.3.1 – FEN 1 crystallisation

An aliquot of FEN 1 was used for the purposes of growing crystals in order to determine an X-ray crystal structure of the protein. Crystallisation trials were performed with FEN 1 at a concentration of 3.1 mg/mL, with a 260 nm/280 nm ratio of 0.58. Using the Gryphon LCP liquid handling robot (Art Robbins Instruments), in a sitting drop conformation, 65 µL precipitant was placed into the reservoir. The drop in the well consisted of 100 nL of precipitant and 100 nL of protein solution. The first commercial screen to be tested was the Structure Screen 1 & 2 (Molecular Dimensions). The plates were stored at

18°C and were checked periodically every 24 hours for approximately 1 week, at which point they were checked once per week.

After 48 hours the most promising conditions to yield crystals were A4 (100 mM sodium acetate, pH 4.6, 2 M sodium formate), B7 (200 mM zinc acetate, 100 mM sodium cacodylate, pH 6.5, 18% w/v PEG 8000) and C9 (200 mM magnesium chloride, 100 mM Tris, pH 8.5, 30% w/v PEG 4000). A typical example of crystals grown in these conditions can be seen in figure 4.8:

Figure 4.8 – Grown in 100 mM sodium acetate, pH 4.6, 2 M sodium formate (A4). The crystals appear to be small, nucleated needles that were indicative of crystals grown in the other conditions tested. Each crystal is roughly 10-20 μm long.

The crystals shown in this figure, although promising, are small and highly nucleated. Neither the crystal size nor the shape are suitable for data collection but are an excellent

starting point for optimisation. The optimisation began by varying the concentrations of the precipitant and buffer, a schematic of which can be seen in figure 4.9.

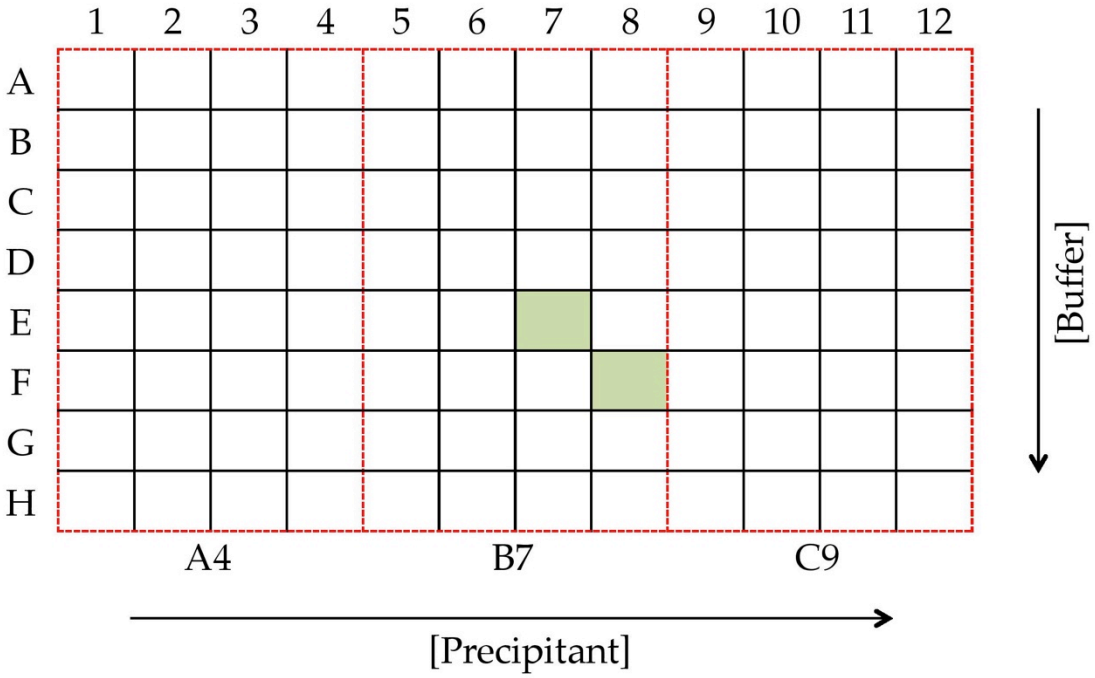


Figure 4.9 - Schematic of the larger scale crystallisation screens expanding on A4 (100 mM sodium acetate, pH 4.6, 2 M sodium formate), B7 (200 mM zinc acetate, 100 mM sodium cacodylate, pH 6.5, 18% w/v PEG 8000) and C9 (200 mM magnesium chloride, 100 mM Tris, pH 8.5, 30% w/v PEG 4000) conditions varying both the precipitant and buffer concentrations above and below screen levels. Highlighted in green are the two most promising results to come out of this round of optimisation. These conditions are: E7 - 200mM zinc acetate, 100mM sodium cacodylate, 20% w/v PEG 8000; F8 – 200mM zinc acetate, 200mM sodium cacodylate, 22% PEG 8000.

After several more iterations and improvements around the conditions a final set of conditions were obtained. These conditions were used for larger scale experiments using 24-well Linbro plates (Hampton Research), which made use of larger reservoir and drop volumes as well as a hanging drop vapour diffusion method rather than sitting drop used previously. The final conditions were 200 mM zinc acetate, 100 mM sodium cacodylate, 11% w/v PEG 8000. FEN 1 at an initial concentration of 2.4 mg/mL was added to the drop and the crystals were left at 18°C to allow crystals to grow. After 14 days crystals had formed

that were large enough to be used for data collection. An image of these crystals can be found in figure 4.10:

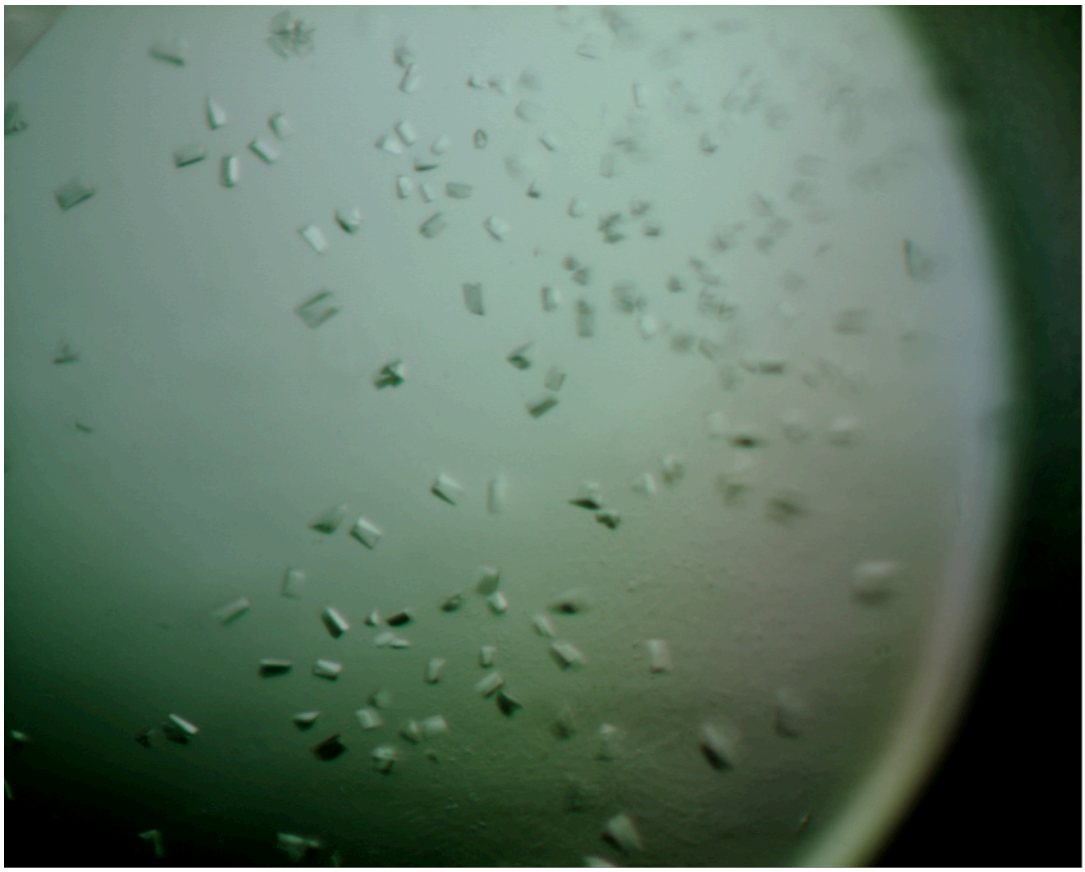


Figure 4.10 - Crystals grown after 14 days in 200 mM zinc acetate, pH 4.6, 100 mM sodium cacodylate, 11% PEG 8000, conditions obtained through several rounds of refinement. The crystals have an approximate length of between 40-50 µm.

Crystals were used for data collection at beamline I04-1 at the Diamond Light Source Synchrotron. The beam-line makes use of cryogenic temperatures during the data collection to reduce X-ray associated free radical formation, which can damage the crystal and reduce the quality of the data. It can also improve the X-ray diffraction due to lower levels of thermal motion. The crystals shown in figure 4.10 were harvested using 0.5mm loops and cryo-cooled in liquid N_2.

4.1.4 – Crystallography

4.1.4.1 – Space Group Determination

The crystals seen in figure 4.10 diffracted to 2.27Å resolution and a typical diffraction image can be seen in figure 4.11. The data were collated and processed using MOSFLM (113) in the space group $C222_1$. The reflections were scaled using SCALA (109).

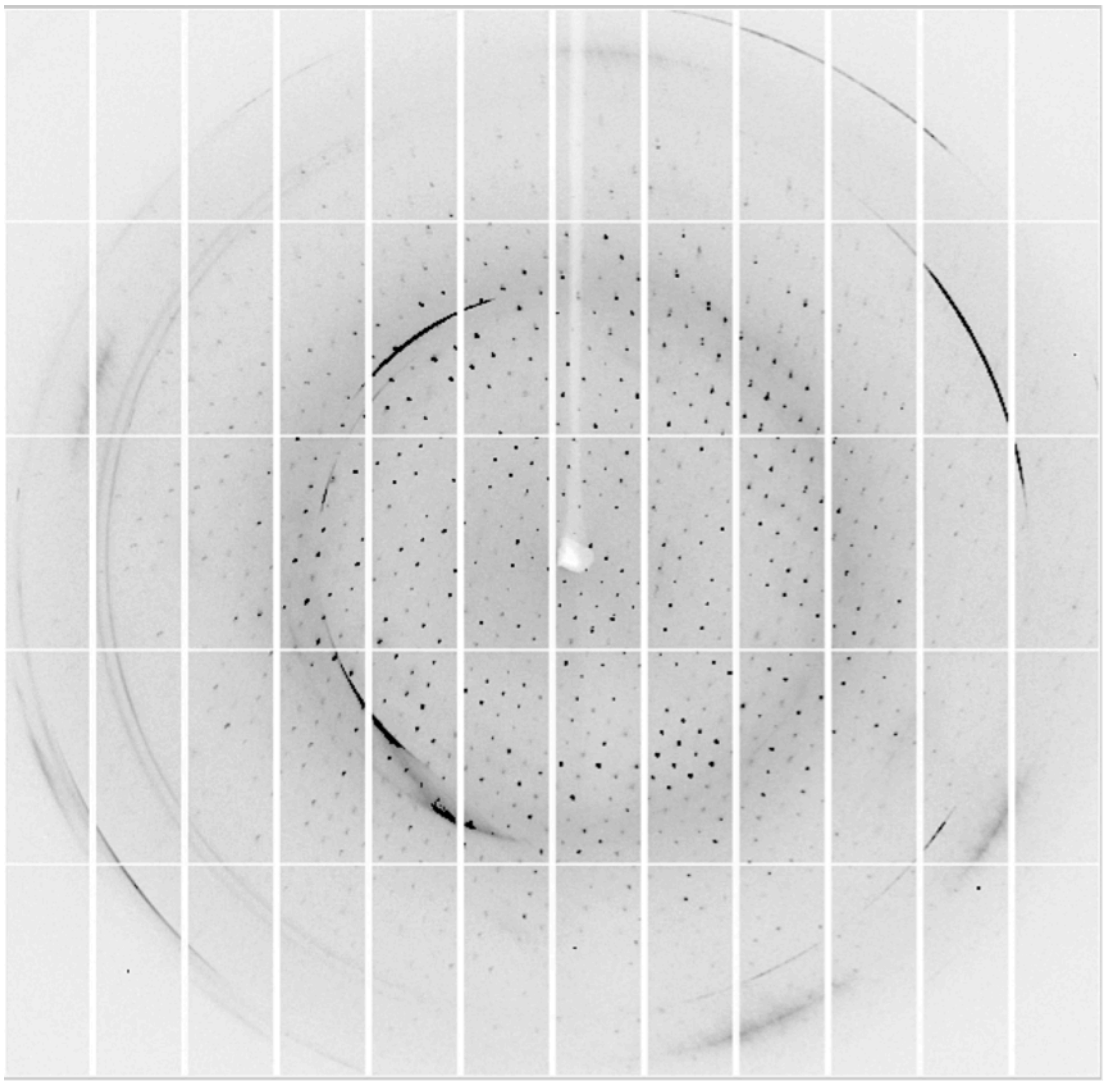


Figure 4.11 - Typical X-ray diffraction pattern from a crystal shown in figure 4.10. The data diffracted to 2.27 Å

Data collection	FEN-1
Space group	C2221
Cell dimensions	
a, *b*, *c* (Å)	65.23, 86.40, 238.99
α, β, γ (°)	90, 90, 90
Resolution Range (Å)	50.26 - 2.27
R_{merge}	0.143 (0.317)
$I/\sigma I$	6.0 (2.7)
Solvent Content (%)	43.1
Molecules / AU	2
Completeness (%)	96.1 (99.8)
Redundancy	4.2
Refinement	
Resolution (Å)	2.27
No. Reflections	30549
R_{work} / R_{free}	0.2160/0.2755
No. Atoms	4507
Protein	4345
Water	162
B-factors **(Å^2)**	
Protein	14.38
Water	16.62
R.m.s deviations	
Bond lengths (Å)	0.008
Bond angles (°)	1.143
Ramachandran Plot	
Preferred	535 (98.7%)
Allowed	542 (100%)
Outliers	0 (0%)

Table 4.1 – Crystallographic statistics for *Pab* FEN 1

The R_{merge} (Linear merging R-value) in the outer shell at this resolution was 0.317. The Rmerge is a measure of reliability of the data, comparing intensity measurements for all

of the reflections to an averaged intensity value, however this does not take multiplicity into consideration in so far as the more a particular reflection is measured, the more accurate the average intensity for that reflection will become. The I/σI, a statistic used to define a potential resolution limit, is a measure of how well spots can be defined compared to background. This signal to noise parameter can show when data has become too weak to accurately define its intensity. The value obtained for the I/σI in the outer shell at this resolution was 2.7, where an acceptable cut off for this is approximately 2.0.

4.1.4.2 – Molecular Replacement of FEN 1

Molecular replacement was performed using FEN 1 from *Sulfolobus solfataricus (Sso)* (PDB code: 2IZO) (139). Figure 4.12 shows a sequence alignment, performed with T-Coffee (138) between FEN 1 from *Sso* and from *Pab*. They share 57% sequence identity and it has been proposed that a solution can be derived using a model with as low as 30% sequence identity compared with the unknown protein (140).

```
Sso            1 MDLADLVKDVKRELSFSELKGKRVSIDGYNALYQFLAAIRQPDGTPLMDS
Pab            1 MGVPIGELIPRKEIELENLYGKKIAIDALNAIYQFLSTIRQRDGTPLMDS
consensus      1 * .       ..*.    * **.. **. **.****  *** ********

Sso           51 QGRVTSHLSGLFYRTINILEEGVIPIYVFDGKPPEQKSEELERRRKAKEE
Pab           51 KGRITSHLSGLFYRTINLMEAGIKPVYVFDGKPPAFKKKELEKRREAREE
consensus     51  **.*************..* *. *.********  *  ***.** *.**

Sso          101 AERKLERAKSEGKIEELRKYSQAILRLSNIMVEESKKLLRAMGIPIVQAP
Pab          101 AEIKWKEALAKGDIEEARKYAQRATKVNEMLIEDAKKLLQLMGIPIVQAP
consensus    101 ** *   *   * *** *** *   ..  ...*. ****  *********

Sso          151 SEGEAEAAYLNKLGLSWAAASQDYDAILFGAKRLVRNLTITGKRKLPNKD
Pab          151 SEGEAQAAYMAGKGDVYASASQDYDSLLFGTPRLVRNLTITGKRKMPGKD
consensus    151 ***** ***.   *  .* ****** .***  *************.* **

Sso          201 VYVEIKPELIETEILLKKLGITREQLIDIGILIGTDYNPDGIRGIGPERA
Pab          201 IYVEIKPELIVLEEVLKELKITREKLIELAILVGTDYNPGGIKGIGPKKA
consensus    201 .*********  * .** * **** **...**.****** **.**** .*

Sso          251 LKIIKKYGKIEKAMEYGEISKKDINFNIDEIRGLFLNPQVVKPEEALDLN
Pab          251 LEIVK-YSKDPLA-KF----QRQSDVDLYAIKEFFLNPPTTD-DYSLKWK
consensus    251 * *.*.* *   *. ..... .     .  *.  ****    ..  *

Sso          301 EPNGEDIINILVYEHNFSEERVKNGIERLTKAIKEAKGASRQTGLDRW--
Pab          294 EPDEEGIIRFLCDEHDFSEERVKNGLERLKKAIKA----GKQSTLESWFI
consensus    301 **  * **  *  ** *********.*** **** .... .*. *. *..

Sso          346 ---F
Pab          340 KKKP
consensus    351 ...

M (*) = Matching amino acid
L (.)= Different amino acids within the same side chain group
D = No match
(.) Also shows points where an insertion/deletion have occurred
```

Figure 4.12 - Sequence alignment of FEN 1 from *Sso* and *Pab* using the online software Clustal Omega and T-Coffee(60, 141). Highlighted in green is the region in *Pab* present in the structure absent in the *Sso* structure. Highlighted in yellow is the region present in *Sso* that is absent in the *Pab* structure.

The program PHASER in the CCP4 suite (106, 109) was used to perform the MR, conducting the steps discussed in chapter 2.2.8.1. Firstly the program performed the cell content analysis, which generated a Matthew's coefficient value of 2.16 Å^3/Da, with 2 molecules in the AU. After this was the anisotropy correction. Anisotropy is defined as having data that has a directional dependence, which can affect the overall quality of the

data. PHASER corrects for overall anisotropy from the dataset. This was followed by both a fast rotation function as well as a fast translation function.

One molecule was used as a search model, searching for 2 molecules in the AU, based on the Matthew's coefficient. The fast rotation function was performed only once and this yielded a least-likelihood gain (LLG) score of 44.91 and a corresponding Z-score of 10.83. The fast translation function was also only performed once due only one solution coming out of the fast rotation function analysis. The top translation function (TF) score for this data set was 179.1 with a Z-score of 12.48. Packing analysis was then performed to ensure that as a result of the rotation and translation, there were no structural clashes. If a significant number of clashes were detected it would imply an incorrect MR solution.

The output from PHASER was then used to perform TLS and restrained refinement using isotropic B factors, using the program REFMAC. 10 cycles of TLS refinement were performed, followed by 10 cycles of maximum likelihood restrained refinement to a maximum resolution of 2.27Å. The R factor after this round of refinement was 29.47%, with an R free of 37.25%.

4.1.4.4 – Refinement & Model Building

Refinement involves iterative rounds of positional improvement in both real and reciprocal space in order to comply more with the observed structure factors that were generated during the data collection increasing the overall accuracy All refinement was performed using REFMAC, part of the CCP4 suite (109, 142). There are several quality control methods that are used throughout the refinement process to ensure any model building and editing improves the accuracy. The first of these are the R factor and free R factor (Rfree), which are a means of measuring the agreement between the model that is being built, with the crystallographic data that was collected.

$$R = \frac{\Sigma \| F_{obs} | - | F_{calc} \|}{\Sigma | F_{obs} |}$$

Where F_{obs} is the observed structure factor amplitude and F_{calc} is the calculated structure factor amplitudes. The Rfree is calculated from a portion of data (around 5%) that is excluded from any refinement steps. It is then used as a marker against how the refinement process is proceeding. It is used to test how well the model predicts experimental changes in the structure without being used when fitting the model.

Another example of a quality control indicator is the Pearson correlation coefficient (CC), which provides a direct assessment of relative proportions between the signal and noise and their contributions to the data for a particular resolution. The data is split into two 'half' data sets ($CC_{1/2}$), each obtained by averaging half of the observations for any given reflection, and is reported to be a more accurate means of determining which data are useful, particularly at higher resolution than R factors (139). The CC can be separated between CCwork and CCfree, which are correlation coefficients between calculated and experimental intensities. The CCwork for this data at 2.27Å is 0.925, with a CCfree of 0.887, which according to Diederichs and Karplus (143, 144) is an acceptable value at this resolution. At higher resolution the acceptable values for CC drop and a proposed acceptable cut off is 0.125, however performing a Student's T test can determine if the data, when compared to 0, is statistically significant and can therefore be included in the dataset (143).

The output of the MR was placed directly in to REFMAC and 20 cycles of TLS with restrained refinement were performed. Translation, Libration, Screw (TLS) is a means of deriving anisotropic motion for collective groups of atoms. The definitions relate to biologically relevant domains or regions of a structure that then allow the modelling of correlated thermal motion that will be common to all atoms within that domain or region. This refinement adds many fewer parameters than anisotropic refinement but can

significantly improve the R and R*free*, The R factor of the input coordinates was 38.07% with an R_{free} of 36.93%. After the 20 cycles the R factor had been reduced to 25.27% and the R_{free} had been reduced to 29.13%. Using the software COOT (145) in combination with REFMAC, backbones and side-chains that were not within the visible electron density were either modified (such as by rotation or by using an alternative rotamer) or completely removed in order to generate coordinates that better matched the density provided. The electron density was contoured to 2.0σ. These coordinates were then saved and were put into REFMAC for a second run of 20 cycles of restrained refinement. REFMAC was run a total of 14 individual times, all running 20 cycles. This number of cycles was performed as refinement is an iterative process, and only when the R factor and R_{free} values remain relatively static is it known that no further improvements can be made and the structure best represents the data collected. The final R factor after all refinement was 21.6% with an R_{free} of 27.55%. Generally the difference between the R factor and R_{free} should be no larger than 5%. Although the difference between the R factor and R_{free} generated for this dataset being greater than 5%, any issues would become apparent during the structural validation.

4.1.4.5 – Validation

The goal of refinement is to generate a model that is stereochemically sound and that best describes the data. This is characterised by a low R factor that remains at the same level after continuous attempts at further refinement. Once this has been achieved, the coordinates have to be validated to ensure that they are correct. All validation of the structure was performed using COOT (145) and the various validation tools it has at its disposal. The first, and possibly most powerful means of validating the structure is the Ramachandran plot. It was first used as a means of visualising dihedral backbone torsion angles (ϕ, ψ) of the amino acids present in the structure (146). The ϕ angle relates to the torsion between the N and Cα and the ψ angle relates to the C and the Cα. The

Ramachandran plot is a powerful tool used to visualise the torsion angles for each individual amino acid in the structure.

There are six individual plots that are used together with one another and they vary based on the particular amino acid. Firstly there is a general plot, which covers every amino acid and plots ψ along the x-axis with ϕ along the y-axis. It then highlights regions within a +180° to -180° range that are favoured energetically. There is also a plot for all isoleucines and valines, aliphatic residues that have moderate length side chains that are more permissive than proline, but more constrained than glycine. There is then a plot dedicated to every glycine within the structure. This is due to the fact that glycine has no side chain. This therefore means it has a greater range of acceptable torsion angles, shown by the larger allowed regions on the plot. Next there are two plots dedicated to every proline within the structure. One for *Cis* prolines and one for *Trans* prolines. Proline contains an aliphatic side chain that is covalently bonded to the nitrogen atom of the α-amino group, forming an imide bond and leading to a constrained 5-membered ring. This constraint is visualised by the very small allowed regions. In fact it is so restricted in its torsion angles that it has a tendency to affect the amino acid prior to it in the structure. This gives rise to the pre-proline plot, which shows every amino acid in the structure that directly precedes a proline.

The Ramachandran plot for this dataset showed that, of the 542 amino acids present in the structure, there were 535 (98%) in the preferred regions, and all 542 (100%) being within the allowed regions with 0 as outliers.

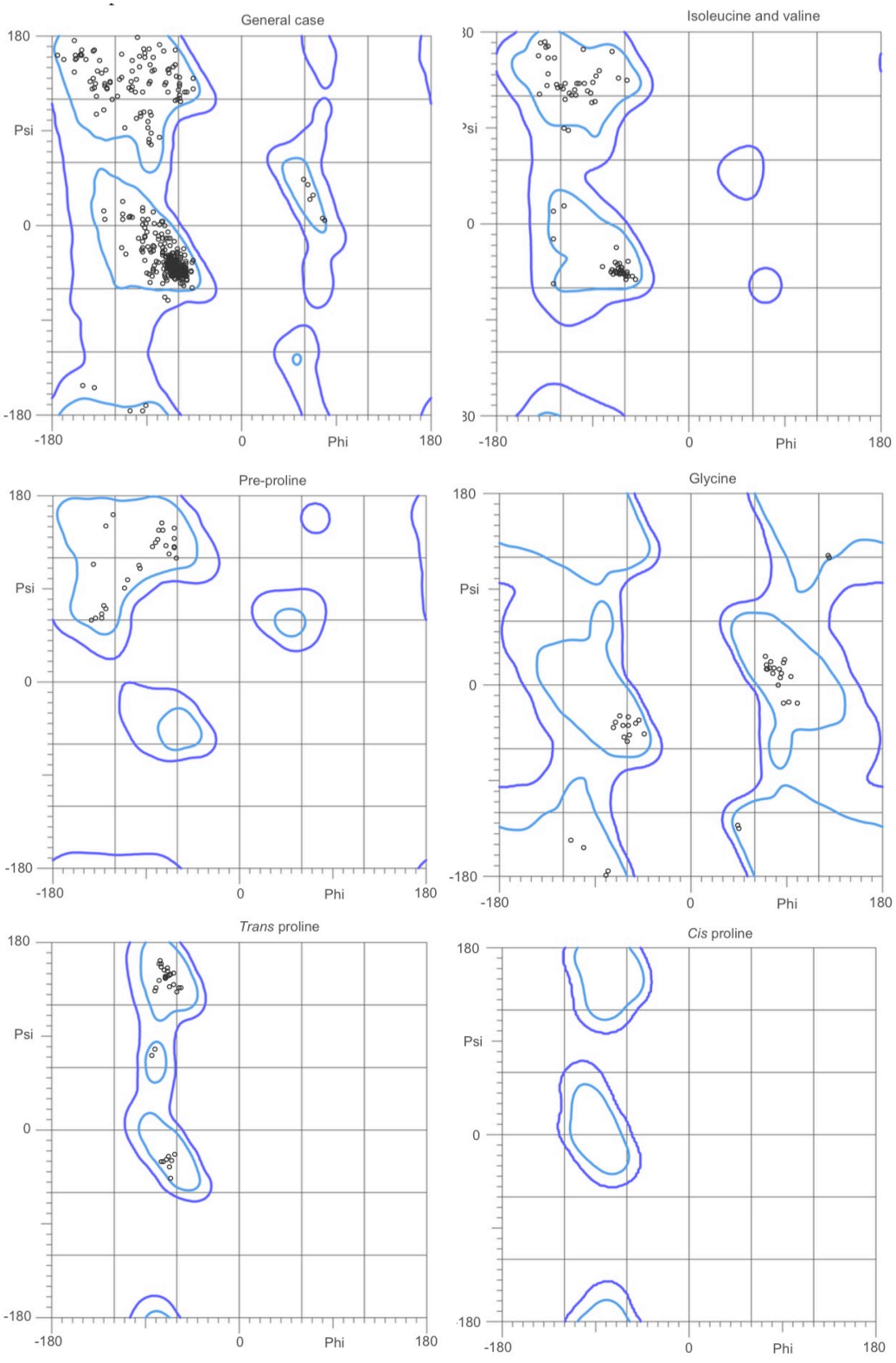


Figure 4.13 - Ramachandran plot generated by MolProbity (147) for this 2.27Å structure of *Pab* FEN 1. It indicates 98.7% of residues were in favoured regions and 100% of residues being in allowed regions.

4.1.4.6 – Conclusions

There are two molecules in the AU, which are organised into an orthorhombic unit cell, as seen in figure 4.15. The secondary structure shows 12 α-helices and 6 β-pleated sheet strands, arranged within the core of the structure, as well as several unstructured loop regions. The core and 'head' of the protein have more of the residues forming more complex secondary structures such as sheets and helices, whereas these secondary structure elements are less well defined in the lower 'tail' region, which could potentially imply that the lower portion of the protein has greater flexibility, and could theoretically move to allow incorporation of the DNA. Using the online software PDBSum (148), a schematic of the secondary structure can be generated.

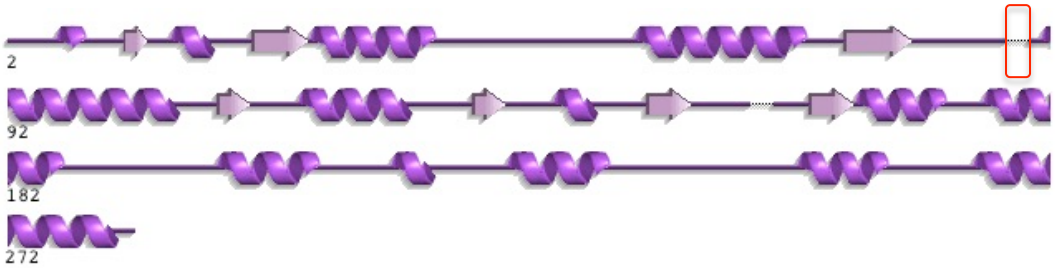


Figure 4.14 - Schematic of the secondary structure elements of the 2.27Å structure of FEN 1 from *Pab*, highlighting unstructured regions, alpha helices and beta sheets. Highlighted in the red box is the location in the structure, between K87 and K124, which is missing.

This schematic serves to illustrate how the head region, from the first half of the protein has a greater number of secondary structure elements, whereas the latter half of the protein, referring to the tail region has a greater number of unstructured regions. This could however be due to disorder within structure meaning regions of the protein couldn't be accurately built.

The region between K87 and K124 is missing from the structure. This is could not be modelled due to poor electron density, therefore it has been excluded from the structure. Based on the construct map in figure 4.1, this is within the internal (I) domain of the protein.

It is believed that this domain of the protein contains the active site (136) and this particular area of the domain encodes for the helical archway (137). Structural disorder is most likely to be the reason as to the lack of electron density for this region of the protein. Examples of the helical archway can be observed in both the *Pfu* structure and the human structure (figures 4.18 and 4.20 respectively). What can be observed are unstructured regions at the base of the activation loop. This therefore has the potential for a great deal of flexibility. It can therefore be surmised that when crystallising this flexibility lead to positional disorder within the crystal.

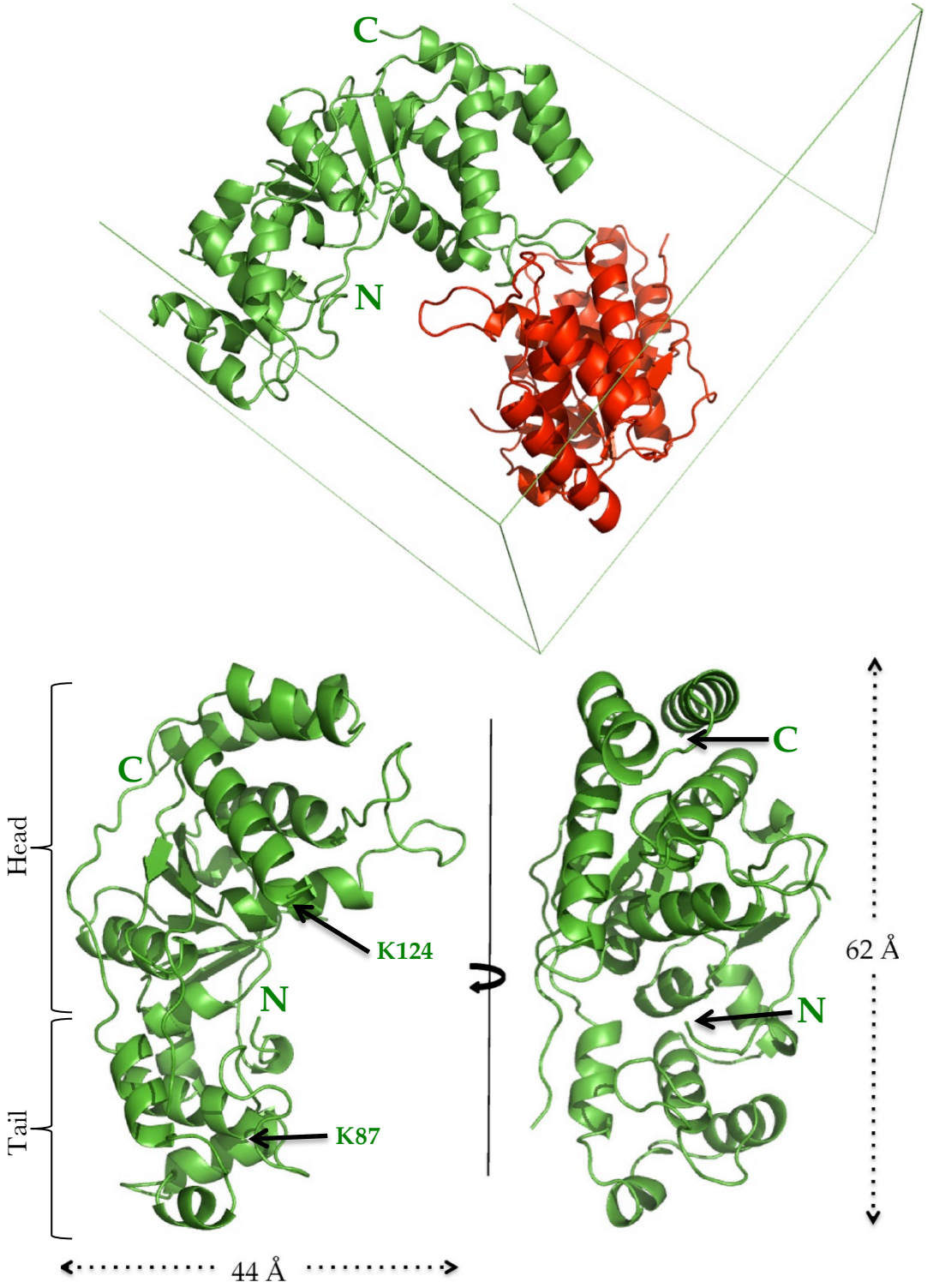


Figure 4.15 – The upper image shows the two molecules within the AU, coloured green and red indicating the two distinct chains, as well as the unit cell. The lower image indicates the dimensions of FEN 1 from *Pab* at the longest and widest points of the structure. Also shown in the figure are K87 and K124, the two residues at either end of the missing activation loop, present in structures of FEN 1 from other organisms.

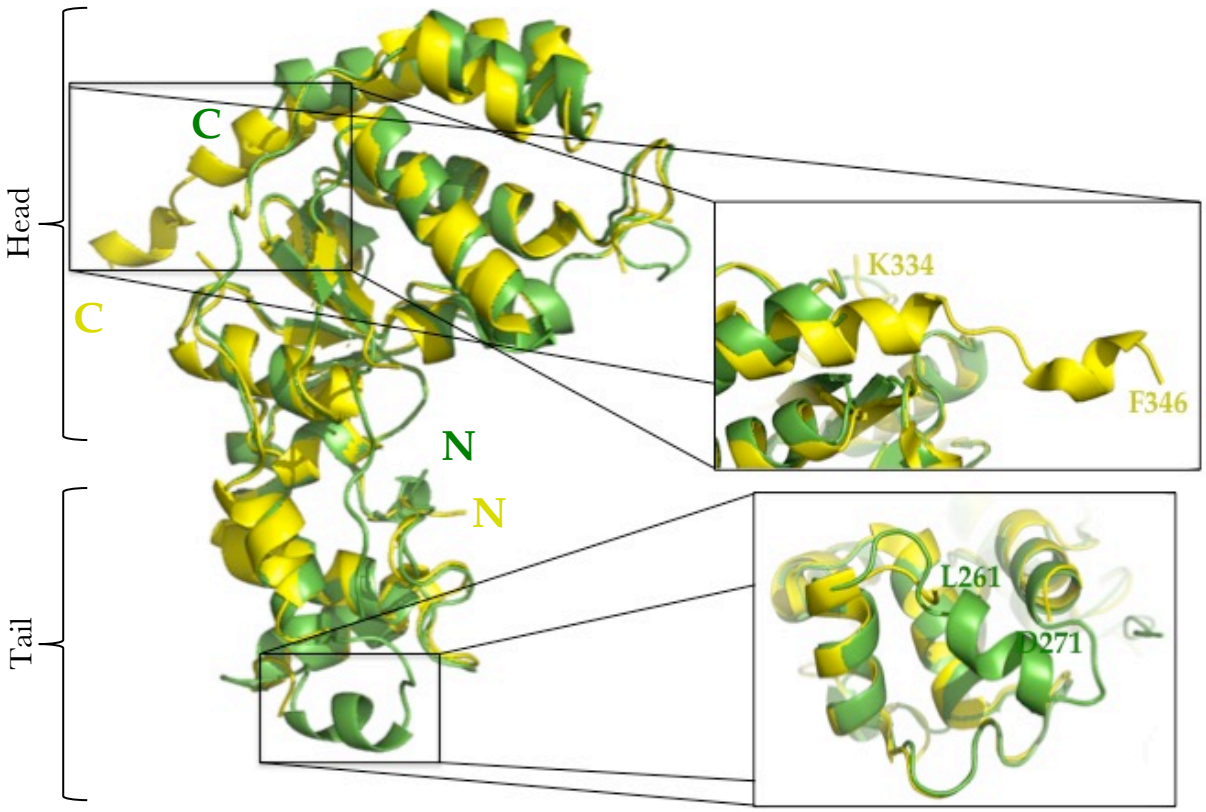


Figure 4.16 – Alignment of FEN 1 from *Pab* (green) and *Sso* (2IZO) (yellow), the model used as the search model for MR. Highlighted are the two main differences between the structures. The yellow text from K334 to F346 indicates the portion of structure present in the *Sso* structure yet absent in the *Pab* structure. The green text from L261 to D271 is a helical portion at the base of the structure that is present in the *Pab* structure and absent in the *Sso* structure.

In order to further validate the solution of MR, the refined *Pab* structure was aligned with the structure used as the search ensemble, FEN 1 from *Sso* (PDB code: 2IZO). In terms of sequence the two structures share 55% identity, but using the online software SSAP (Sequential Structure Alignment Program) (149) to compare the two PDB coordinates, they have 91% structural similarity with an RMSD of 1.43Å. The two main differences have been highlighted and exploded in the figure. The first difference is a loss of 17 amino acids from K334 to F346. The reason for this is most likely due to it being relatively unstructured. This lack of organised structure leads to heterogeneity, evidenced by a lack of electron density meaning it cannot be accurately added to the structural coordinates. This reason can also be

applied to the loss of the loop-helix-loop region between L261 and D271, however this is a lack of density in the search ensemble and not in the final structure. This will merely be as a result of a lack of homogeneity but it does give credence to the fact that the structure has limited levels of model bias.

4.1.4.7 Oligomeric State

All of the purification steps indicated that FEN 1 behaved as a monomer in solution as well as the DLS data, which showed a small proportion of the protein dimerising, however it appeared that this dimerisation was concentration dependent.

The online software PISA (150), a server designed to analyse the interface between macromolecules within their crystallographic environment, determined that of the 227 residues within the protein, 20 belong at the interface between the two molecules. These residues have been highlighted in the table below.

Chain A	HSDC	ASA	BSA	ΔiG	Chain C	HSDC	ASA	BSA	ΔiG
A:TYR 33	H	122.95	52.76	0.31	C:TYR 33		125.22	45.29	0.33
A:LEU 36		49.10	38.47	0.36	C:LEU 36		53.23	40.50	0.29
A:SER 37		35.54	25.79	0.40	C:SER 37		42.10	28.11	0.44
A:ARG 40		131.46	48.40	-0.88	C:ARG 40		125.59	43.39	-0.35
A:ASP 43		90.93	41.52	0.13	C:GLN 41		64.97	2.31	-0.03
A:GLY 44		43.53	22.18	-0.04	C:ASP 43	H	107.65	39.08	-0.09
A:THR 45		55.94	44.42	0.71	C:GLY 44		41.22	26.24	0.05
A:PRO 46		57.15	52.79	0.84	C:THR 45		61.32	51.26	0.82
A:MET 48		96.21	65.31	1.67	C:PRO 46		60.39	54.86	0.88
A:ASP 49		8.99	6.73	-0.03	C:MET 48		95.83	68.16	1.71
A:SER 50		97.61	66.95	0.43	C:ASP 49		9.72	7.46	-0.03
A:LYS 51		151.72	43.34	0.13	C:SER 50		93.97	60.28	0.26
A:GLY 52	H	51.89	49.55	0.02	C:LYS 51		145.16	41.97	0.12
A:ARG 53		95.28	35.66	0.57	C:GLY 52	H	52.32	50.48	0.02
A:ILE 54	H	55.31	50.32	0.68	C:ARG 53		101.72	35.89	0.57
A:LYS 126		101.76	22.28	0.28	C:ILE 54	H	57.47	53.65	0.75
A:VAL 127		107.70	64.78	0.96	C:LYS 126		210.54	32.24	0.33
A:MET 130		132.10	53.88	0.87	C:VAL 127		100.00	60.64	0.96
A:LEU 131		12.48	9.30	0.15	C:MET 130		119.83	55.60	0.90
A:ASP 309		75.27	3.71	-0.00	C:LEU 131		14.34	9.49	0.15

Table 4.2 - PISA output showing all 20 residues shown to play a role in the formation of an interface. In green are the residues that are predicted to be forming hydrogen bonds to the alternate chain.

This table shows all interfacing residues between the two chains within the PDB coordinates for the 2.27Å structure of FEN 1 from *Pab*. Highlighted in yellow are all of the interface residues. In green are the residues on each respective chain that form hydrogen

bonds. HSDC refers to interactions between the two chains, as either hydrogen (H), salt bridge (S), disulphide bonds (D) or covalent links (C). The ASA is the available surface area, measured in Å^2. The BSA is the buried surface area, also measured in Å^2. The green tally marks indicate the buried area percentage, with each tally mark representing 10% of the total surface available surface area. The ΔiG is the solvation energy effect, measured in kcal/M.

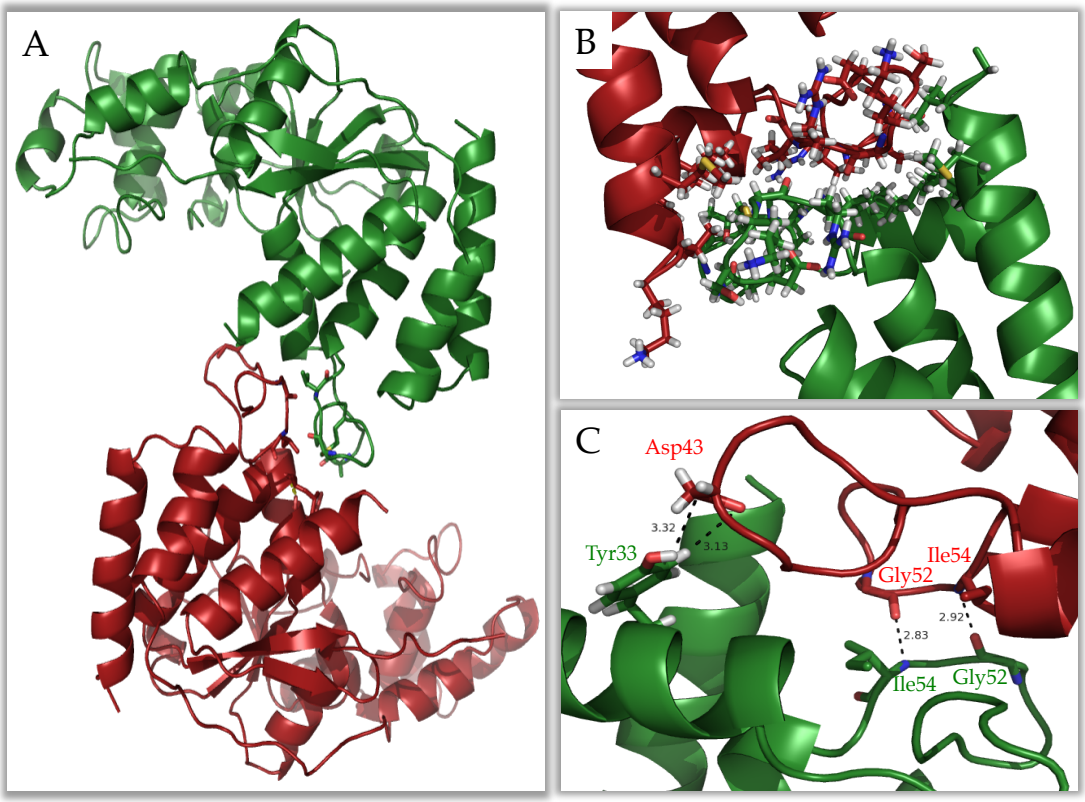


Figure 4.17 – (A) 2.27Å structure of FEN 1 from *Pab*, including stick representations of the residues predicted to be forming hydrogen bonds. (B) Expanded view of the interface between the two chains, with hydrogen atoms presented to aid visualisation of hydrogen bond interactions. All 20 residues forming the interface have been identified with a stick representation in Pymol. (C) An expanded view of the 6 residues on both chains that have been predicted to form hydrogen bonds. Residues in chain A are coloured green, residues in chain C are coloured red.

The carbonyl group of Gly52 from chain A forms a hydrogen bond with the main chain amide group of Ile54 from chain C and vice versa. These interactions fall around the symmetry axis for the two chains. Tyr33 from chain A forms two hydrogen bonds with Asp43 from chain C as seen in figure 4.17. An interesting observation from this analysis is the fact that, excluding two instances (Gln 41 from chain C and Asp 309 from chain A), each residue is present in both chains. PISA calculated that of the interfacing residues, only 5.9% is solvent accessible from chain A, with only 6.0% being accessible from chain C.

PISA awarded the interface a CSS (Complex formation Significance Score) of 1.000, implying that the interface plays a significant role in complex formation. Through searching the protein data bank for other FEN 1 structures from human, *Sso,* and *Pfu*, it is apparent that none of these structures exhibit this dimerisation event. As mentioned previously, there has been evidence of a concentration dependent oligomerisation during this research. There is therefore a strong likelihood that the concentration of the protein during the crystallisation process has led to the formation of a dimer during crystal formation. This does not preclude the possibility that this dimerisation is biologically significant, however various searches in the literature (90, 94, 96, 137, 151) indicate that only one molecule of FEN 1 binds to the DNA in order for successful cleavage of the flap.

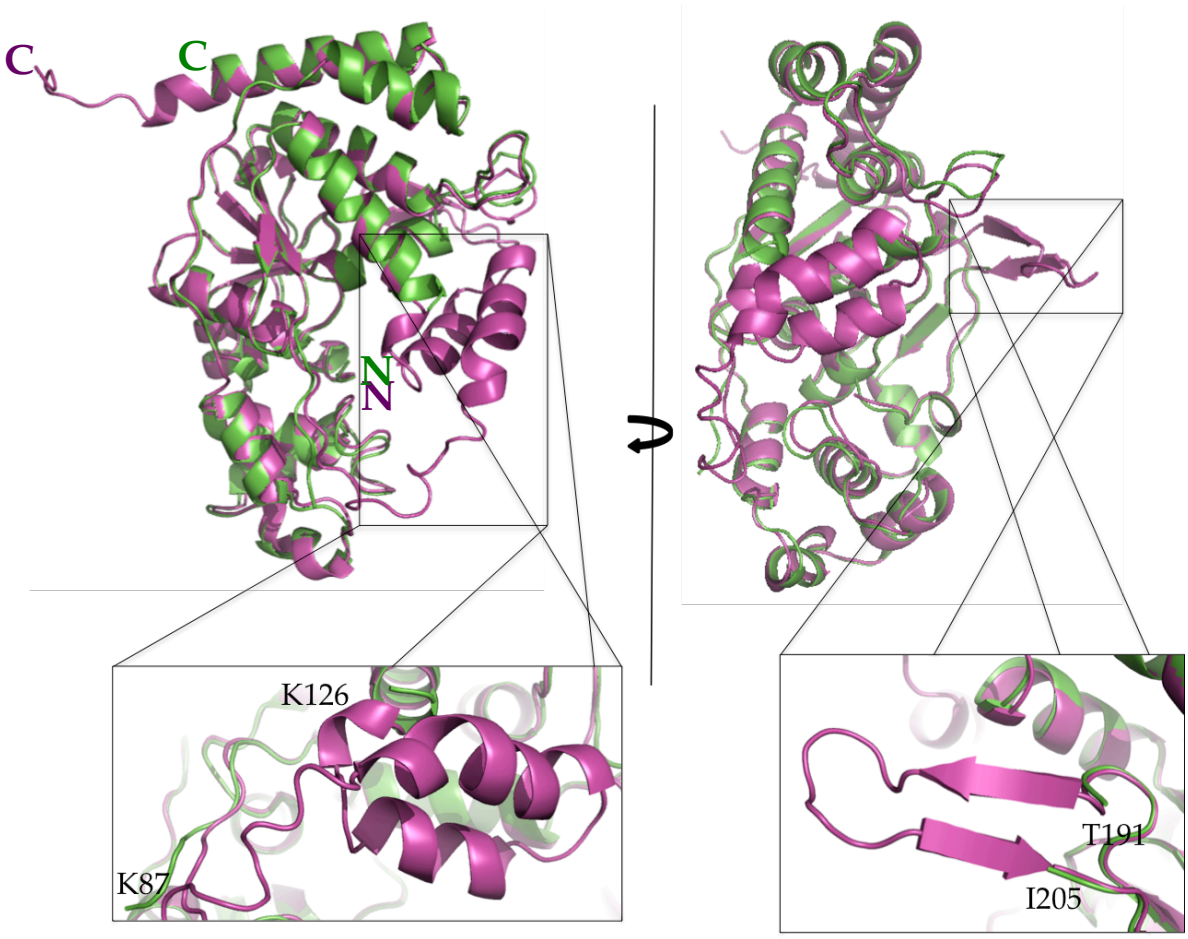


Figure 4.18 – Alignment of FEN 1 from *Pab* (green) and *Pfu* (1B43) (magenta). Although *Pfu* FEN1 has 90.3% sequence identity with *Pab* FEN1, a MR solution could not be found with *Pfu* FEN1 due to packing clashes. Exploded views of the two main differences between the two structures.

The alignment of the two structures in figure 4.18 are interesting, because the *Pfu* structure shares a 90.3% sequence identity with FEN 1 from *Pab*. When attempts were made to use the *Pfu* structure as a search model during MR, significant clashes were generated during the rotation and translation functions. According to the online software SSAP (149), the two sets of PDB coordinates share a structure similarity of 81% with an RMSD of 0.79Å,. The first difference between the two is the lack of the helical archway in the *Pab* structure compared to the *Pfu* structure. When looking at the sequences together in figure 4.19, it can be observed that the *Pab* sequence does account for this helical archway.

```
sp|Q9V0P9|FEN_PYRAB        PAFKKKELEKRREAREEAEIKWKEALAKGDIEEARKYAQRAT 42
sp|O93634|FEN_PYRFU        PEFKKKELEKRREAREEAEEKWREALEKGEIEEARKYAQRAT 42
sp|Q980U8|FEN_SULSO        PEQKSEELERRRKAKEEAERKLERAKSEGKIEELRKYSQAIL 42
sp|P39748|FEN1_HUMAN       PQLKSGELAKRSERRAEAEKQLQQAQAAGAEQEVEKFTKRLV 42
                           *  *. ** :* : : *** : ..*   *  :* .*:::
```

Figure 4.19 - Multiple-sequence alignment of the region believed to encode for the helical archway

The second difference between the two structures is that the *Pab* structure is missing a loop region containing two β-strands. This is most likely due to increased heterogeneity within the crystal caused by the lack of an organised structure.

The structures seen in figure 4.20 show how the human and archaeal FEN 1 structures align with one another. Using the online software SSAP again, the two PDB files were uploaded and along with the 42% sequence identity it showed an 80% sequence similarity with an RMSD of 2.06Å, when comparing chain A from each structure. The biggest difference between the two structures is a difference common with the *Pfu* structure, in a lack of the helical archway. As shown in figure 4.19, the *Pab*, *Pfu* and human sequences all possess this region and a reason it is lacking in the *Pab* structure could simply be due to increased disorder as a result of a lack of a stabilising effect from the presence of DNA.

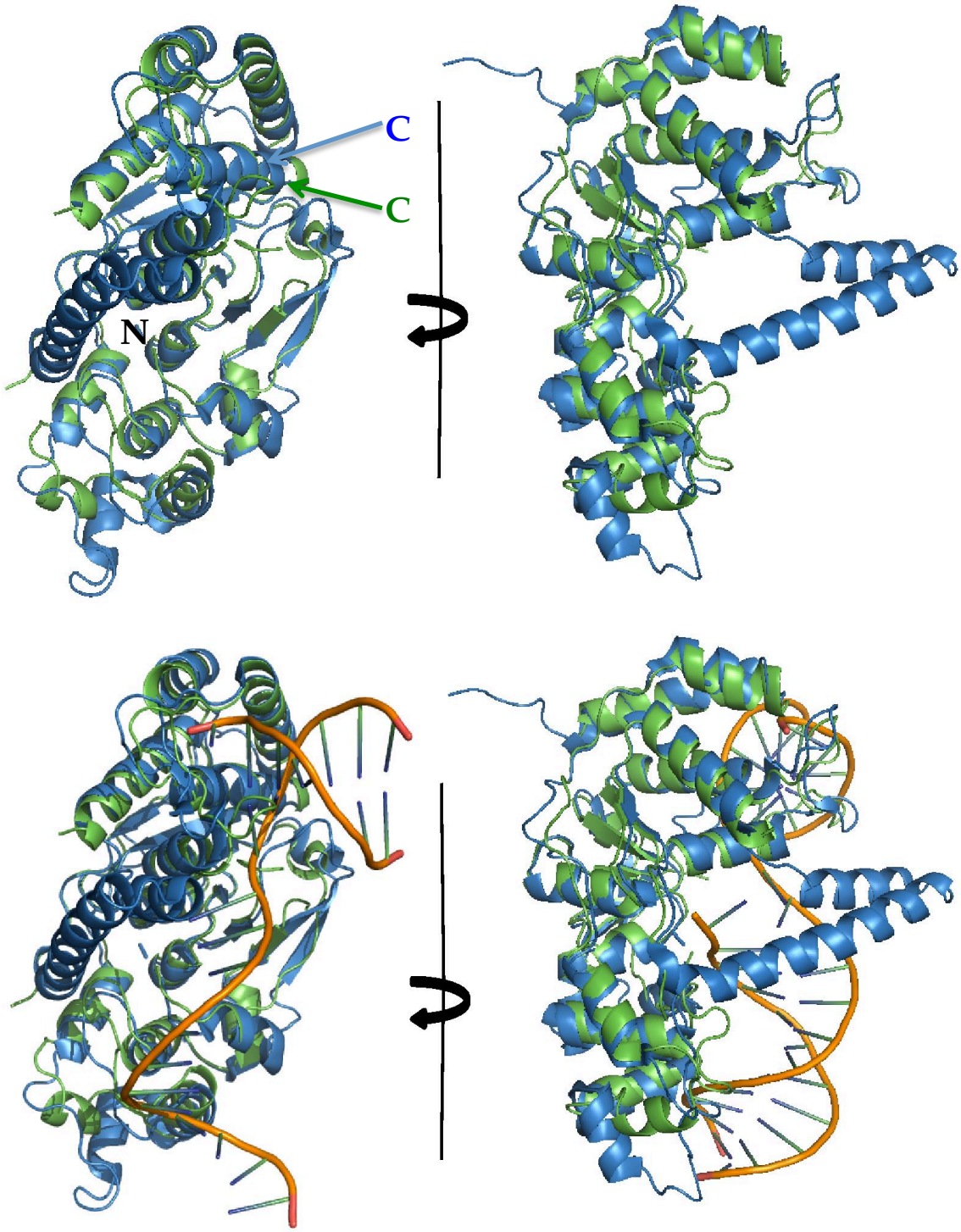


Figure 4.20 – Alignment of FEN1 from Pab (green) and human (PDB code: PDB code: 3Q8K) (blue). The two sequences share 42% identity and seem to share higher structural similarity. The N termini of the two chains are both identified with the single N in black. Their respective C termini are coloured according to the colour of their respective chains .The two main differences are shared with the *Pfu* structure and a purpose for these α-helices seems to be apparent upon the addition of the DNA substrate present in the PDB.

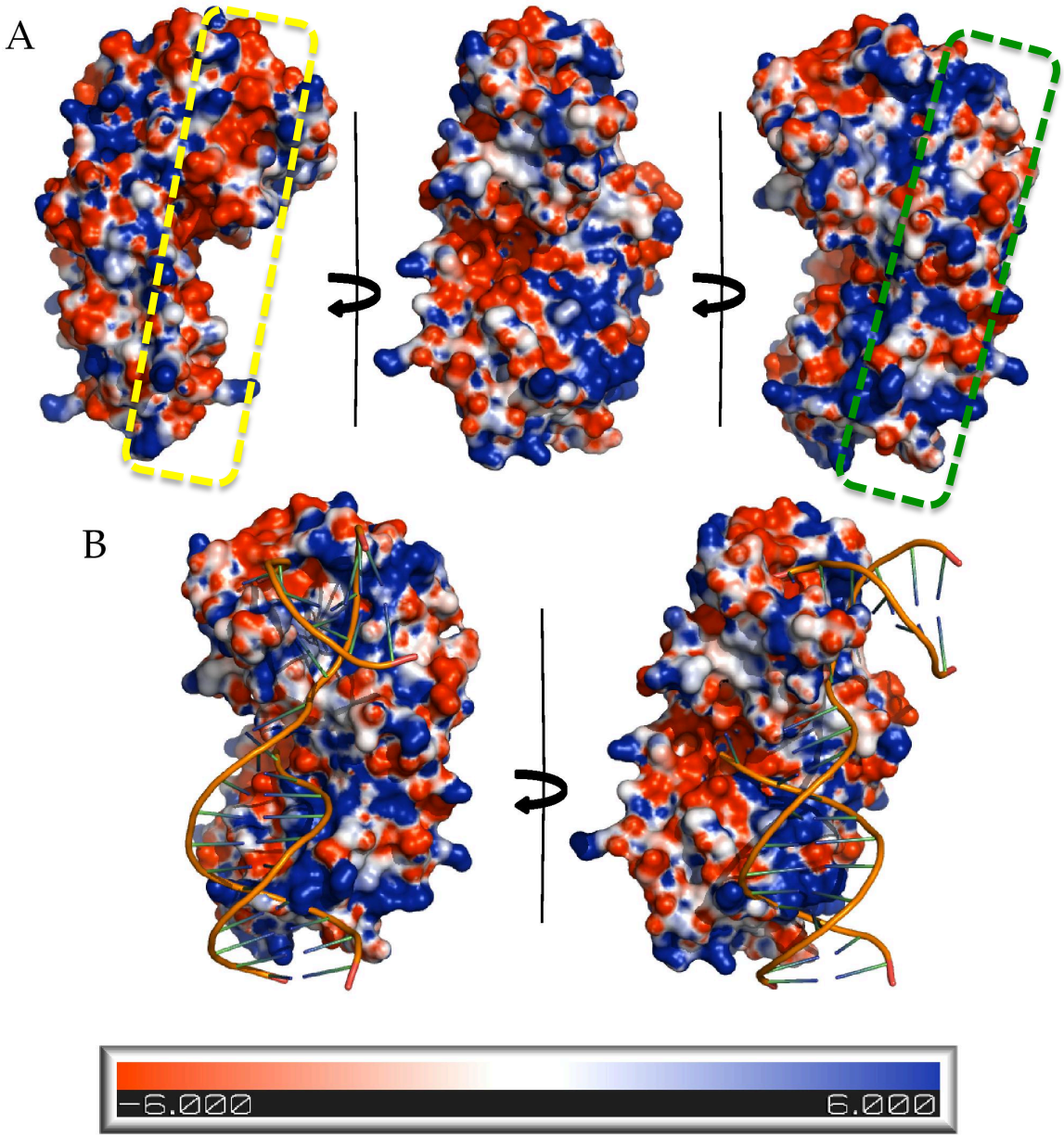


Figure 4.21 – (A) Electrostatic potential representation of FEN 1 from *Pab*. Both the left hand image and central image show a wide array of both positive and negatively charged areas. The right hand image shows a positively charged channel in an area that could bind DNA. This channel has been highlighted with a green dashed box, with the positively charged region being coloured blue. The left hand image shows the negatively charged strip, highlighted with the yellow box that is postulated to confer some degree of directionality. (B) Shows Pab FEN 1 modelled with the DNA coordinates from the human structure (PDB code: 3Q8K) as a theoretical binding mode for DNA, binding in the same proposed area based on surface charges. Due to only taking the DNA coordinates from the human PDB code: 3Q8K structure and overlaying this to the *Pab* FEN 1 structure, the helical archway cannot be observed.

Using the software Pymol, the surface of the protein was shown, giving rise to the basic shape seen in figure 4.19. Using an external plugin through Pymol known as APBS Tools 2.1(152), it is possible to determine the surface electrostatic potential. The PDB coordinates were first converted to PQR format, a modified version of the PDB that contains both the charge and radius parameters, using the in-built software within APBS Tools. Then using the macro 'Set Grid' the software determines the grid spacings for its calculations. Once completed various parameters can be altered to produce an image such as the one in figure 4.19. The surface electrostatic potential is measured in $k_bTe_c^{-1}$, where k_b is the Boltzmann constant, T is temperature in Kelvin and e_c^{-1} is the charge of an electron.

What is noticeable about the charge distribution for this structure are the two charged regions that run parallel with one another either side of the 'head' of the molecule, most visible in the far left and far right images from figure 4.21 (A). One theory could be that the positively charged strip is the means by which the DNA binds to FEN 1 in order for it to perform its function. To counter that the negatively charged strip adjacent to that could be charged in this fashion in order to confer a level of directionality to the protein, ensuring it can only bind in one orientation. This however is merely supposition and further experimentation is required.

A hypothetical model of the interaction with DNA is shown in figure 4.21 (B), using the DNA coordinates from the human FEN 1 structure, PDB code: 3Q8K after its alignment with the *Pab* structure. With this region being highly conserved, it is likely to have a similar mode of action elsewhere. A further aim of this thesis is to empirically determine the binding mode of *Pab* FEN 1 to DNA, to determine if it is in fact similar to the human interaction.

4.2 FEN 1 / DNA Complex Structure

4.2.1 – Introduction

Leading on from the information presented in the previous section, it is apparent that in order to prove the hypothesised binding mode of DNA with FEN 1 from *Pab*, they have to be co-crystallised and the complexed structure has to be solved.

Firstly the absence of the helical archway that can be seen in the *Pfu* FEN 1 structure but perhaps more importantly the human FEN 1 structure could be significant. However as previously alluded to, the sequences for this portion of the structure match up with both the *Pfu* and human sequences, implying its presence. A working hypothesis to explain its absence is a high degree of disorder, with a lack of organised structure at the base of the loop, it has fewer structural constraints and is therefore very flexible, meaning that it is unlikely to remain in one orientation in every unit cell within the crystal. Solving the crystal structure in the presence of DNA however could alleviate this issue. With DNA in place, this could act to stabilise the helical archway into a single conformation and therefore generating electron density.

4.2.2 – Oligonucleotide Design

The Tainer group crystallised human FEN1 and DNA using synthetic oligos and solved for the structure of the complex (PDB code: 3Q8K) (137), this was taken as an initial model for oligonucleotide design. For the protein to efficiently perform its function, it is self explanatory that it will not be sequence specific, however it has been shown that it acts in a structure specific manner (153), recognising the 5' flap that it cleaves.

The DNA needed to be long enough to be able to physically bind but not significantly longer than the protein itself, which would hinder crystal growth. As shown in figure 4.13, the length of FEN 1 from *Pab* is approximately 62 Å. The distance between two bases of

DNA is approximately 3.4 Å, which means that approximately 18 nucleotides will equate to the length of the protein (154). The template strand was therefore made to be 18 nucleotides long.

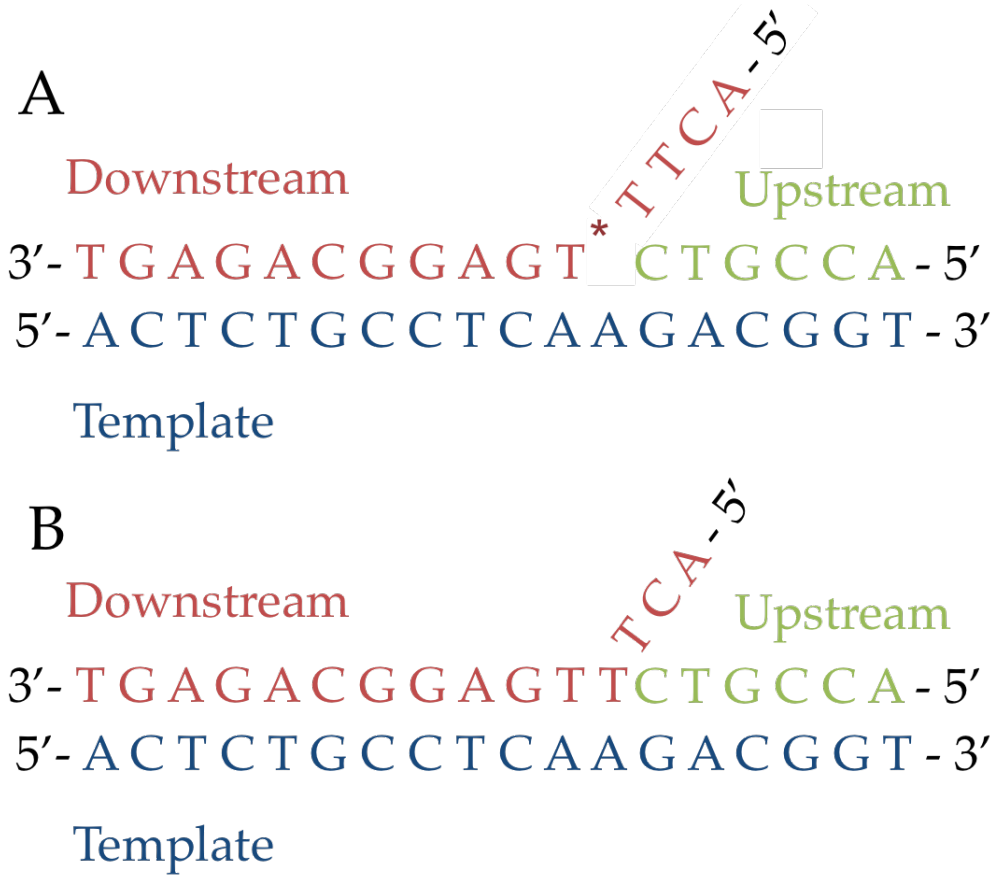


Figure 4.22 – (A) Oligonucleotide design used by Tsutukawa *et al* (137) including the introduction of a non-hydrolysable phosphoramidate, indicated by an asterisk (*). (B) Oligonucleotide design including the template, downstream and upstream portions that were ordered as single stranded DNA that annealed to one another for this study. A standard phosphate at the scissile phosphate position was used in this study.

The Tainer group replaced the phosphodiester bond of the 5' flap oligonucleotide with a 3' phosphoramidate group, thus meaning that the flap could no longer be cleaved. This was due to the fact that they were attempting to capture the complex, intermediate steps and products. The phosphoramidate was present as a control to additional base removal (137) (figure 4.22 (A)). In this study the initial complex was the structure to be determined and the experiments were designed in order to trap the complex before any

cleavage occurred, in order to determine binding mode and to increase likelihood of uniform crystal production.

4.2.3 – Annealing

Using the protocol mentioned in chapter 2, the three single stranded oligos were combined to form the double stranded sequence of DNA that contained the designed 'flap', in order to test its interaction with FEN 1 and ideally to crystallise the complex and gain structural insight into the mechanism.

After the annealing process, checks were performed to determine a successful protocol and formation of the desired product. One theoretical method would be to run the samples on an agarose gel. This however would be an issue when running such small oligos (6nt, 15nt, 18nt) and it would be extremely difficult to differentiate between them, especially the 15mer and 18mer. The other possibility would be to run a non denaturing acrylamide gel, which would allow the resolution required to detect the difference in size of the oligos, but to get accurate results for the single stranded DNA, the gel would have to be denaturing to inhibit duplex formation, which would vary greatly depending upon the sequence.

This therefore means that there is not really a technique that can be performed to check both the relative size of the individual oligos as well as the complex. However the samples were analysed *via* high performance liquid chromatography (HPLC) using a Mini-Q PE column (GE Healthcare). Each individual component was run over the column to determine both the elution volume and elution conductivity. Once values were obtained for each component, the complex was passed over the column to determine if the complex had in fact formed, or if it was simply a mix of the constitutive parts. The results of this experiment can be seen in figure 4.23.

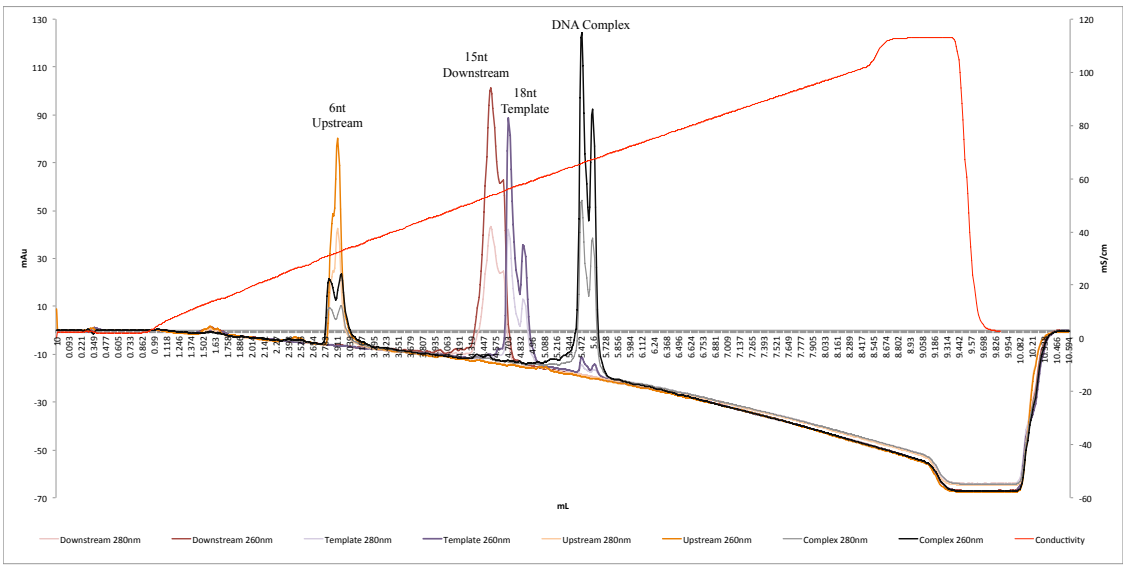


Figure 4.23 – Chromatography traces of each individual oligonucleotide, as well as the formed complex passing down a mini-Q PE™ column. Orange peak – 6nt Upstream, Red peak – 15nt downstream, Purple peak – 18nt template, Black peak – DNA complex.

As can be seen by the chromatogram, the smallest oligo (6 nt - upstream) is eluted first at 2.95 mL and a conductivity of 32.49 mS/cm. The next smallest oligo (15 nt - downstream) is eluted at 4.56 mL and 53.77 mS/cm. The largest of the three (18 nt – template) is then eluted at 4.75 mL and 56.25 mS/cm. It stands to reason that the two oligos with only 3 nt difference will elute at similar times and conductivities, especially in relation to the 6 nt oligo. The complex of all three eluted at 5.51 mL and 65.79 mS/cm. With its length being the same as the template the observable shift can be attributed to the fact that this is now double stranded, complexed DNA, rather than the single stranded DNA seen in the template.

Although this is the case it can be seen that in the DNA complex sample (Black trace), has a smaller peak that correlates to the same elution volume (2.95 mL) and conductivity (32.49 mS/cm) as the upstream only. One possible reason could be that the 6nt upstream fragment hasn't annealed and the complex is merely template and downstream combined. Although this is possible it is more likely that this peak can be attributed to an excess of upstream. This is due to the fact that all oligonucleotides were loaded onto the column at a

final concentration 100 μM and if there was no annealing event involving the upstream then the peak size would correlate with the upstream only peak, whereas it is significantly smaller. For the template strand there is also a small peak that correlates with the same elution of the complex. This is most likely going to be some self-complementarity of the template strand, as indicated in figure 4.24.

Figure 4.24 – A diagram showing the areas where the template strand could dimerise with itself

The fact that there are two peaks adjacent to one another for each sample could possibly be indicative of minor nucleotide errors in the synthesis and purification of the oligo from the manufacturer (Sigma Aldrich). This could be in the case of the 18 nt template that there is either some 19 or 17 nt variants present in the sample. For future experiments this HPLC could be used as a purification step rather than an analytical one, and the correct peaks could be collected and used in subsequent experiments. This however will most likely not be an issue due to the fact that a one nucleotide discrepancy at either the 5' or 3' end of the complexed DNA is unlikely to affect the central 'flapped' portion of the complex and it is this structure, rather than the sequence, that is required for effective binding to the FEN 1 protein.

4.2.4 – Binding to FEN 1

Once a complex containing the three oligos had been produced, it was then taken and added to purified FEN 1 in order to form the protein-DNA complex. The DNA was added to the protein in slight excess, in order for the equilibrium to favour the formation of the complex if the interaction is transient, which is most likely due to the proteins mode of action. This also correlates with previous examples in literature for similar protein-DNA

complexes (137). The final molar ratio of protein to DNA was 1:1.2. This was worked out using the following calculations:

$$\text{MW of FEN 1} = 38949 \text{ (Da)}$$

$$\text{Concentration} = 6.4 \text{ mgmL}^{-1}$$

$$n = 1.643 \times 10^{-4} \text{ moles}$$

$$= \underline{164.3 \text{ µM}}$$

Therefore to maintain the molar ratio required for the slight excess of DNA as previously mentioned:

$$164.3 \times 1.2 = \underline{197.2 \text{ µM}}$$

The complexed DNA was concentrated to a final concentration of 100 µM. Therefore for every 1 µL of FEN 1 added, 1.97 µL (~2 µL) of DNA was added. Once the desired amount of FEN 1 and DNA were added together, EDTA was added at a final concentration of 5 mM.

Mg^{2+} has been shown to bind to sites of two metal acidic residue clusters within archaeal FEN 1, known as M-1 (Asp-27, Asp-80, Glu-152, Glu-154) and M-2 (Asp-173, Asp-175, Asp-236), shown to be essential for catalytic activity. It is theorised that the M-1 site engages in nucleophilic attack of the scissile phosphodiester bond at the junction of the flap, whereas M-2 induces conformational changes resulting in the formation of the active complex (155). By chelating the Mg^{2+} with EDTA the theory is that the FEN 1 will no longer be catalytically active and therefore it will be easier to trap the protein and nucleic acid in complex with one another.

The mixture was then incubated for 1 h at 25°C at which point it was then centrifuged for 60 seconds at 13,000g to remove any precipitated protein and particulate matter ready for further experiments.

The first of these experiments was a native-PAGE gel in order to detect any changes in running height of the FEN 1 band +/- DNA. The results of this experiment can be seen in figure 4.25.

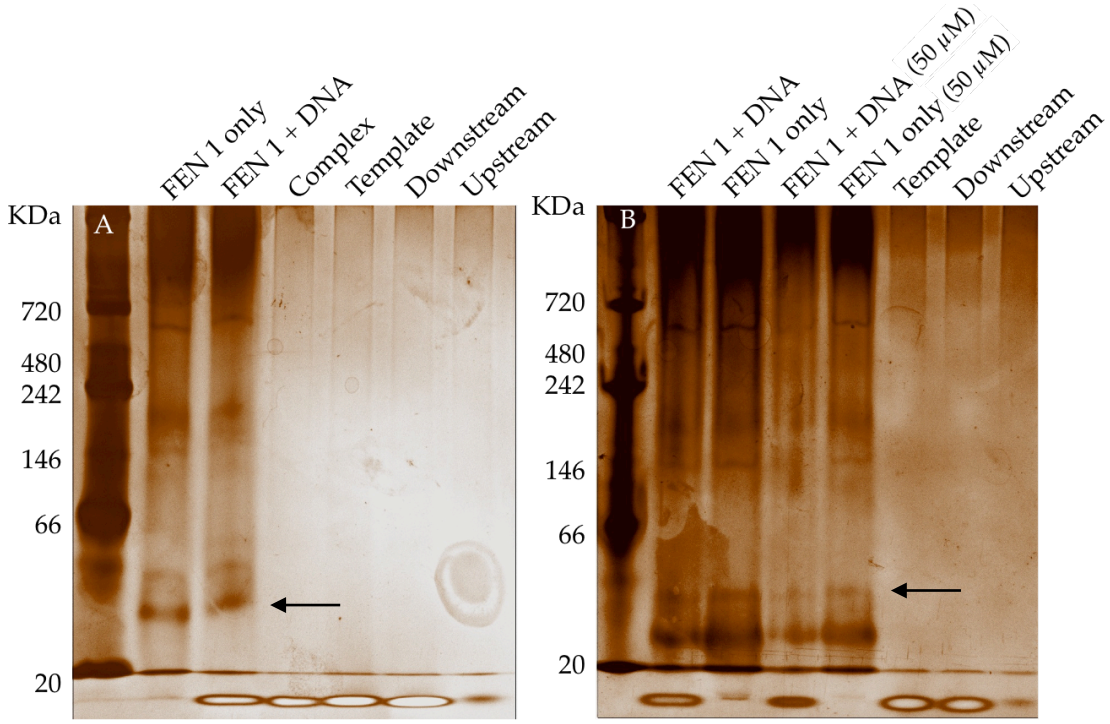


Figure 4.25 – (A) 15% Native-PAGE gel testing a possible change in the height of the protein +/- DNA. (B) Repeat of the gel to confirm a potential increase in size after complex formation. Two lanes were filled with half of the protein solution so the final concentration was 50 μM to reduce the smear levels often seen with native gel electrophoresis. The arrows indicate proposed FEN 1 bands based on their migration and theoretical molecular weight. The molecular weight of the double stranded DNA is approximately 11.1 kDa therefore the MW of the complex is approximately 49.9 kDa.

Based on gel A in figure 4.25, it could be concluded that the band assumed to be FEN 1 (38.9 kDa) does in fact shift upward, indicating a larger complex, upon addition of the DNA. This however could be refuted due to the fact that each of the other 'contaminant' bands present in the solution also shift upward implying that this is merely an artefact of the

gel, simply a running anomaly. Identical samples were then analysed on another freshly cast 15% native-PAGE gel which can be seen as gel B in figure 4.25. In this gel there is no upward shift of any bands in the gel upon addition of the DNA, either contaminant or FEN 1. This therefore implies that there is no complex formation.

This can be further corroborated by the presence of the over-saturated bands < 20 kDa. For reference purposes there are lanes dedicated to each individual oligo as well as the complex. The bands present in all 3 FEN 1 + DNA lanes are of a similar intensity to these reference lanes. Although the presence of these bands could indicate an excess of DNA, at these similar intensities it is more likely that it is the entirety of the DNA sample.

Another experiment that was performed on the FEN 1 + DNA sample was analytical size exclusion chromatography. Using a Superdex 200 10/300 GL column (GE Healthcare) a small volume of the sample was loaded and the chromatographic trace was recorded. Samples of FEN 1 *apo*, as well the DNA only as comparisons were also loaded. The results of this experiment can be seen in figure 4.26.

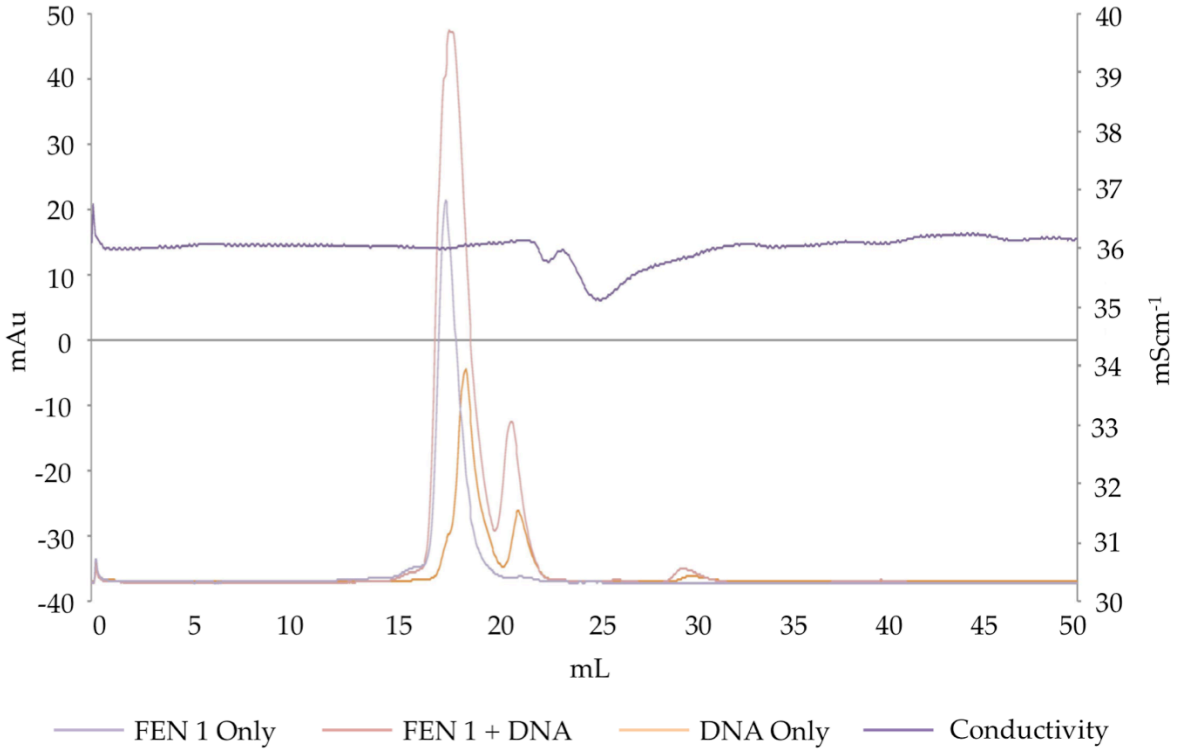


Figure 4.26 - A chromatogram showing the results of an analytical size exclusion chromatography experiment with FEN 1 +/- DNA as well as the DNA only.

As can be seen in figure 4.26, the purple FEN 1 only peak had an elution volume of 17.93 mL, which, based on previous results and calibration curves created in the same conditions, indicate that this is monomeric FEN 1 and is the control that will act as a comparison at the point when the DNA is added. The orange DNA only trace has two peaks at 18.95 mL and 21.61 mL. The presence of two peaks is most likely due to two different species. The first peak corresponds to a molecular weight of ~ 9.1 kDa, which is smaller than the double stranded complex, but this is based on calibration data for globular proteins, not double stranded DNA in a helical arrangement. The smaller peak at 21.61 mL roughly corresponds to ~ 2.3 kDa, again, which is smaller than single stranded template but based on calibration data that is not suitable for helical DNA. Future work will include a size exclusion chromatography trace for nucleic acids to help determine the identities of these two peaks.

Perhaps if the sample had been purified when performing the mini-Q experiment in figure 4.21 this secondary peak would be removed, a hypothesis that can be tested in future experiments.

When looking at the FEN 1 + DNA trace in red it can be observed that the elution volume of the large peak is at 18.15 mL. Although this is shifted very slightly right of the FEN 1 *apo* sample, the elution actually begins at the same point, and the largest DNA only peak is also contained under the peak. The increase in mAu can also be attributed to the fact that the two species are combined in a mixture rather than a complex. It would make sense that if the higher DNA peak were the complex, and the lower peak was either single stranded template or some other incorrect complex, then even if the FEN 1-DNA complex were forming, this second peak would be present in the FEN 1 + DNA sample, which can be seen in the trace. Overall this chromatographic trace further suggests that no complex has been formed.

Another means of detecting complex formation is to perform a thermal denaturation assay. When comparing a melting profile of FEN 1 *apo* and the FEN 1-DNA complex, the complex should theoretically be more stable than the *apo* protein and therefore melt at a higher temperature. If this shift is observed it will indicate complex formation. The inherent problem with this technique is the fact that *Pyrococcus abyssi* is a thermostable organism and the protein is less susceptible to thermal denaturation than human proteins. This problem can be observed in figure 4.27.

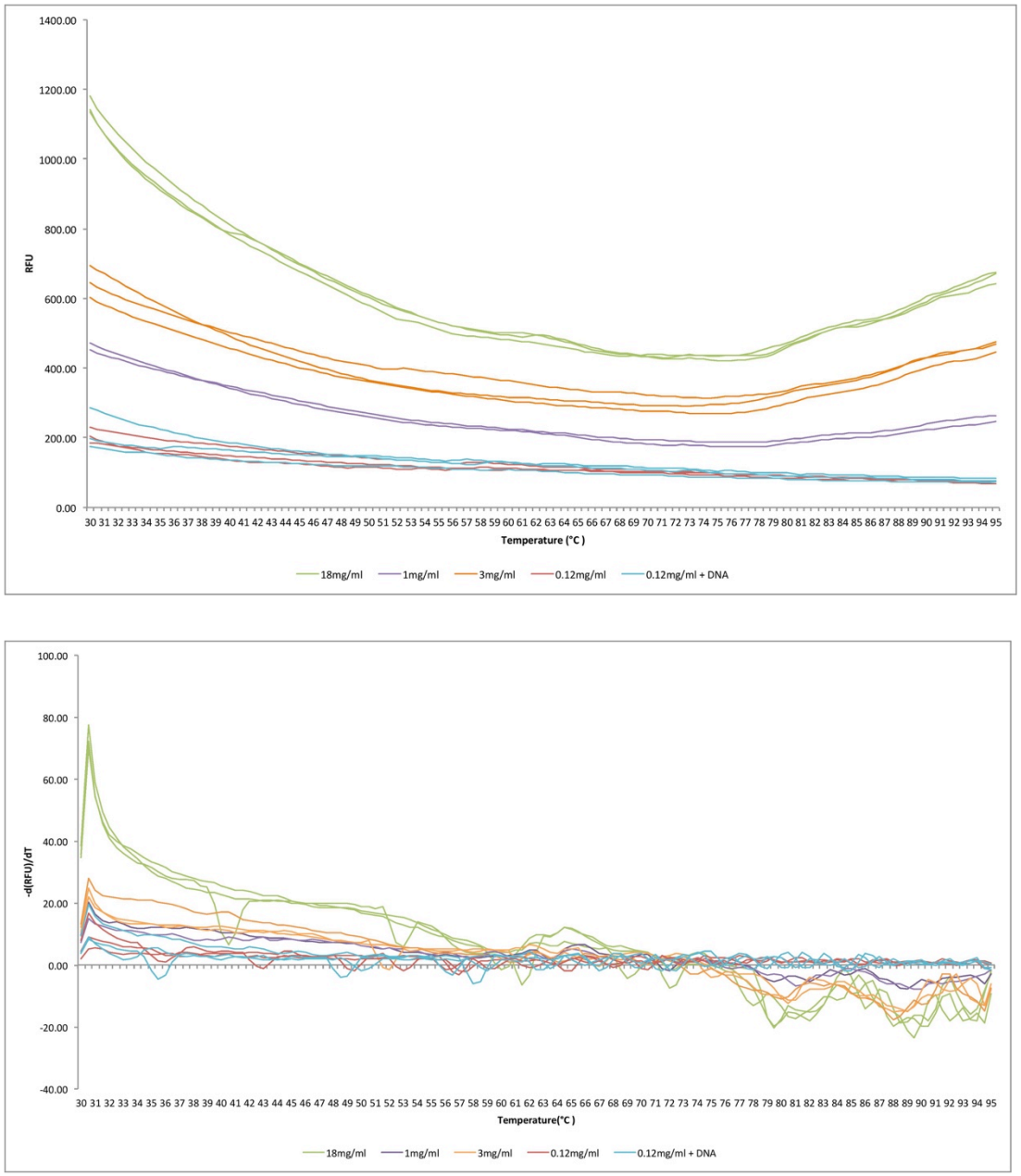


Figure 4.27 - Thermal denaturation assay showing both the relative fluorescent units and their derivative values.

Based on the information in both the raw data as well as the derivative data, at various concentrations from 18 $mgmL^{-1}$ to 0.12 $mgmL^{-1}$ there does not appear to be a suitable melt. At both ~80°C and ~89°C there appears to be signs of initial melts but they aren't significant enough to be considered the melting point of the protein. This therefore hasn't

provided any more conclusive information regarding the formation of a FEN 1:DNA complex.

4.2.5 – Crystallisation

After performing the previous experiments, and determining neither positively nor negatively if the FEN 1-DNA complex had indeed formed, a sample of the mixture was used in an attempt to produce crystals containing the complex. The FEN 1-DNA solution was used at its current concentration of 2.2 $mgmL^{-1}$. Using the Gryphon LCP liquid handling robot, the precipitant and FEN 1-DNA solution were placed in a sitting drop orientation in a 1:1 ratio in two 96 well plates testing the Structure Screen 1 + 2 (Molecular Dimensions) and the Nucleix (Qiagen) screens. The first screen was chosen due to the previous success with the FEN 1 *apo* crystal formation. The Nucleix screen was chosen due to its apparent increased likelihood of growth of crystals containing protein-DNA complexes based on previous successful examples, such as from within the PDB, checking for RNA, DNA and protein/nucleic acid complexes, analysing the crystallisation conditions and checking for how recurrent the components were (156). Crystals grew in 7 days in 200 mM zinc acetate, 100 mM sodium cacodylate, 18 % PEG 8000, pH 6.5. These crystals grew in the same conditions as the FEN 1 *apo* crystals.

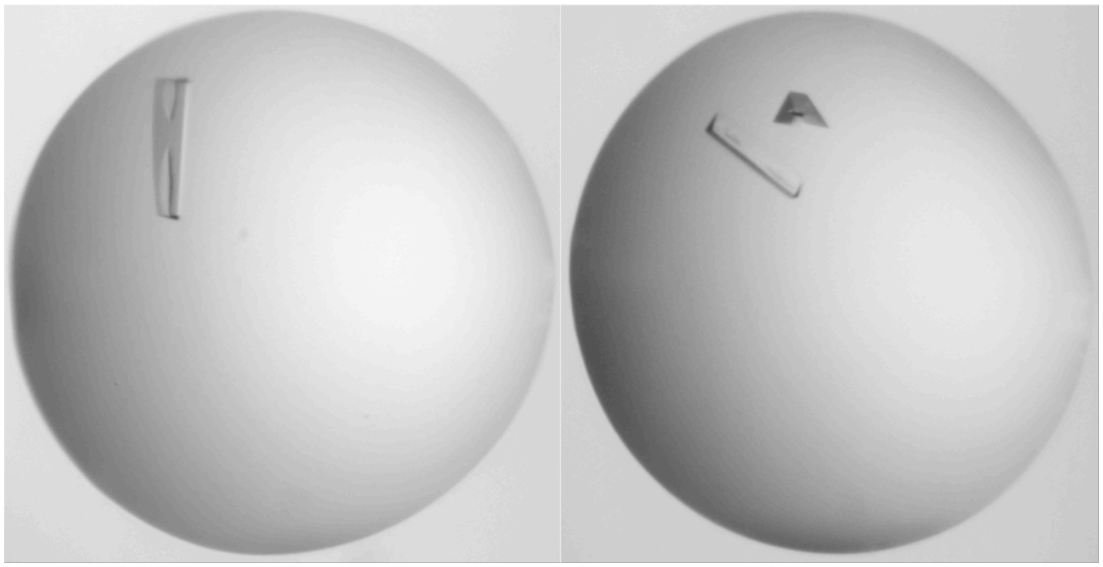


Figure 4.28 - Crystals in sitting drop, containing FEN 1 + DNA grown in 200 mM Zinc Acetate, 100 mM Sodium Cacodylate, 18% PEG 8000, pH 6.5

The crystals take the form of a triangular prism and all possessed small inwardly facing growth defects. This was reproducible but an adequate reason for this has not been determined. The crystals were harvested and cryo-cooled in liquid N_2 without the use of cryo-protectant due to the high concentration of PEG 8000 thought to be suitable to reduce cooling damage and the presence of ice rings in data collection. The data were collected remotely using the I04-1 beam-line at the Diamond Light Source Synchrotron.

4.2.6 – Crystallography

4.2.6.1 – Space Group Determination

The crystals seen in figure 4.28 were shot and diffracted to 2.83 Å resolution. Prior to any crystal growth refinement, the data were processed to determine the presence or absence of DNA within the structure.

The data were collected and processed using MOSFLM (113) in the space group $P3_112$. As mentioned previously the space group for the FEN 1 *apo* was $C222_1$. The reflections were scaled using SCALA (109). Although the crystal diffracted to 2.83Å, the

R_{merge} in the outer shell at this resolution was 4.776 and the R_{pim} was 3.519. The R_{pim} has within it an additional weighting accounting for multiplicity, which the R_{merge} doesn't, and can therefore be considered a more accurate guide to the quality of the dataset. It is believed that a value for R_{pim} of 0.5 is an acceptable cut-off for data quality. The value obtained for the $I/\sigma I$ in the outer shell at this resolution was 0.2, where an acceptable cut off for this is approximately 2.0.

The resolution was therefore manually cut back until acceptable values for these statistics were obtained. At a resolution of 3.62Å, the R_{merge} was 0.519, the R_{pim} was 0.370 and the $I/\sigma I$ was 2.3. This resolution limit was then used for the remaining data processing.

The crystal gave a total of 82395 observed reflections. The unit cell dimensions were shown to be a = b = 127.21 Å, c = 177.47 Å; $\alpha = \beta = 90°$, $\gamma = 120°$. The Rmerge values were 11.7%. The Matthew's Coefficient was calculated for both the protein alone as well as the protein in complex with the DNA. The value for the protein alone was 2.66 $Å^3$/Da, calculating 4 molecules in the AU with a solvent content of 53.8%. By assuming that the molecular weight of the protein-DNA complex is 38.9 kDa (FEN 1) + 11.2 kDa (DNA) = 50.1 kDa, the Matthew's Coefficient was 2.76 $Å^3$/Da, predicting three molecules in the AU with a solvent content of 55.5%. The data were collected to a completeness of 92.8% (79.9%). There was a multiplicity of 4.7 and the $(I) / \sigma(I)$ was 6.1 (2.3).

	FEN 1	FEN 1 + DNA
Data collection		
Space group	C2221	$P3_112$
Cell dimensions		
a, *b*, *c* (Å)	65.23, 86.40, 238.99	127.21, 127.21, 177.47
α, β, γ (°)	90, 90, 90	90, 90, 120
Resolution Range (Å)	50.26 - 2.27	51.34 – 3.62
R_{merge}	0.143 (0.317)	0.117 (0.519)
$I/\sigma I$	6.0 (2.7)	6.1 (2.3)
Solvent Content (%)	43.1	53.8 / 55.5
Molecules / AU	2	4 / 3
Completeness (%)	96.1 (99.8)	92.8 (79.9)
Redundancy	4.2	4.7

Table 4.3 - A table comparing the data collection statistics for FEN 1 apo crystals and crystals containing FEN 1 in complex with the DNA. Values in parentheses indicate the statistic for the highest resolution shell

The table shows that despite the two sets of crystals growing in the same conditions, they have in fact changed upon addition of the DNA. The space group has changed from a centred orthorhombic space group to a primitive trigonal one. The unit cell is considerably larger when the DNA is added when compared to the protein only. This has allowed 4 molecules within the asymmetric unit compared to 2 previously. The presence of the DNA could have allowed the protein to adopt an alternate conformation, meaning it packs very differently leading to these outcomes.

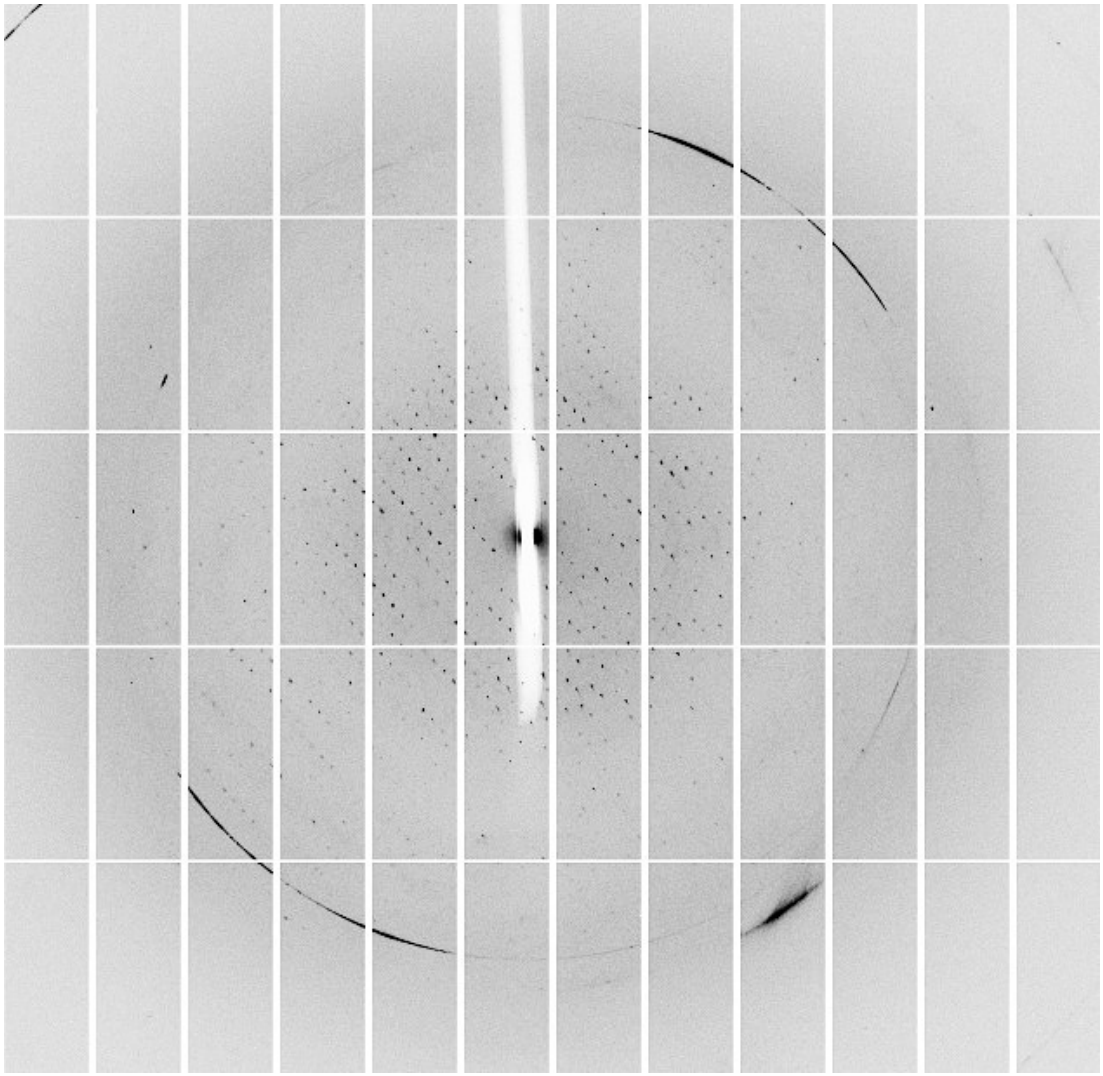


Figure 4.29 - Diffraction pattern of the FEN 1+ DNA crystals shown in figure 4.28, which diffracted to 2.83 Å resolution. The pattern indicates one partial ice ring and another very feint one, indicating the lack of cryo-protectant during crystal harvesting was not too detrimental, however crystal quality could have been improved by including this step.

4.2.6.2 – Molecular Replacement

The FEN 1 *apo* structure solved to 2.27 Å was the first search model used in order to solve the phase problem using PHASER. A single molecule was used, searching for 4 moleucles within the AU, based on the Matthew's coefficient. The cell content analysis revealed similar values for the Matthews coefficient as the values obtained prior to any MR. The rotation function generated a top RF score of 60.47 with a Z-score of 6.2. The

translation function was then performed, generating a TF score of 143.3 and a Z-score of 8.9. Packing analysis was performed and there were 2 possible orientations, both of which were deemed acceptable. Automated molecular replacement was then carried out and the best solution obtained had an LLG of 1536.4 with an R factor of 41.45% prior to any refinement.

The output from PHASER was then used to perform TLS and restrained refinement using isotropic B factors, using the program REFMAC. 10 cycles of TLS refinement were performed, followed by 10 cycles of maximum likelihood restrained refinement to a maximum resolution of 3.62Å. The R factor after this round of refinement was 29.47%, with an R free of 37.25%.

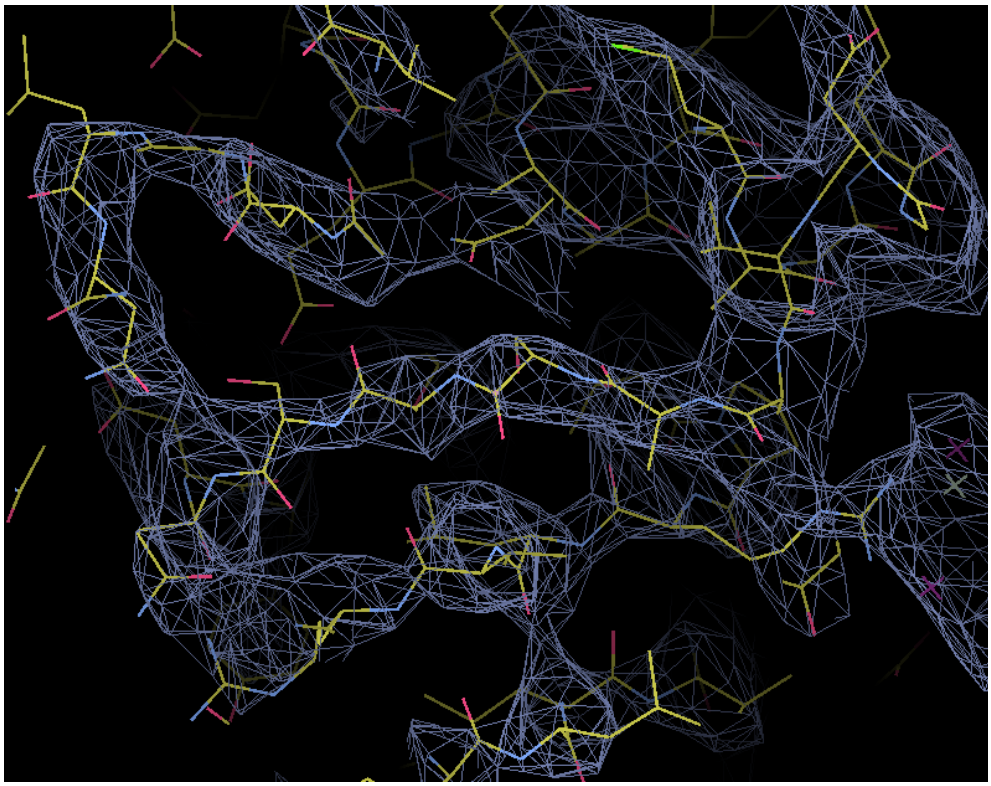


Figure 4.30 - Electron density map (2Fo-Fc) of a portion of the 3.62Å structure of FEN 1 contoured to 1.8σ. The density strongly correlates with the atomic coordinates.

The electron density shown in figure 4.30 is a representative example of the density throughout the structure. It correlates well with the coordinates of the protein structure meaning that any conclusions drawn from the structure in its current form can be done so with a certain degree of confidence. At 3.62Å however side chain positions and rotamers cannot be accurately identified leading to levels of ambiguity. The purpose at this stage of the process is to determine the presence or absence of DNA within the structure and at this resolution density for a double helix should be observable.

The four molecules within the asymmetric unit, seen in figure 4.31 (A), are coloured according to their chain ID, with the N and C termini highlighted on chain A (green). Chain A and chain B (cyan) sit end to end within the AU, whereas chain C (yellow) and chain D (magenta) sit side by side facing inward toward one other. However, packing of asymmetric units within the unit cell allows for chains A and B, which sit end to end in the AU, to effectively pair with their corresponding chain from the adjacent AU. This allows them to also sit side-by-side, facing inward toward one another, shown in figure 4.29 (B).

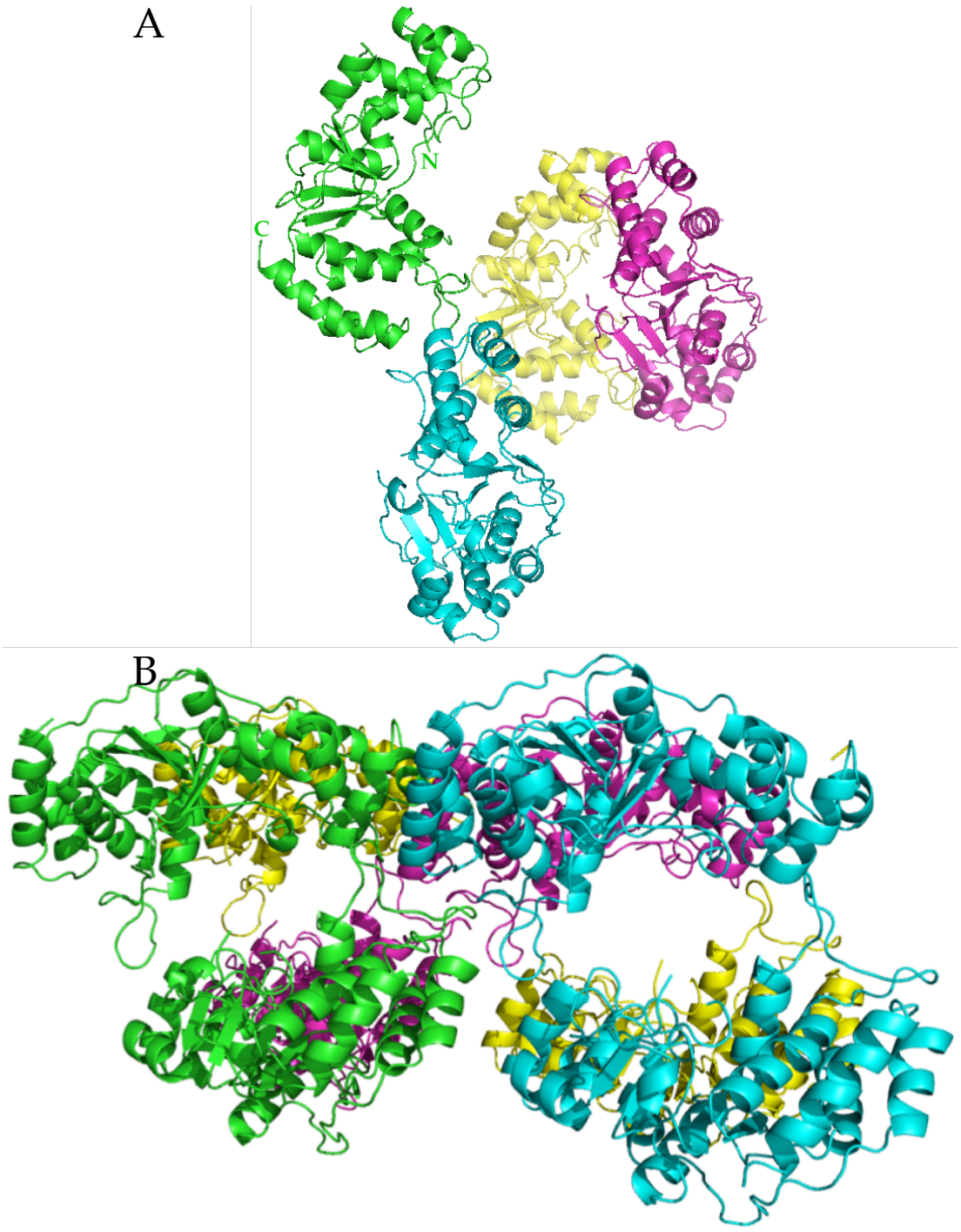


Figure 4.31 - (A) The 4 molecules within the AU coloured by chain. Chain A (green), chain B (cyan), chain C (yellow), chain D (magenta). (B) - Inclusion of molecules from adjacent AUs showing how each asymmetric unit packs within the unit cell.

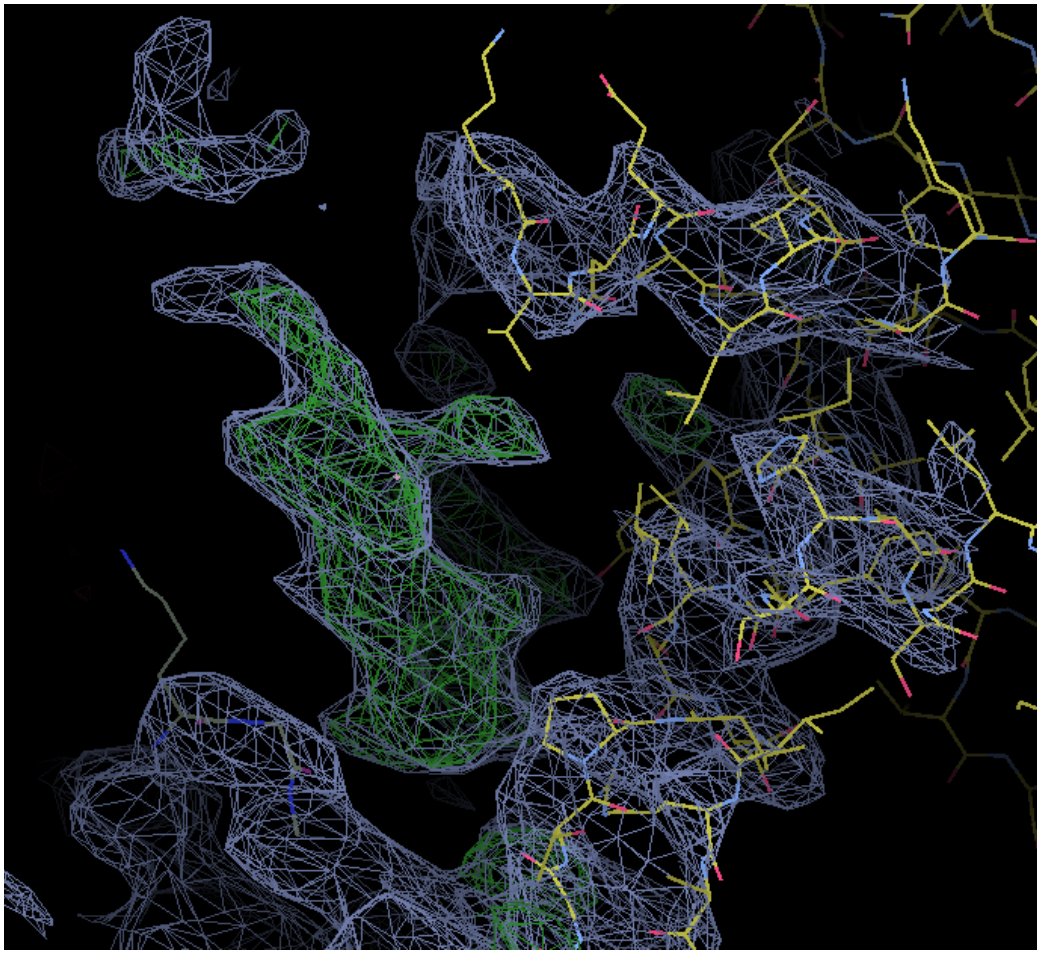


Figure 4.32 - 2Fo-Fc map contoured to 1.8σ in blue showing atomic coordinates from FEN 1 as well as unoccupied density adjacent to the protein. Fo-Fc map contoured to 2.8σ showing positive density in green.

Adjacent to the protein coordinates is a region of positive density that runs alongside the protein in a manner similar to that of DNA binding to human FEN 1 in the PDB PDB code: 3Q8K (137). This positive density is approximately 18Å in length and 8Å at its widest point, which is not wide enough to contain a DNA double helix, which is approximately 20Å, and not long enough to contain even one helical turn, which is 10.5 base pairs, the equivalent to 34Å (157). It also doesn't exhibit any helical nature that could be attributed to DNA. It does exhibit features of an α helix, however a helix in this position would not relate to any residues within the sequence of FEN 1 from *Pab*, including the missing activation loop previously discussed.

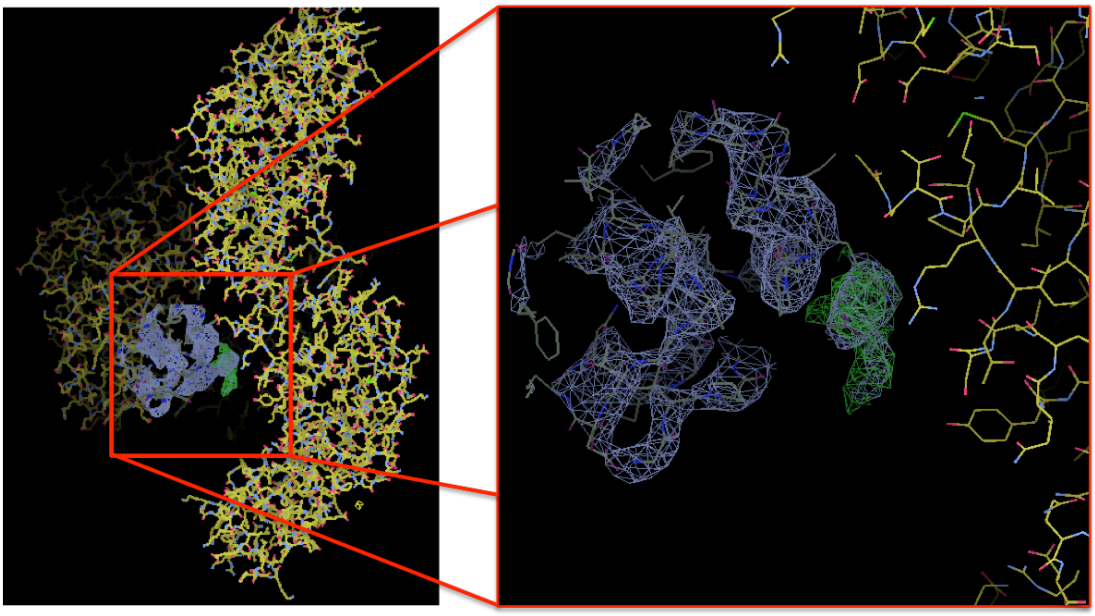


Figure 4.33 - The left hand image shows the PDB coordinates for FEN 1 from Pab at 3.62Å. The 2Fo-Fc map in blue, contoured to 1.2 σ, is adjacent to the protein. The right hand image is an expanded version of the first image highlighting both the 2Fo-Fc map and the Fo-Fc map, contoured to 2.8σ. Also shown in the density are the symmetry atoms indicating that any density in these regions is due to symmetry mates from adjacent unit cells within the crystal.

In the lower left hand corner of figure 4.31 as well as the right hand image of figure 4.33 additional regions of density can be seen to occupy space surrounding the protein, as well as areas of positive density. They can be seen to be containing symmetry mates of the visible AU. The particular symmetry mates seen in figure 4.33 were previously pointed out in figure 4.31 when discussing how the chains interact with one another, whether it be through symmetry mate interactions or with alternative chains within the AU. Symmetry atoms account for a significant proportion of density adjacent to the protein, however there are distinct regions that cannot be attributed to this, a prime example of which is the positive density in figure 4.30. Before making an attempt to fit nucleic acids to the density, modelling was performed to determine if there was space for DNA to bind to the protein whilst it occupied this particular orientation within the AU.

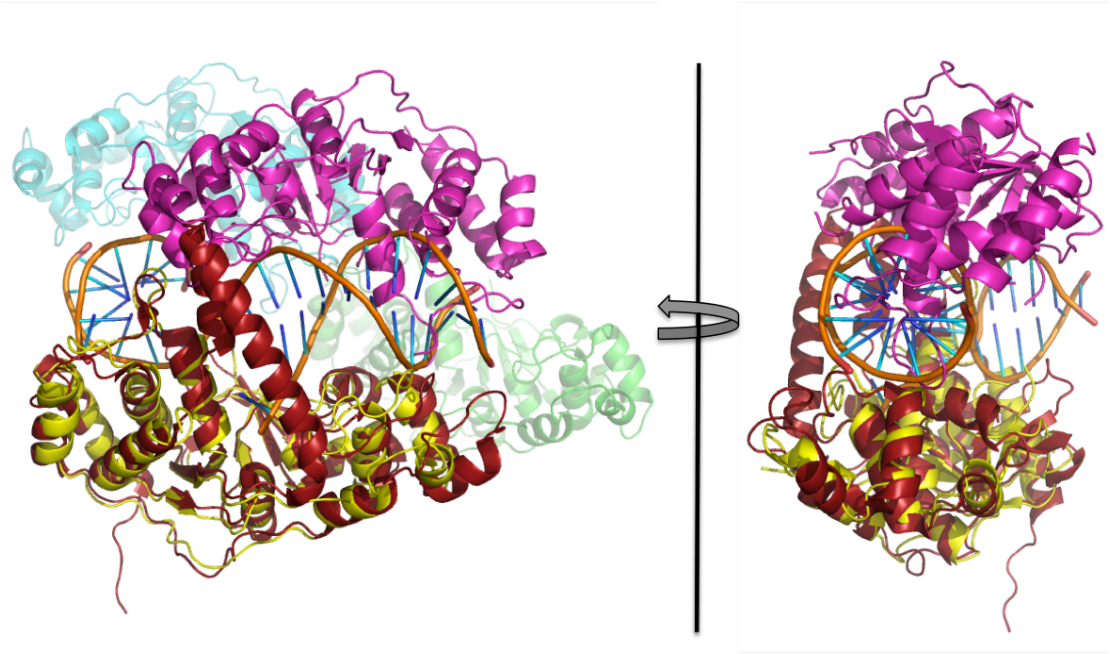


Figure 4.34 – model of the coordinates from the 3.62Å structure of FEN 1. In the figure the coordinates from the human FEN 1 structure (PDB code: 3Q8K) were aligned to chain C from the *Pab* structure. Making the assumption that the DNA will bind to *Pab* FEN 1 in the same manner as it does in human, it is apparent that there is insufficient space when the protein adopts this orientation within the crystal for double stranded DNA. This lack of space is apparent when observing the clashes between the DNA and chain D (magenta)

In order to theoretically determine if this AU could allow for DNA to form a complex with the protein, the coordinates from the human FEN1-DNA complex (PDB code: 3Q8K) (137), shown in red, were aligned to one of the four molecules within the AU. FEN 1 binds to DNA in a structure-specific manner rather than a sequence-specific one (153). Assuming that the FEN 1 from *Pab* binds to DNA in the same manner as it does in humans, this suggests that due to the protein packing in this particular orientation within the crystal, there is not enough space in between the two chains to allow for this particular DNA. Meaning that even if only one double stranded molecule of DNA occupied the space between the chains, it would bind to one of the two chains but, as can be seen in the figure, the clashes observed from this alignment indicate a distinct lack of space.

4.2.6.3 – Refinement

With an initial R factor after one round of TLS and restrained refinement of 29.47% it can be surmised that the phases obtained through molecular replacement accurately relate to the experimental data obtained.

The purpose of collecting data from these particular crystals was to determine a binding mode for DNA with *Pab* FEN 1, in order to empirically determine if the binding mode is similar to that proposed for human FEN 1 by Tsutakawa *et al* (137). As discussed in the previous section however, there is a lack of evidence for any DNA to exist within the crystal structure, therefore this aim cannot be met.

Due to the fact that the data was only collected to 2.83Å, but was manually cut back to 3.62Å to ensure only suitable data was used, any data obtained from this experiment, has a lower resolution than the 2.27Å achieved when crystallising *Pab* FEN 1 *apo*. This combination of a lack of electron density for any DNA, as well as significantly lower resolution than previous experiments means that further refinement of this structure would serve little purpose and therefore was discontinued.

4.2.7 – Concluding Remarks

Based on the results discussed in chapter 4.2, it is believed that the FEN 1 and DNA are not co-crystallising in complex with one another. One possible reason could be that the interaction is transient and although a complex may be forming, it may be breaking away again due to the fact that either the released DNA has had the flap removed or perhaps the interaction is not strong enough to maintain the complex.

Another possible reason is that the concentration of EDTA is not high enough to chelate all the divalent cations present in the solution. If the Mg^{2+} concentration is sufficient for the enzyme to turn over ATP and cleave the flap, releasing the cleaved DNA product then it won't be visible in either the assays or the crystal structure.

It could also be that the DNA being used was not pure enough for crystallisation purposes. If the doubling of peaks present in figure 4.21 indicates several species, perhaps this heterogeneity meant that crystals could not pack effectively and only crystals without the DNA could grow.

4.3 FEN 1 High Pressure Crystallography

4.3.1 – Introduction

The replicative machinery of *Pab* is poorly understood at the high pressures that the piezophilic archaeon thrives at, but structural studies could generate valuable insight (158). That being said it has been shown that *Photobacterium profundum*, another piezophilic archaeon, and *Xenopus* cell-free systems, which can survive at pressures up to 80 MPa, both halt DNA replication at high pressure (158, 159). One previous example of an HPMX experiment, using the enzyme urate oxidase from *Aspergillus flavus,* showed that upon pressure being applied at 150MPa, they observed a conformational sub-state, where the uric acid substrate binding pocket increased in size at the expense of a localised hydrophobic pocket, which the authors suggest was required to be able to accommodate both intermediate and final products (160). This therefore could be a useful technique to determine potential structural changes that occur within FEN 1 from *Pab* at high pressure.

The mechanisms by which the *Photobacterium profundum* replicates under pressure in these highlighted studies indicate a role of both DiaA and SeqA (positive and negative replication regulators respectively) in controlling levels of replication at high-pressure. However once replication has been initiated at pressure, what changes in the replicative machinery occur to allow it to continue error free? By studying FEN 1 at pressure, one can determine if there are structural changes that could affect function, or at least show that it is tolerant to these pressures.

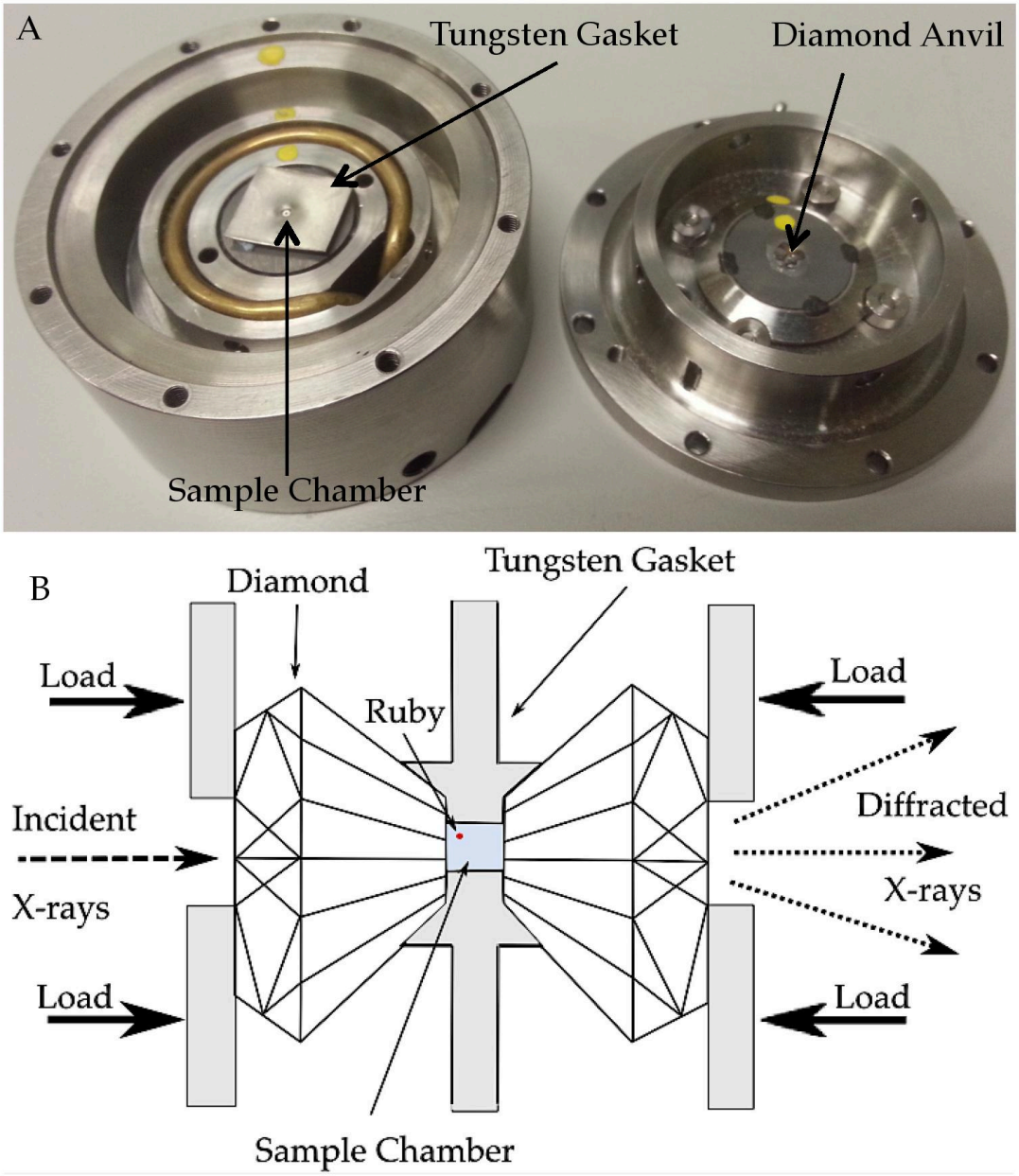


Figure 4.35 - (A) Photograph taken of the DAC during the data collection experiment. Highlighted are the Tungsten gasket, the Diamond Anvil and the Sample Chamber. Also seen in the image is a brass ring that encircles the sample chamber. This plays a role in the temperature regulation of the DAC. (B) Schematic showing the DAC and how it contains the sample during data collection.

Pressure was measured using the ruby shown in Figure 4.35, as discussed in section 2.2.9.1. Due to the position in which the DAC sits in relation to the beam, there was only a 90º rotation possible from -45º to +45º around the vertical axis. This meant that in order to generate a complete data set in order to produce a structure, more crystals had to be used,

and they were placed within the sample chamber in as many varied positions as possible. Upon processing the data, due to the necessity for a greater number of crystals being used, there are more images to merge from different crystals, which led to greater Rmerge values. The intensity of the spots collected is also considerably smaller than that of standard cryo data collection. The Mar555 detector is sensitive enough to detect this reduction, but MOSFLM is not powerful enough to process the data and therefore XDS was used (108). A total of 12 crystals were shot in order to collect a complete dataset. The first crystal was used as a precursor to determine the maximum pressure to which these particular crystals could be subjected to, whilst still diffracting X-rays. This value was 8.6 kBar. Using this value as a benchmark the pressure for the remaining crystals was set at just below 50 % of this value, which was 3.2 kBar.

4.3.2 – Data Collection

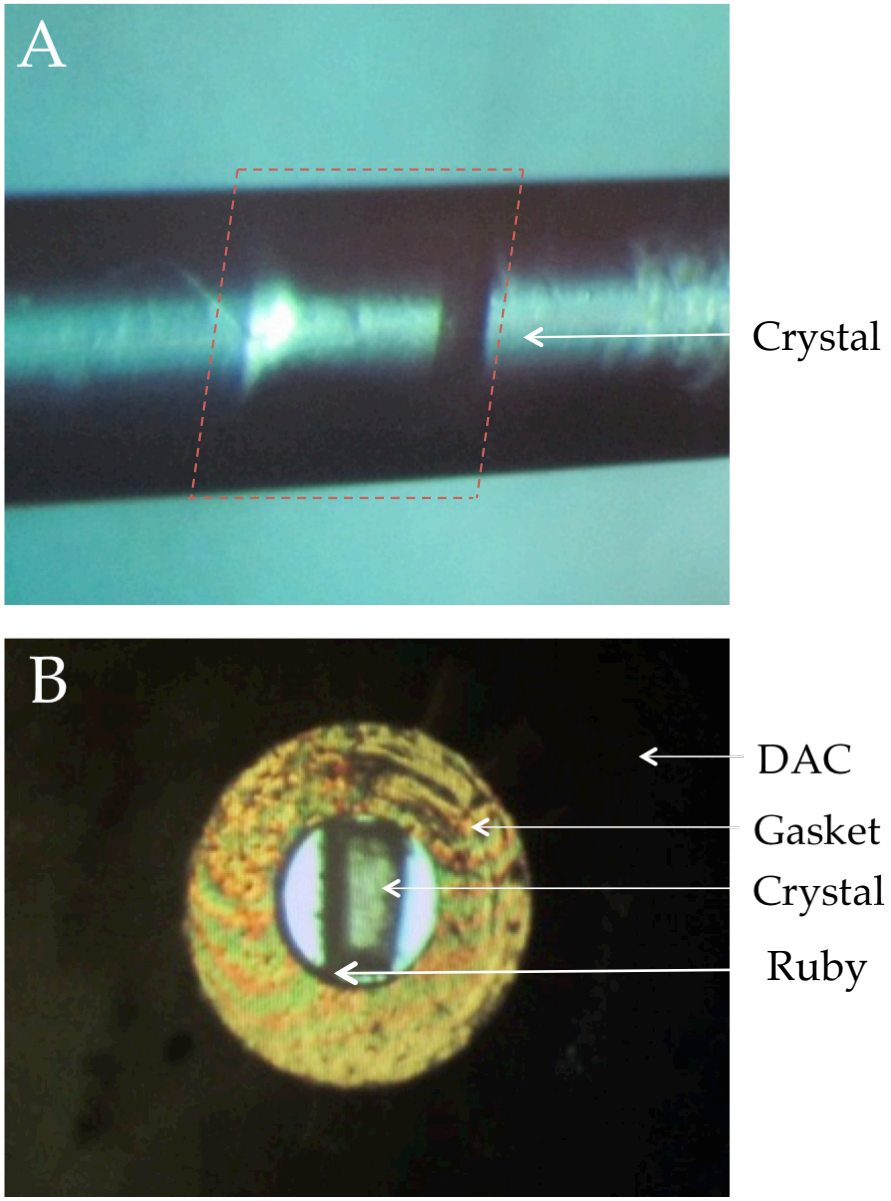


Figure 4.36 - (A) One of the crystals used for the high-pressure data collection housed within a capillary used for transport. (B) The same crystal after being loaded within the DAC prior to the data collection

The crystals were transported in 1 mm capillaries in order to ensure minimal damage to the crystal, whilst maintaining the surrounding mother liquor that it grew in to potentially increase the crystal's stability. Figure 4.36 (A) shows a typical example of the crystal within the capillary. Extraction from the crystal was performed by placing a small drop of mother

liquor onto a cover slip allowing capillary action to move the crystal out, where it was then picked *via* a loop and placed within the DAC. An image of the same crystal placed within the DAC can be seen in figure 4.36 (B), which also highlights the ruby used for pressure measurements. Data collection was then performed as described in the chapter 2.

4.3.3 – Data Processing

Figure 4.37 - Typical example of collected data during the high-pressure experiments. This particular image is from the 7th crystal shot during the experiment. The resolution limit was approximately 3.5 Å.

The crystals used in this experiment were grown in 50 mM $MgCl_2$, 100 mM imidazole/maleate buffer pH 6.5, 15% PEG 8000. Crystals appeared after one week. Although 12 crystals were shot, they had varying degrees of success and a final solution could not be obtained. A large factor in this was the fact that the crystals that were taken for the experiment had larger dimensions than could be accommodated within the DAC. In order

to attempt data collection they were manually broken up into smaller pieces, that were placed inside the chamber one at a time. This introduced significant structural damage and inherent internal damage that affected the quality of the data.

Promising data were collected from one crystal and the statistics for this particular crystal will be discussed. The unit cell dimensions for this particular crystal system were a = 65.18, b = 83.04, c = 235.64, $\alpha = \beta = \gamma = 90°$ with the space group $C222_1$. There was a resolution range from 14.8 to 3.3 Å. The total number of observed reflections was 3436, of which 2607 were unique. It was collected to a completeness of 26.1% with an $I/\sigma I$ of 3.2. The data had an observed R factor of 12.7% with an Rmeas value of 17.5%.

These preliminary high-pressure experiments proved to be a good stepping-stone for further experimentation. Conditions were determined that yielded crystals amenable to both the high pressures and ambient temperatures. Optimisation could include crystal harvesting at an earlier stage, resulting in crystals with the same unit cell dimensions, but smaller overall to fit within the DAC chamber.

Chapter 5

Conclusions & Future Work

5.1 DNA-PKcs

5.1.1 – Overall conclusions of this work

What can be observed based on the results presented throughout this thesis, is that these particular DNA-PKcs constructs are inherently insoluble and even with archetypal techniques employed to increase the solubility; no improvement could be obtained. One particular method that is frequently used to improve the solubility of proteins is to use alternative expression systems. These include yeast, which has the advantage of being a eukaryotic system. Baculovirus infected insect cells can also be used, which are able to process more complex post-translational modifications, which *E. coli* cannot perform. Mammalian cells such as HEK or CHO cells can also be used. They have the ability to offer increased productivity and less variation, however expression levels are lower and the process is more time consuming. Future work would certainly involve the use of alternative expression systems to potentially increase levels of soluble protein. Despite these problems, relative levels of success were achieved using a denature-renature approach to refold the protein in the absence of cellular components in favourable conditions.

Based on the structure published by Sibanda *et al* (47), shown in figure 1.3, it can be seen that the purple kinase domain is nestled within the surrounding domains and it could be

postulated that the outer portions of this domain are actually hydrophobic, preferring to lie within the folded structure. Therefore if this domain is isolated from the surrounding domains, its hydrophobic regions will be exposed to the aqueous buffer and would fold inward on itself in an attempt to avoid the buffer. In order to test this theory, one could generate considerably larger constructs to potentially contain these surrounding regions. One possible new domain boundary could be from S2883 rather than P3600. This would contain the entirety of the FAT domain, rather than approximately half of it in the case of A1. Taking a domain in its entirety like this is more likely to lead to successfully folded protein as the areas in between domains can often be disordered and it can be considered a lot more preferable to start a protein from an unstructured region such as this, rather than from the middle of a structured region, such as an α-helix.

Although solubilisation experiments failed to yield any improvement this result can still be built upon. This is due to the fact that these failed options can be removed from future experiments as variables to be tested. Although a considerable number of solubilisation techniques were tested to no avail, soluble domains have been produced, evidenced by Sajish *et al* (126), discussed in chapter 3.3.3, although these results could not be replicated in this study, through various attempts.

Another result that requires further analysis is the co-purification of both the 60 kDa chaperonin as well as GroEL with the two codon optimised constructs used in this thesis. It stands to reason that within the cellular environment, if a protein requires assistance to maintain its conformation they would exist as some form of complex, but for this complex to be maintained out of the cellular environment and after a barrage of stringent conditions is an interesting observation.

One method that was not performed during this research that could be employed to remove this contamination was described by Rohman *et al* (161) and it involves the addition of denatured host protein to the lysate. The theory being that the unfolded protein will

compete for the binding of the chaperone and therefore allow the chaperone to dissociate from the codon optimised construct.

In future, experiments could be performed to ensure catalytically active protein is refolded. An ATPase assay monitoring the turnover of radio labelled ATP into ADP detecting the release of the γ-^{32}P could be used. Stability assays could involve testing the thermal shift profile at various stages after the purification to determine the stability of the protein over time.

5.1.2 – Optimisation of the purification protocol

As mentioned in chapter 4, a protocol was developed that allowed for the denaturing and refolding of the longest A1 construct (P3600 – M4128). Preliminary biophysical characterisation also showed that one of the refolded species was approximately the same size as the theoretical size of that particular construct. However there were several issues with this protocol, mainly the low yields obtained.

The initial concentration of A1 in the presence of AMP-PNP and 8M Urea was around 38 mg/mL but after refolding this was reduced to 0.3 mg/mL in approximately 100 µL, thus not being enough to perform important biophysical experiments or crystal screens. The largest rates of attrition were consistently seen during the dialysis steps and even though various modifications were tested this could not be improved upon. That being said, there was some degree of improvement by using a combination of drop-wise dilution into an intermediate buffer, containing 2 M guanidine, and then using a Ni-Sepharose HisTrap column as a scaffold for the remaining refold. Time constraints limited the levels of improvement that could be performed on this step but future work would include this optimisation. By decreasing the levels of protein loss at this stage will lead to greater amounts of seemingly correctly folded protein. Adequate biophysical characterisation could then be performed as well as crystal growth trials.

The other issue with this protocol was also the 260 nm/ 280 nm ratio after treatment. This was still at high levels, typically indicating contamination, either from nucleic acids or from pyrimidine nucleotides such as ATP. The former of which however, would be unlikely due to the procedure that the protein went through. Normal protein-DNA interactions would not be able to withstand denaturing, purification and then refolding. The other observation seen during this research was a high 260 nm / 280 nm ratio in relation to chaperone contamination, although a solution to this problem could not be determined. This too is highly unlikely for the same reasons mentioned previously. Therefore this would require adequate investigation, to both determine the source of this 260 nm / 280 nm ratio as well as removing it.

ATP contamination is a more likely in this scenario due to the fact that AMP-PNP is required as a scaffold for the successful refolding of the protein. Due to the protocol failing in the absence of AMP-PNP it will be difficult to test if this hypothesis is correct. One possible technique could involve a radiolabelled AMP-PNP in the refolding process, and subsequent testing of the protein to detect its presence as a reason for the high 260 nm / 280 nm ratio.

5.2 FEN 1

5.2.1 – Overall conclusions of this work

Also presented in this thesis is a novel X-ray crystal structure of FEN 1 from *Pyrococcus abyssi* at 2.27 Å. The structure had two molecules of FEN 1 in the asymmetric unit with each molecule being 44 Å at its widest dimension, and 62 Å at its longest dimension. The results showed that in solution the *Pab* FEN 1 behaves as a monomer. This result was evidenced by both size exclusion chromatography, as well as dynamic light scattering experiments. Although this is an *in vitro* result it would make sense that this is observed *in vivo* too, based on the theorised mode of action. Due to FEN 1 being a structure-specific endonuclease, then

only one FEN 1 molecule will be able to bind to one particular flap at any given time, therefore if it were a dimer, the second copy of the FEN 1 would not be able to interact with the DNA at the same time as the first, in fact being a large globular protein it is more likely that it would in fact sterically hinder the FEN 1 from accessing the DNA, a reason that could potentially indicate why no FEN 1 – DNA complex was being seen in this research. This theory however is something that would require experimentation.

Subsequent improvements on this structure would start with attempting to generate a structure containing the helical archway. This portion of the structure was observed in the *Sso* and human structures and is thought to be integral to the function of the enzyme, specifically the DNA recognition and binding. The lack of this region has been attributed to disorder.

Future work will include further attempts at using the coordinates from *Pfu* FEN 1 as a search model for MR. Due to the 90.3% sequence identity it stands to reason that this would have been the most suitable search model, however when attempts were made during this research, significant clashes were observed during rotation and translation functions. This will most likely be due to the presence of the two regions shown in figure 4.18 in the search model, but with no corresponding density obtained from the experimental data. By removing these portions from the search model prior to performing the MR the phases from the remaining structure would correlate better with the experimental data, when compared to the *Sso* structure used. The removed regions could then be replaced during the iterative cycles of refinement to potentially improve the final structure.

Structural alignments between the *Pab* structure and the human showed that they potentially share their mode of interaction with DNA with one another. This theory was strengthened based on the surface charge distribution, which showed a positively charged channel running along the proposed region where the DNA would interact with the protein.

This theory would therefore need to be validated, which was why attempts were made to generate a FEN 1–DNA complexed crystal structure.

5.2.2 – Determining a crystal structure of FEN 1 in complex with DNA.

The results in this thesis show that pure FEN 1 was being produced and that a DNA complex was being formed from the three oligonucleotides designed, based on similar DNA published by Tsutakawa *et al* (137). However the required complex could not be formed between the protein and nucleic acid species. There could be a variety of reasons for this.

One of these reasons could be that the DNA being used was not pure enough for crystallisation purposes. The doubling of peaks present in figure 4.21 most likely indicates the presence of several species and this heterogeneity could lead to ineffective crystal packing. Future work would include an additional purification step after the DNA complex formation prior to any crystallisation.

Future work would also include additional analysis performed on the crystals being tested in order to determine whether or not any DNA was actually present within the crystal. This analysis would include mass spectrometry on a dissolved crystal as well as detecting the 260 nm / 280 nm ratio, also of a dissolved crystal to detect the presence of nucleic acid.

The data presented in chapter 4.2.6 shows the data collection and processing strategy to solve the structure for a FEN 1-DNA complex. After molecular replacement the top solution had a Z-score of 8.9, with an LLG of 1536.4. This in combination with an R factor after one round of TLS and restrained refinement of 29.47% indicates that the solution is correct. In light of this, it can be stated with confidence that the structure lacks electron density for DNA. Theoretical models using coordinates from the human FEN 1 structure containing DNA also showed that there was not enough space for a DNA double helix, assuming that the binding mode of action is similar.

One interesting aspect of this solution, however, is that even though a complex between the protein and nucleic acid did not form, the crystal grew in an alternative space group upon addition of the DNA even though they were grown in identical conditions. This is also something that can be expanded upon in future work.

5.2.3 – Determining a crystal structure of FEN 1 at high pressure.

The crystallography experiments performed at high pressure were only preliminary, and they serve as a springboard for future work attempting to solve the structure at high pressure. Important results obtained included the determination of crystals that favoured ambient high-pressure data collection but were poor for cryo data collection and *vice versa*. Although the crystal growth conditions were different for both the ambient and high-pressure experiments, the initial statistics indicated that they had the same space group and unit cell dimensions. That being said the high-pressure conditions should be subjected to further optimisation in any future experimentation.

Aside from archetypal optimisation, whereby slight modifications in the conditions are made to generate more suitable crystal types, optimisation could include an earlier crystal harvest. The crystals that were harvested for these preliminary experiments had dimensions larger than that of the DAC sample chamber. This led to the crystals having to be manually broken up to fit into the chamber. This introduced cracks and other deformities that compromised the integrity and quality of the data collected. Harvesting the crystals earlier would mean that the crystals would be smaller, and therefore more likely to fit within the chamber. This would mean that no damage is introduced to the crystal before the data collection generates radiation damage potentially increasing the quality of the data as well as the lifespan of the crystal.

References

1. Lodish H, Berk A, Matsudaira P, Caiser CA, Krieger M, Scott MP, et al. Molecular Cell Biology. 5th ed ed. Freeman WH, editor. New York, NY2004.

2. Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. Journal of Biological Chemistry. 1992;267(1):166-72.

3. Yarosh DB. The role of O6-methylguanine-DNA methyltransferase in cell survival, mutagenesis and carcinogenesis. Mutation Research/DNA Repair Reports. 1985;145(1–2):1-16.

4. Norbury CJ, Hickson ID. CELLULAR RESPONSES TO DNA DAMAGE. Annual Review of Pharmacology and Toxicology. 2001;41(1):367-401.

5. Deans AJ, West SC. DNA interstrand crosslink repair and cancer. Nat Rev Cancer. 2011;11(7):467-80.

6. Dronkert MLG, Kanaar R. Repair of DNA interstrand cross-links. Mutation Research/DNA Repair. 2001;486(4):217-47.

7. Noll DM, Noronha AM, Wilds CJ, Miller PS. Preparation of interstrand cross-linked DNA oligonucleotide duplexes. Front Biosci [Internet]. 2004 2004/01//; 9:[421-37 pp.]. Available from: http://europepmc.org/abstract/MED/14766379 http://dx.doi.org/10.2741/1246.

8. Xue W, Warshawsky D. Metabolic activation of polycyclic and heterocyclic aromatic hydrocarbons and DNA damage: A review. Toxicology and Applied Pharmacology. 2005;206(1):73-93.

9. Formenton-Catai AP, Machado RP, Lanças FM, Carrilho E. Solid-Phase purification of deoxyguanosine-benzo[a]pyrene diol epoxide adducts from genomic DNA adduct synthesis. Journal of the Brazilian Chemical Society. 2005;16:808-14.

10. Friedberg EC. DNA damage and repair. Nature. 2003;421(6921):436-40.

11. Featherstone C, Jackson SP. DNA double-strand break repair. Current biology. 1999;9(20):R759-R61.

12. Jackson SP. Sensing and repairing DNA double-strand breaks. Carcinogenesis. 2002;23(5):687-96.

13. Rothkamm K, Krüger I, Thompson LH, Löbrich M. Pathways of DNA Double-Strand Break Repair during the Mammalian Cell Cycle. Molecular and Cellular Biology. 2003;23(16):5706-15.

14. Schatz DG. V(D)J recombination. Immunological Reviews. 2004;200(1):5-11.

15. ADAMSON ED. Oncogenes in development. Development. 1987;99(4):449-71.

16. Chial H. Proto-oncogenes to oncogenes to cancer. Nature Education. 2008;1(1):33.

17. Hirohashi S, Kanai Y. Cell adhesion system and human cancer morphogenesis. Cancer Science. 2003;94(7):575-81.

18. Buermeyer AB, Deschênes SM, Baker SM, Liskay RM. MAMMALIAN DNA MISMATCH REPAIR. Annual Review of Genetics. 1999;33(1):533-64.

19. Kunkel TA, Erie DA. DNA MISMATCH REPAIR*. Annual Review of Biochemistry. 2005;74(1):681-710.

20. de Laat WL, Jaspers NGJ, Hoeijmakers JHJ. Molecular mechanism of nucleotide excision repair. Genes & Development. 1999;13(7):768-85.

21. Aboussekhra A, Biggerstaff M, Shivji MK, Vilpo JA, Moncollin V, Podust VN, et al. Mammalian DNA nucleotide excision repair reconstituted with purified protein components. Cell. 1995;80(6):859-68.

22. Krokan HE, Bjørås M. Base Excision Repair. Cold Spring Harbor Perspectives in Biology. 2013;5(4).

23. Shibata A, Barton O, Noon AT, Dahm K, Deckbar D, Goodarzi AA, et al. Role of ATM and the Damage Response Mediator Proteins 53BP1 and MDC1 in the Maintenance of G2/M Checkpoint Arrest. Molecular and Cellular Biology. 2010;30(13):3371-83.

24. Wyman C, Kanaar R. DNA Double-Strand Break Repair: All's Well that Ends Well. Annual Review of Genetics. 2006;40(1):363-83.

25. Bartek J, Lukas C, Lukas J. Checking on DNA damage in S phase. Nat Rev Mol Cell Biol. 2004;5(10):792-804.

26. O'Connell MJ, Walworth NC, Carr AM. The G2-phase DNA-damage checkpoint. Trends in Cell Biology. 2000;10(7):296-303.

27. Sørensen CS, Hansen LT, Dziegielewski J, Syljuåsen RG, Lundin C, Bartek J, et al. The cell-cycle checkpoint kinase Chk1 is required for mammalian homologous recombination repair. Nature Cell Biology. 2005;7(2):195-201.

28. Khanna KK, Jackson SP. DNA double-strand breaks: signaling, repair and the cancer connection. Nat Genet. 2001;27(3):247-54.

29. Salles-Passador I, Fotedar A, Fotedara R. Cellular response to DNA damage. Link between p53 and DNA-PK. Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie. 1999;322(2):113-20.

30. Critchlow SE, Jackson SP. DNA end-joining: from yeast to man. Trends in Biochemical Sciences. 1998;23(10):394-8.

31. Walker JR, Corpina RA, Goldberg J. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature. 2001;412(6847):607-14.

32. Spagnolo L, Rivera-Calzada A, Pearl LH, Llorca O. Three-Dimensional Structure of the Human DNA-PKcs/Ku70/Ku80 Complex Assembled on DNA and Its Implications for DNA DSB Repair. Molecular Cell. 2006;22(4):511-9.

33. Dynan WS, Yoo S. Interaction of Ku protein and DNA-dependent protein kinase catalytic subunit with nucleic acids. Nucleic Acids Research. 1998;26(7):1551-9.

34. Yaneva M, Kowalewski T, Lieber MR. Interaction of DNA-dependent protein kinase with DNA and with Ku: biochemical and atomic-force microscopy studies. The EMBO journal. 1997;16(16):5098-112.

35. Dvir A, Peterson SR, Knuth MW, Lu H, Dynan WS. Ku autoantigen is the regulatory component of a template-associated protein kinase that phosphorylates RNA polymerase II. Proceedings of the National Academy of Sciences. 1992;89(24):11920-4.

36. Davis AJ, Chen BPC, Chen DJ. DNA-PK: A dynamic enzyme in a versatile DSB repair pathway. DNA repair. 2014;17(0):21-9.

37. Neal JA, Meek K. Choosing the right path: does DNA-PK help make the decision? Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2011;711(1):73-86.

38. Ma Y, Pannicke U, Schwarz K, Lieber MR. Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination. Cell. 2002;108(6):781-94.

39. Goodarzi AA, Yu Y, Riballo E, Douglas P, Walker SA, Ye R, et al. DNA-PK autophosphorylation facilitates Artemis endonuclease activity2006 2006-08-23 00:00:00. 3880-9 p.

40. Chappell C, Hanakahi LA, Karimi-Busheri F, Weinfeld M, West SC. Involvement of human polynucleotide kinase in double-strand break repair by non-homologous end joining2002 2002-06-03 00:00:00. 2827-32 p.

41. De Ioannes P, Malu S, Cortes P, Aggarwal Aneel K. Structural Basis of DNA Ligase IV-Artemis Interaction in Nonhomologous End-Joining. Cell Reports. 2012;2(6):1505-12.

42. Costantini S, Woodbine L, Andreoli L, Jeggo PA, Vindigni A. Interaction of the Ku heterodimer with the DNA ligase IV/Xrcc4 complex and its regulation by DNA-PK. DNA repair. 2007;6(6):712-22.

43. Hsu H-L, Yannone SM, Chen DJ. Defining interactions between DNA-PK and ligase IV/XRCC4. DNA repair. 2002;1(3):225-35.

44. Ahnesorg P, Smith P, Jackson SP. XLF Interacts with the XRCC4-DNA Ligase IV Complex to Promote DNA Nonhomologous End-Joining. Cell. 2006;124(2):301-13.

45. Hammel M, Yu Y, Mahaney BL, Cai B, Ye R, Phipps BM, et al. Ku and DNA-dependent Protein Kinase Dynamic Conformations and Assembly Regulate DNA Binding and the Initial Non-homologous End Joining Complex. Journal of Biological Chemistry. 2010;285(2):1414-23.

46. Ding Q, Reddy YV, Wang W, Woods T, Douglas P, Ramsden DA, et al. Autophosphorylation of the catalytic subunit of the DNA-dependent protein kinase is required for efficient end processing during DNA double-strand break repair. Molecular and cellular biology. 2003;23(16):5836-48.

47. Sibanda BL, Chirgadze DY, Blundell TL. Crystal structure of DNA-PKcs reveals a large open-ring cradle comprised of HEAT repeats. Nature. 2010;463(7277):118-21.

48. Abraham RT. Cell cycle checkpoint signaling through the ATM and ATR kinases. Genes & Development. 2001;15(17):2177-96.

49. Chen H, Ma Z, Vanderwaal RP, Feng Z, Gonzalez-Suarez I, Wang S, et al. The mTOR Inhibitor Rapamycin Suppresses DNA Double-Strand Break Repair. Radiation Research. 2010;175(2):214-24.

50. Hefferin ML, Tomkinson AE. Mechanism of DNA double-strand break repair by non-homologous end joining. DNA repair. 2005;4(6):639-48.

51. Summers KC, Shen F, Sierra Potchanant EA, Phipps EA, Hickey RJ, Malkas LH. Phosphorylation: The Molecular Switch of Double-Strand Break Repair. International Journal of Proteomics. 2011;2011.

52. Cui X, Yu Y, Gupta S, Cho Y-M, Lees-Miller SP, Meek K. Autophosphorylation of DNA-Dependent Protein Kinase Regulates DNA End Processing and May Also Alter Double-Strand Break Repair Pathway Choice. Molecular and Cellular Biology. 2005;25(24):10842-52.

53. Graham KL, Gustin KE, Rivera C, Kuyumcu-Martinez NM, Choe SS, Lloyd RE, et al. Proteolytic cleavage of the catalytic subunit of DNA-dependent protein kinase during poliovirus infection. Journal of virology. 2004;78(12):6313-21.

54. Sarbassov dD, Ali SM, Sabatini DM. Growing roles for the mTOR pathway. Current opinion in cell biology. 2005;17(6):596-603.

55. Stiff T, O'Driscoll M, Rief N, Iwabuchi K, Löbrich M, Jeggo PA. ATM and DNA-PK Function Redundantly to Phosphorylate H2AX after Exposure to Ionizing Radiation. Cancer Research. 2004;64(7):2390-6.

56. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research. 2011;39(suppl 1):D225-D9.

57. Douglas P, Cui X, Block WD, Yu Y, Gupta S, Ding Q, et al. The DNA-dependent protein kinase catalytic subunit is phosphorylated in vivo on threonine 3950, a highly conserved amino acid in the protein kinase domain. Molecular and cellular biology. 2007;27(5):1581-91.

58. Rivera-Calzada A, Maman JP, Spagnolo L, Pearl LH, Llorca O. Three-Dimensional Structure and Regulation of the DNA-Dependent Protein Kinase Catalytic Subunit (DNA-PKcs). Structure. 2005;13(2):243-55.

59. Schrödinger. The PyMOL Molecular Graphics System.Version 1.5.0.4

60. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, et al. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Research. 2013;41(W1):W597-W600.

61. Williams DR, Lee K-J, Shi J, Chen DJ, Stewart PL. Cryo-EM structure of the DNA-dependent protein kinase catalytic subunit at subnanometer resolution reveals α helices and insight into DNA binding. Structure. 2008;16(3):468-77.

62. Kurimasa A, Kumano S, Boubnov NV, Story MD, Tung C-S, Peterson SR, et al. Requirement for the kinase activity of human DNA-dependent protein kinase catalytic subunit in DNA strand break rejoining. Molecular and cellular biology. 1999;19(5):3877-84.

63. Neal JA, Dang V, Douglas P, Wold MS, Lees-Miller SP, Meek K. Inhibition of homologous recombination by DNA-dependent protein kinase requires kinase activity, is titratable, and is modulated by autophosphorylation. Molecular and cellular biology. 2011;31(8):1719-33.

64. Bailey SM, Brenneman MA, Halbrook J, Nickoloff JA, Ullrich RL, Goodwin EH. The kinase activity of DNA-PK is required to protect mammalian telomeres. DNA repair. 2004;3(3):225-33.

65. de Lange T. Protection of mammalian telomeres. Oncogene. 2002;21(4):532-40.

66. Bailey SM, Meyne J, Chen DJ, Kurimasa A, Li GC, Lehnert BE, et al. DNA double-strand break repair proteins are required to cap the ends of mammalian chromosomes. Proceedings of the National Academy of Sciences. 1999;96(26):14899-904.

67. Davies TG, Bentley J, Arris CE, Boyle FT, Curtin NJ, Endicott JA, et al. Structure-based design of a potent purine-based cyclin-dependent kinase inhibitor. Nature Structural & Molecular Biology. 2002;9(10):745-9.

68. Hollick JJ, Rigoreau LJ, Cano-Soumillac C, Cockcroft X, Curtin NJ, Frigerio M, et al. Pyranone, thiopyranone, and pyridone inhibitors of phosphatidylinositol 3-kinase related kinases. Structure-activity relationships for DNA-dependent protein kinase inhibition, and identification of the first potent and selective inhibitor of the ataxia telangiectasia mutated kinase. Journal of medicinal chemistry. 2007;50(8):1958-72.

69. Bain J, Plater L, Elliott M, Shpiro N, Hastie C J, McLauchlan H, et al. The selectivity of protein kinase inhibitors: a further update. The Biochemical Journal. 2007;408(Pt 3):297-315.

70. Sun X, Yang C, Liu H, Wang Q, Wu S-X, Li X, et al. Identification and Characterization of a Small Inhibitory Peptide That Can Target DNA-PKcs Autophosphorylation and Increase Tumor Radiosensitivity. International Journal of Radiation Oncology*Biology*Physics. 2012;84(5):1212-9.

71. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences. 1990;87(12):4576-9.

72. M.T. M, J.M. M, D.A. S, D.P C. Brock Biology of Microorganisms 2013.

73. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 1977;74(11):5088-90.

74. Barns SM, Delwiche CF, Palmer JD, Pace NR. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proceedings of the National Academy of Sciences. 1996;93(17):9188-93.

75. Olsen GJ, Woese CR. Archaeal Genomics: An Overview. Cell. 1997;89(7):991-4.

76. Erauso G, Reysenbach A-L, Godfroy A, Meunier J-R, Crump B, Partensky F, et al. Pyrococcus abyssi sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. Archives of microbiology. 1993;160(5):338-49.

77. Bramhill D, Kornberg A. Duplex opening by dnaA protein at novel sequences in initiation of replication at the origin of the E. coli chromosome. Cell. 1988;52(5):743-55.

78. Mott ML, Berger JM. DNA replication initiation: mechanisms and regulation in bacteria. Nat Rev Micro. 2007;5(5):343-54.

79. Robinson A, van Oijen AM. Bacterial replication, transcription and translation: mechanistic insights from single-molecule biochemical studies. Nat Rev Micro. 2013;11(5):303-15.

80. Bell SP, Dutta A. DNA REPLICATION IN EUKARYOTIC CELLS. Annual Review of Biochemistry. 2002;71(1):333-74.

81. Takeda DY, Dutta A. DNA replication and progression through S phase. 2005;24(17):2827-43.

82. Matsunaga F, Forterre P, Ishino Y, Myllykallio H. In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. Proceedings of the National Academy of Sciences. 2001;98(20):11152-7.

83. Robinson NP, Dionne I, Lundgren M, Marsh VL, Bernander R, Bell SD. Identification of Two Origins of Replication in the Single Chromosome of the Archaeon< i> Sulfolobus solfataricus</i>. Cell. 2004;116(1):25-38.

84. Kelman LM, Kelman Z. Multiple origins of replication in archaea. Trends in Microbiology. 2004;12(9):399-401.

85. Barry ER, Bell SD. DNA Replication in the Archaea. Microbiology and Molecular Biology Reviews. 2006;70(4):876-87.

86. Shin J-H, Heo GY, Kelman Z. The Methanothermobacter thermautotrophicus Cdc6-2 Protein, the Putative Helicase Loader, Dissociates the Minichromosome Maintenance Helicase. Journal of Bacteriology. 2008;190(11):4091-4.

87. Marinsek NBERMKSDIKEVBSD. GINS, a central nexus in the archaeal DNA replication fork. EMBO reports. 2006;7(5):539-45.

88. Waga S, Stillman B. THE DNA REPLICATION FORK IN EUKARYOTIC CELLS. Annual Review of Biochemistry. 1998;67(1):721-51.

89. Ishino Y, Ishino S. DNA Replication in Archaea, the Third Domain of Life. 2013.

90. Craggs TD, Hutton RD, Brenlla A, White MF, Penedo JC. Single-molecule characterization of Fen1 and Fen1/PCNA complexes acting on flap substrates. Nucleic Acids Research. 2014;42(3):1857-72.

91. Kucherlapati M, Yang K, Kuraguchi M, Zhao J, Lia M, Heyer J, et al. Haploinsufficiency of Flap endonuclease (Fen1) leads to rapid tumor progression. Proceedings of the National Academy of Sciences. 2002;99(15):9924-9.

92. MacNeill SA. DNA replication: Partners in the Okazaki two-step. Current Biology. 2001;11(20):R842-R4.

93. Balakrishnan L, Bambara RA. Flap endonuclease 1. Annual review of biochemistry. 2013;82:119.

94. Chapados BR, Hosfield DJ, Han S, Qiu J, Yelent B, Shen B, et al. Structural Basis for FEN-1 Substrate Specificity and PCNA-Mediated Activation in DNA Replication and Repair. Cell. 2004;116(1):39-50.

95. Kao H-I, Henricksen LA, Liu Y, Bambara RA. Cleavage specificity of Saccharomyces cerevisiae flap endonuclease 1 suggests a double-flap structure as the cellular substrate. Journal of Biological Chemistry. 2002;277(17):14379-89.

96. Rossi ML, Bambara RA. Reconstituted Okazaki Fragment Processing Indicates Two Pathways of Primer Removal. Journal of Biological Chemistry. 2006;281(36):26051-61.

97. Ryu G-H, Tanaka H, Kim D-H, Kim J-H, Bae S-H, Kwon Y-N, et al. Genetic and biochemical analyses of Pfh1 DNA helicase function in fission yeast. Nucleic Acids Research. 2004;32(14):4205-16.

98. Asagoshi K, Tano K, Chastain PD, Adachi N, Sonoda E, Kikuchi K, et al. FEN1 functions in long patch base excision repair under conditions of oxidative stress in vertebrate cells. Molecular Cancer Research. 2010;8(2):204-15.

99. Meersman F, McMillan PF. High hydrostatic pressure: a probing tool and a necessary parameter in biophysical chemistry. Chemical Communications. 2014;50(7):766-75.

100. Kharakoz DP. Protein compressibility, dynamics, and pressure. Biophysical journal. 2000;79(1):511-25.

101. Millipore M. 69663 | pET-17b DNA - Novagen 2015 [cited 2015 02/03/2015]. Available from: http://www.merckmillipore.com/GB/en/product/pET-17b-DNA---Novagen,EMD_BIO-69663 - anchor_CITATION.

102. Asherie N. Protein crystallization and phase diagrams. Methods. 2004;34(3):266-72.

103. Matthews BW. Solvent Content of Protein Crystals. J Mol Bio. 1968;33:491-7.

104. Kantardjieff KA, Rupp B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. Protein science : a publication of the Protein Society. 2003;12(9):1865-71.

105. Chirgadze D. The theory of the Molecular Replacement method http://www.xray.bioc.cam.ac.uk/xray_resources/whitepapers/mr-in-action/node4.html2001 [cited 2015 18/02/2015].

106. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. Journal of Applied Crystallography. 2007;40(4):658-74.

107. Fourme R, Kahn R, Mezouar M, Girard E, Hoerentrup C, Prangè T, et al. J Synchrotron Rad. 2001;8:1149-56.

108. Kabsch W. Xds. Acta Crystallographica Section D: Biological Crystallography. 2010;66(2):125-32.

109. Bailey S. The CCP4 suite: programs for protein crystallography: Daresbury Laboratory; 1993.

110. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. Acta Crystallographica Section D. 2011;67(4):235-42.

111. Mao HK, Bell PM, Shaner JW, Steinberg DJ. Specific volume measurements of Cu, Mo, Pd, and Ag and calibration of the ruby R1 fluorescence pressure gauge from 0.06 to 1 Mbar. Journal of Applied Physics. 1978;49(6):3276-83.

112. Fourme R, Girard E, Kahn R, Dhaussy A-C, Mezouar M, Colloc'h N, et al. High-Pressure Macromolecular Crystallography (HPMX): Status and prospects. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 2006;1764(3):384-90.

113. Powell AGWLaHR. Processing Diffraction Data with Mosflm. Evolving Methods for Macromolecular Crystallography. 2007;245:41-51.

114. Lawrence AK, Michael JES. Protein structure prediction on the Web: a case study using the Phyre server. Nature Protocols. 2009;4(3):363-71.

115. Connelly MA, Zhang H, Kieleczawa J, Anderson CW. Alternate splice-site utilization in the gene for the catalytic subunit of the DNA-activated protein kinase, DNA-PKcs. Gene. 1996;175(1–2):271-3.

116. Rocha EPC. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. Genome Research. 2004;14(11):2279-86.

117. Kane JF. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. Current opinion in biotechnology. 1995;6(5):494-500.

118. Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O. Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. Protein Expression and Purification. 2008;59(1):94-102.

119. Fisher T. GeneOptimizer® Process for Multi-parameter Gene Optimization 2011 [cited 2012 02/05/2012].

120. Kishi A, Nakamura T, Nishio Y, Maegawa H, Kashiwagi A. Sumoylation of Pdx1 is associated with its nuclear localization and insulin gene activation. American Journal of Physiology - Endocrinology and Metabolism. 2003;284(4):E830-E40.

121. Costa S, Almeida A, Castro A, Domingues L. Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system. Frontiers in Microbiology. 2014;5:63.

122. Terpe K. Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. Appl Microbiol Biotechnol. 2003;60:523-33.

123. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research. 2003;31(13):3784-8.

124. DOUDNA JA, Lucast LJ, Batey RT. Mutant proteinase with reduced self-cleavage activity and method of purification. Google Patents; 2012.

125. Holler C, Vaughan D, Zhang C. Polyethyleneimine precipitation versus anion exchange chromatography in fractionating recombinant β-glucuronidase from transgenic tobacco extract. Journal of Chromatography A. 2007;1142(1):98-105.

126. Sajish M, Zhou Q, Kishi S, Valdez DM, Jr., Kapoor M, Guo M, et al. Trp-tRNA synthetase bridges DNA-PKcs to PARP-1 to link IFN-gamma and p53 signaling. Nature chemical biology. 2012;8(6):547-54.

127. Lei S-P, Lin H, Wang S-S, Callaway J, Wilcox G. Characterization of the Erwinia carotovora pelB gene and its product pectate lyase. Journal of bacteriology. 1987;169(9):4379-83.

128. Singh P, Sharma L, Kulothungan SR, Adkar BV, Prajapati RS, Ali PSS, et al. Effect of Signal Peptide on Stability and Folding of Escherichia coli Thioredoxin. PLoS ONE. 2013;8(5):e63442.

129. Natividad R, Daniel K, Thomas JS. Advances in understanding bacterial outer-membrane biogenesis. Nature Reviews Microbiology. 2006;4(1):57-66.

130. Reynaud E. Protein Misfolding and Degenerative Diseases. Nature Education. 2010;3(9):28.

131. Lin MM, Mohammed OF, Jas GS, Zewail AH. Speed limit of protein folding evidenced in secondary structure dynamics. Proceedings of the National Academy of Sciences. 2011;108(40):16622-7.

132. Upadhyay AK, Murmu A, Singh A, Panda AK. Kinetics of Inclusion Body Formation and Its Correlation with the Characteristics of Protein Aggregates in Escherichia coli. PLoS ONE. 2012;7(3):e33951.

133. Salvi G, De Los Rios P, Vendruscolo M. Effective interactions between chaotropic agents and proteins. Proteins: Structure, Function, and Bioinformatics. 2005;61(3):492-9.

134. Greene RF, Pace CN. Urea and Guanidine Hydrochloride Denaturation of Ribonuclease, Lysozyme, α-Chymotrypsin, and β-Lactoglobulin. Journal of Biological Chemistry. 1974;249(17):5388-93.

135. Rashid F, Sharma S, Bano B. Comparison of guanidine hydrochloride (GdnHCl) and urea denaturation on inactivation and unfolding of human placental cystatin (HPC). The protein journal. 2005;24(5):283-92.

136. Hohl M, Dunand-Sauthier I, Staresincic L, Jaquier-Gubler P, Thorel F, Modesti M, et al. Domain swapping between FEN-1 and XPG defines regions in XPG that mediate nucleotide excision repair activity and substrate specificity. Nucleic Acids Research. 2007;35(9):3053-63.

137. Tsutakawa Susan E, Classen S, Chapados Brian R, Arvai AS, Finger LD, Guenther G, et al. Human Flap Endonuclease Structures, DNA Double-Base Flipping, and a Unified Understanding of the FEN1 Superfamily. Cell. 2011;145(2):198-211.

138. Selleck W, Tan S. Recombinant Protein Complex Expression in E. coli. Current Protocols in Protein Science: John Wiley & Sons, Inc.; 2001.

139. Dore AS, Kilkenny ML, Jones SA, Oliver AW, Roe SM, Bell SD, et al. Structure of an archaeal PCNA1-PCNA2-FEN1 complex: elucidating PCNA subunit and client enzyme specificity. Nucleic Acids Res. 2006;34(16):4515-26.

140. Schwarzenbacher R, Godzik A, Grzechnik SK, Jaroszewski L. The importance of alignment accuracy for molecular replacement. Acta crystallographica Section D, Biological crystallography. 2004;60(Pt 7):1229-36.

141. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302(1):205-17.

142. Murshudov GN, Vagin AA, Dodson EJ. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. Acta Crystallographica Section D. 1997;53(3):240-55.

143. Karplus PA, Diederichs K. Linking Crystallographic Model and Data Quality. Science. 2012;336(6084):1030-3.

144. Diederichs K, Karplus PA. Better models by discarding data? Acta crystallographica Section D, Biological crystallography. 2013;69(Pt 7):1215-22.

145. Emsley P CK. Coot: model-building tools for molecular graphics. . Acta Crystallogr D. 2004:2126-32.

146. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology. 1963;7(1):95-9.

147. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D: Biological Crystallography. 2010;66(Pt 1):12-21.

148. de Beer TA, Berka K, Thornton JM, Laskowski RA. PDBsum additions. Nucleic Acids Res. 2014;42(Database issue):D292-6.

149. Orengo C A FTP, Taylor W R & Thornton J M. Identification and classification of protein fold families. Protein Eng. 1993;6:485-500.

150. Henrick EKaK. Protein interfaces, surfaces and assemblies' service PISA at the European Bioinformatics Institute 'Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007;372:774-97.

151. Kim K, Biade S, Matsumoto Y. Involvement of flap endonuclease 1 in base excision DNA repair. Journal of Biological Chemistry. 1998;273(15):8842-8.

152. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: Application to microtubules and the ribosome. Proceedings of the National Academy of Sciences. 2001;98(18):10037-41.

153. Lieber MR. The FEN-1 family of structure-specific nucleases in eukaryotic dna replication, recombination and repair. BioEssays. 1997;19(3):233-40.

154. Sinden RR. DNA structure and Function. 1st ed: Academic Press; 1994.

155. Hosfield DJ, Mol CD, Shen B, Tainer JA. Structure of the DNA repair and replication endonuclease and exonuclease FEN-1: coupling DNA and PCNA binding to FEN-1 activity. Cell. 1998;95(1):135-46.

156. Qiagen. Critical Factors For Successful Protein Crystallisation2010 19.06.14.

157. James D. Watson, Tania A. Baker, Stephen P. Bell, Alexander Gann, Michael Levine, Losick R. Molecular Biology of the Gene. 5th ed. Pearson, editor: Benjamin Cummins; 2003.

158. El-Hajj ZW, Tryfona T, Allcock DJ, Hasan F, Lauro FM, Sawyer L, et al. Importance of Proteins Controlling Initiation of DNA Replication in the Growth of the High-Pressure-Loving Bacterium Photobacterium profundum SS9. Journal of Bacteriology. 2009;191(20):6383-93.

159. Takahashi H, Yamaguchi T, Koga M, Kageura H, Terada S. DNA replication reaction in Xenopus cell-free system is suppressed by high pressure. Cellular & molecular biology letters. 2004;9(3):423-7.

160. Girard E, Marchal S, Perez J, Finet S, Kahn R, Fourme R, et al. Structure-function perturbation and dissociation of tetrameric urate oxidase by high hydrostatic pressure. Biophys J. 2010;98(10):2365-73.

161. Rohman M, Harrison-Lavoie KJ. Separation of Copurifying GroEL from Glutathione-S-Transferase Fusion Proteins. Protein Expression and Purification. 2000;20(1):45-7.

162. Taylor G. The phase problem. Acta Crystallographica Section D. 2003;59(11):1881-90.

163. Taylor GL. Introduction to phasing. Acta Crystallographica Section D. 2010;66(4):325-38.

Appendix

Alternative methods to solve the phase problem

Isomorphous replacement works on the principle of detecting visible changes in intensity as a result of soaking the crystal in a solution containing 'heavy' atoms, by which they are atoms with a large atomic number.

The presence of the heavy atom significantly changes the structure factors associated with the data.

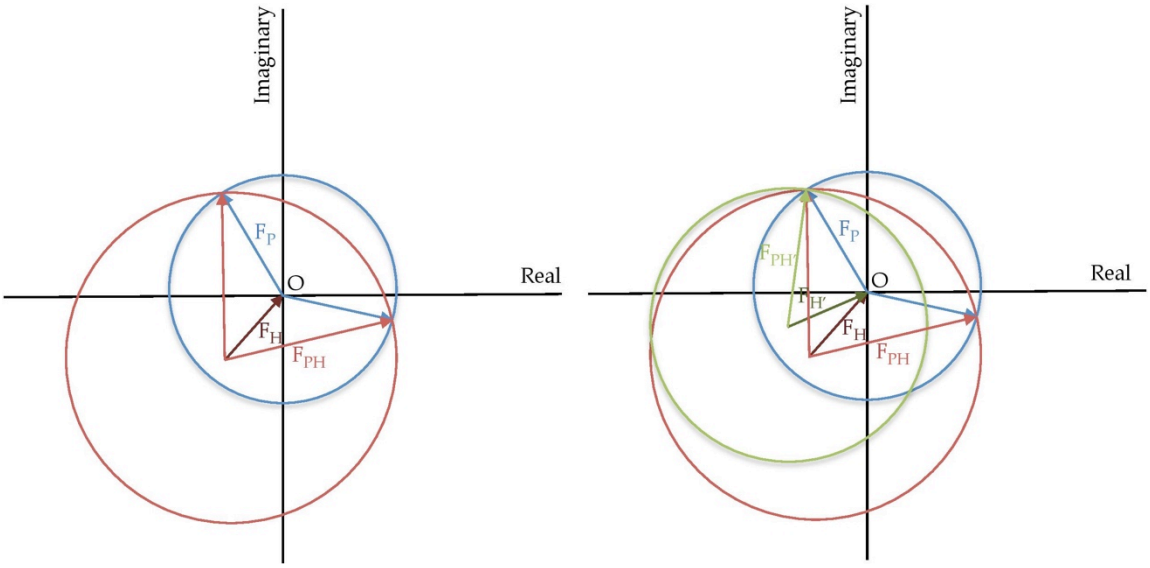


Figure A.1 - Harker constructs demonstrating both single isomorphous replacement (SIR) and multiple isomorphous replacement (MIR). F_P relates to the structure factor for the protein, which remains unknown, F_H relates to the heavy atom structure factor and F_{PH} relates to the structure factor for the derivative crystal containing the heavy atom.

Making use of this phenomenon to collect both a native data set and a derivative set containing a heavy atom set allows direct comparisons to be made and a location for the heavy atoms within the structure to be identified. This in turn then allows for the determination of the structure factor for the heavy atoms (F_H).

Combining this information with the amplitude information about both datasets into a Harker construct as seen in figure 2.5, where F_{PH} is the structure factor for the derivative crystal containing the heavy atom and F_P is the as yet unknown structure factor for the protein.

$$F_{PH} = F_P + F_H$$

Because of this formula, the only possible phase outcomes for the F_P are where the two circles intersect with one another, giving two possible results. In this diagram, the circles represent the maximum amplitude for the data collected and all possible phases.

If a second heavy atom derivative dataset was collected, the structure factor for the second heavy atom could be determined ($F_{H'}$) and therefore a third circle could be drawn, leading on to only one possible phase outcome for the protein structure factor as seen in figure 2.5. In practice, the result is seldom as clear cut, but the principle works in theory

Anomalous dispersion is another technique that can be used to determine the phases, and works on similar principles to isomorphous replacement. One way it improves on SIR and even more so MIR, is that the levels of isomorphism are preserved due to the fact that only one crystal is used. In MAD, the data are collected at several wavelengths, which are shown in figure 2.6 and of which, λ_1 and λ_2 are most important.

With a change in the wavelength there is an alteration in the structure factor amplitude, which can be plotted either on a phase diagram or again on a Harker construction. In SAD only one wavelength is used.

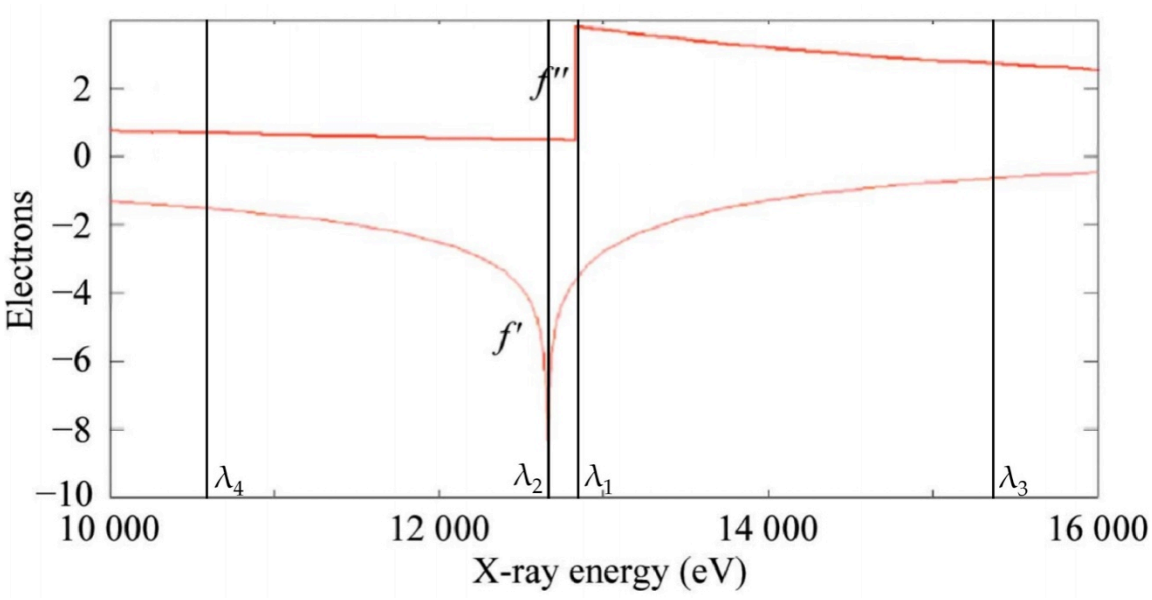


Figure A.2. - A modified graph that illustrates the variation in anomalous scattering against the incident X-ray energy. Showing f' – dispersive term, f'' – absorption term, λ_1 – absorption peak, λ_2 – maximum dispersive difference, $\lambda_3 + \lambda_4$ – remote wavelengths (162, 163)

In order to perform anomalous dispersion, atoms with anomalous scattering are needed. One way is to use a heavy atom that has been incorporated into the structure, such as by replacing the methionine amino acids present in the structure with selenomethionines, replacing the S atoms with Se atoms, which have a significantly larger atomic mass and therefore scatter electrons considerably more.

Wavelength tuning is essential for AD, therefore the data has to be collected at a synchrotron. The anomalous scattering relates to the scattering of both the *hkl* values as well as the *−h-k-l* values, which one would expect to be identical. However they are in fact slightly different.

Once the data have been collected, the structure factors are known for all the atoms. The anomalous scattering component of all the atoms then needs to be determined. Once this is known the difference in phase angle between the normal and anomalous data can be determined and the phase of the data can be calculated.