UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author FADDEN

Andrew John

Structural and functional analysis of the XPF/Mus81 family of endonucleases.

by

Andrew John Fadden

A thesis submitted to the University of London for the degree of Doctor of Philosophy

2006

London Research Institute
44 Lincoln's Inn Fields
London
WC2A 3PX

University College London
Gower Street
London
WC1E 6BT

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This thesis describes structural and functional analysis of members of the XPF/Mus81 family of junction-specific endonucleases. These enzymes recognise and nick double-stranded DNA within a variety of flaps, bubbles and branched DNA substrates. Genetic and biochemical evidence indicates that these endonucleases have important functions in nucleotide excision repair, DNA interstrand crosslink repair, recombination and replication fork repair, and thereby help maintain genomic integrity.

Eukaryotic XPF orthologues are large multidomain proteins that are not suitable for structural studies at present. For this reason, an XPF orthologue from the crenarchaeote *Aeropyrum pernix* has been used to shed light on the mechanism of XPF action. The crystal structure of the core nuclease domain of *A. pernix* XPF has been solved to 2.1 Å resolution. This together with other structures detailed in the literature has provided insight into enzyme architecture and possible mechanisms of substrate binding and cleavage. Attempts to co-crystallise *A. pernix* XPF with short oligonucleotides akin to repair substrates are detailed. Other evidence in the literature has implicated the processivity factor PCNA in stimulating crenarchaeal XPF activity. With this in mind, recombinant heterotrimeric *A. pernix* PCNA has been expressed and been complexed with various *A. pernix* XPF constructs. The repair helicase XPB has also been added to form a ternary repair complex. Structural studies of these complexes by crystallography and electron microscopy have been initiated and some crystals have been produced.

Bioinformatic analysis of eukaryotic Mus81 orthologues has identified a region of conserved sequence of previously unknown structure or function. This putative domain appears to be conserved in most eukaryotic Mus81 orthologues. The human domain was overexpressed for biochemical and structural analyses. Deletion of this domain from within Mus81 appears to affect endonuclease activity against splayed arm substrates in vitro, suggesting it has a role in protein-DNA interaction. Crystals of the human domain have been produced, although diffraction data sufficient for structure determination has not been obtained.

Acknowledgments

Firstly I would like to thank my supervisor Neil McDonald for giving me the opportunity to work in his group in such a stimulating environment. I am also very grateful for his support and advice over the last few years, as well as his patience during the writing of this thesis. Additionally, I would like to say a big thank-you to members of the Structural Biology Laboratory, both past and present for general assistance, encouragement, ideas and making the lab a great place to work. I would especially like to thank Phil for his patience in answering 50 questions per day, as well as my teammates on Project XPF, John, Maureen and Judith for all their help. Thanks also go Svend for teaching me not to buy three Hoegaarden at once, and for teaching us all how to get home from Bethnal Green.

I am grateful to all my friends at Cancer Research UK football club and those at Dinamo Dorigo football club for helping me let off some excess energy and occasionally steam.

Although they are unlikely to ever know, I would like to thank teachers past and present, especially Kevin Vickers, Ken Jones, Russell Wallis, Dan Mitchell, Maureen Taylor and Kurt Drickamer for their guidance and helping me believe in myself.

The love, support and encouragement of my family are probably the biggest reasons that I am even in a position to be examined for a PhD. They have been so important throughout my life. Thanks to Mum, Dad and Gemma for being there when I have needed you.

Finally, I would like to thank Faye for your love and understanding, especially during recent weeks. Next time I get married, I will wait until after the submission date! This thesis is dedicated to you Faye.

List of publications

Lally, J., Newman, M., Murray-Rust, J., Fadden, A., Kawarabayasi, Y., and McDonald, N. (2004). Crystallization of the xeroderma pigmentosum group F endonuclease from Aeropyrum pernix. Acta Crystallogr D Biol Crystallogr **60**, 1658-1661.

Newman, M., Murray-Rust, J., Lally, J., Rudolf, J., Fadden, A., Knowles, P. P., White, M. F., and McDonald, N. Q. (2005). Structure of an XPF endonuclease with and without DNA suggests a model for substrate recognition. EMBO J. **24**, 895-905.

Boëda, B., Briggs, D.C., Garvalov, B.K., Fadden, A.J., McDonald, N.Q., and Way, M. (2006) Tes inhibits Mena-dependent cell migration via a non-canonical EVH1 binding domain. Science - under revision

Fadden, A.J, Biggerstaff, M., Harris, R., and McDonald, N.Q. (2006). Structural and functional analysis of a novel domain of human Mus81. In preparation.

Fadden, A.J., Murray-Rust, J., Lally, J., Newman, M., McDonald, N.Q. (2006). The structure of the Aeropyrum pernix XPF nuclease domain. In preparation.

Declaration

The following figures were kindly provided by other members of the Structural Biology Laboratory, Cancer Research UK

Figure 2.2 – Bacterial overexpression of the XPF-ERCC1 complex.

Figure 2.9 – XPF-ERCC1 activity.

Figure 3.3 – Crystals and diffraction for ApeXPF nuclease domain.

Figure 4.16 – ApeXPF activity.

Figure 5.20 – Effect of hACE deletion on hMus81 activity.

Figure 6.1 - Crystals of an ApePCNA-ApeXPF complex.

Table of contents

List of figures

List of tables

Abbreviations

Ω	Ohm
Δ	Deletion
° C	Temperature in degrees Celsius
2-ME	2-mercaptoethanol
6-4 PP	pyrimidine (6-4) pyrimidone photoproduct
Å	Ångström
ACE	Mus81 Absent in *C.elegans* domain
Amp^R	Ampicillin
AP site	apurinic or apyrimidinic site
Ape	*Aeropyrum pernix*
ApeXPF	*Aeropyrum pernix* XPF homologue
APS	ammonium persulphate
AU	Asymmetric unit
BER	Base excision repair
$C_{12}E_9$	nonaethylene glycol monododecyl ether
Cam^R	Chloramphenicol
CCD	Charge-coupled device
CCP4	Collaborative Computational Project Number 4
CD	Circular dichroism
cDNA	Complementary DNA
CHAPS	3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate
CPD	cyclobutane pyrimidine dimers
CS	Cockayne's syndrome
CTAB	hexadecyltrimethylammonium bromide
Da	Daltons
dHJ	Double Holliday junction
D-loop	Displacement loop
DNA	Deoxyribonucleic acid
dNTP	2'-deoxynucleotide triphosphate
DSB	Double strand break
DTT	Dithiothreitol

E.coli	*Escherichia coli*
Eme1	Essential meiotic endonuclease 1
ERCC1	Excision repair cross-complementing gene 1
ESRF	European Synchrotron Radiation Facility
FPLC	Fast protein liquid chromatography
g	Gravitational force
GGR	Global genome repair
GSH	Glutathione
GST	Glutathione-S-transferase
hACE	human ACE domain
Hef	*Pyrococcus furiosus* Hef endonuclease; XPF homologue
HEPES	N-[2-hydroxyethyl]piperazine-N'-[2-ethanesulphonic acid]
HhH	Helix-hairpin-Helix domain
$(HhH)_2$	Tandem Helix-hairpin-Helix domain
HJ	Holliday junction
HPLC	High performance liquid chromatography
HR	Homologous recombination
ICL	Interstrand crosslink
IPTG	Isopropyl-β-D-thiogalactoside
K	Temperature in Kelvins
Kan^R	Kanamycin
Kb	Kilobase
KDa	Kilodaltons
kV	Kilovolt
LB	Luria-Bertani media
M	Molar/moles per litre
mAU	Milliabsorbance unit
MES	2-morpholinoethanesulphonic acid
mg	Milligram
μF	Microfaraday
μl	Microlitre
μM	Micromolar
mins	Minutes

MIR	Multiple isomorphous replacement
ml	Millilitre
mM	Millimolar
Mms4	Methyl methanesulfonate-sensitive 4
MPD	2-Methyl-2, 4-pentanediol
mRNA	Messenger RNA
Mus81	Methyl methanesulfonate, UV sensitive clone 81
NEB	New England Biolabs
NER	Nucleotide excision repair
ng	Nanogram
NHEJ	Non homologous end joining
nl	Nanolitre
nm	Nanometer
NMR	Nuclear Magnetic Resonance
NTA	Nitrilotriacetic acid
OD	Optical density
ori	Origin of replication
PCNA	Proliferating cell nuclear antigen
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
PEG MME	Polyethylene glycol monomethylether
pI	Isoelectric point
PIM	PCNA Interaction Motif
pmol	Picomole
PMSF	Phenylmethylsulphonylfluoride
RBS	Ribosome binding site
RMSD	Root mean square deviation
RNA	Ribonucleic acid
scACE	*Saccharomyces cerevisiae* ACE domain
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate – polyacrylamide gel electrophoresis
secs	Seconds

SeMet	Selenomethionine
SOC	Super Optimal Catabolite Repression Broth
spACE	*Schizosaccharomyces pombe* ACE domain
Sso	*Sulfolobus solfataricus*
SsoXPF	*Sulfolobus solfataricus* XPF homologue
TCR	Transcription coupled repair
TEMED	tetramethylethylenediamine
Tris	Tris[hydroxymethyl]aminomethane
tRNA	transfer RNA
TTD	Trichothiodystrophy
UV	Ultra violet
V	Volt
XP	Xeroderma pigmentosum
XPF	Xeroderma pigmentosum group F gene product

1 Introduction

1.1 DNA damage

Since deoxyribonucleic acid (DNA) was identified in 1944 as Avery's "transforming principle" (Avery et al., 1944), it has occupied a central position in biochemistry and genetics. As the source of genetic information in virtually all living organisms and the blueprint for cellular operations, maintaining the integrity of DNA is essential. This is underlined by the fact that DNA is the only biomolecule that is specifically repaired, all the others being replaced. If the repair mechanisms fail then this can lead to genomic instability, which can contribute to the development of cancer and other human diseases.

Some types of DNA damage are listed below. Some damage occurs spontaneously (a & b), whilst other damage is caused by radiation or chemical stresses (c, d, e, f; Friedberg et al., 2006)

a) Base loss

A consequence of the presence of 2-deoxyribose in DNA (as opposed to ribose) is that the N-glycosyl bond between sugar and base becomes more labile. Loss of a purine or pyrimidine base produces either an apurinic or apyrimidinic (AP) site (figure 1.1a).

b) Deamination

Apart from thymine, the bases found in DNA contain exocyclic amino groups. The loss of these primary amino groups (deamination) could lead to mutation, since these amino groups are important in the specific base-pairing described by Watson and Crick. The conversions of cytosine to uracil (figure 1.1b), adenine to hypoxanthine, guanine to xanthine and 5-methylcytosine to thymine have all been observed.

c) Oxidative damage

The bases are susceptible to numerous different modifications by a wide variety of chemical agents. Reactive oxygen species generated during oxidative metabolism and by ionizing radiation can modify bases. The most common is the modification of

thymine to thymine glycol (figure 1.1c). However oxidative degradation of the sugar backbone can lead to single strand breaks.

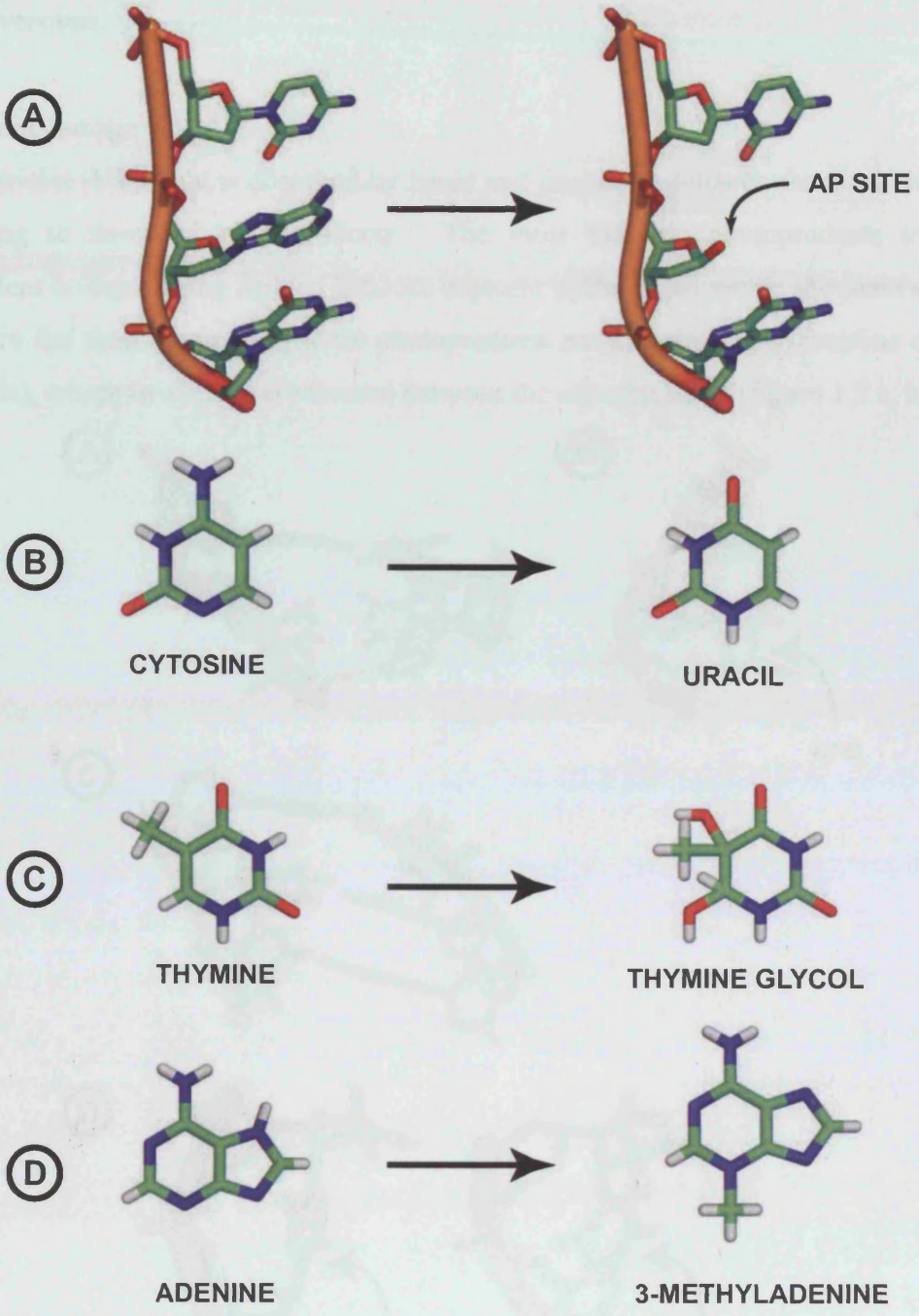

Figure 1.1 – Types of DNA damage. [A] Base loss. [B] Deamination. [C] Chemical modification: oxidation. [D] Chemical modification: methylation. Carbon atoms are in green; nitrogen, blue; oxygen, red and hydrogen, grey. All figures of this type in this thesis were produced using the program PyMol (DeLano, 2002)

In addition other metabolites may accidentally react with bases. S-adenosylmethionine, the normal biological methyl group donor reacts with DNA to

produce alkylated bases like 3-methyladenine (figure 1.1d). These bases must be repaired in order to prevent mutation. Mutations may lead transitions and transversions.

d) Photodamage

Ultraviolet (UV) light is absorbed by bases and frequently induces chemical changes leading to so-called photoproducts. The most frequent photoproducts involve covalent linkages being formed between adjacent pyrimidines within the same strand. In turn the most common of these photoproducts are cyclobutane pyrimidine dimers (CPDs), where two bonds are formed between the adjacent bases (figure 1.2 a, b).

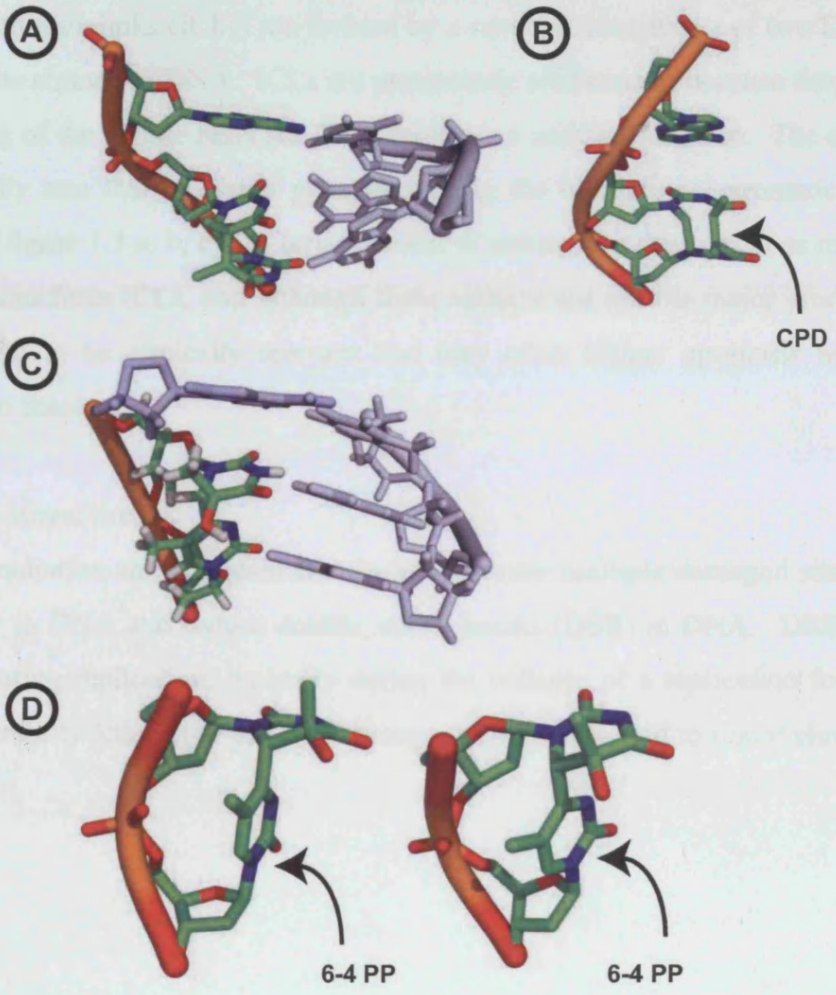

Figure 1.2 – Types of DNA damage caused by UV-light. [A] A CPD in the context of duplex DNA (PDB accession code 1PIB (Lee et al., 2004). [B] Close-up of the CPD. [C] A (6-4) PP in the context of duplex DNA (PDB accession code 1CFL (Lee et al., 1999)). [D] Close-ups of the 6-4 PP

Dimers can also be produced by formation of a single covalent bond between the C6 of one pyrimidine and the C4 of the adjacent downstream pyrimidine. This type of photoproduct is called a pyrimidine (6-4) pyrimidone photoproduct (6-4 PP) (figure 1.2 c, d).

As figure 1.2 shows both type of photoproduct create extensive distortion within the DNA structure. This distortion is an important consequence of forming photoproducts, as this is often the primary recognition signal for the repair apparatus.

e) Inter-strand crosslinks

Inter-strand crosslinks (ICLs) are formed by a covalent connection of two DNA bases on opposite strands of DNA. ICLs are particularly problematic because they block all unwinding of the double helix for DNA replication and transcription. The model ICL is generally one that has been generated using the bifunctional aromatic molecule psoralen (figure 1.3 a, b, c). A large number of antitumour drugs such as mitomycins and cisplatin form ICLs, and although these adducts are not the major products, they are thought to be clinically relevant and they often trigger apoptosis as they are difficult to resolve.

f) Double strand breaks

Ionizing radiation and radiomimetic drugs can cause multiple damaged sites in close proximity in DNA and induce double strand breaks (DSB) in DNA. DSBs are also formed during replication, typically during the collapse of a replication fork. DSBs are a potentially lethal type of DNA damage, as they may lead to major chromosomal aberrations.

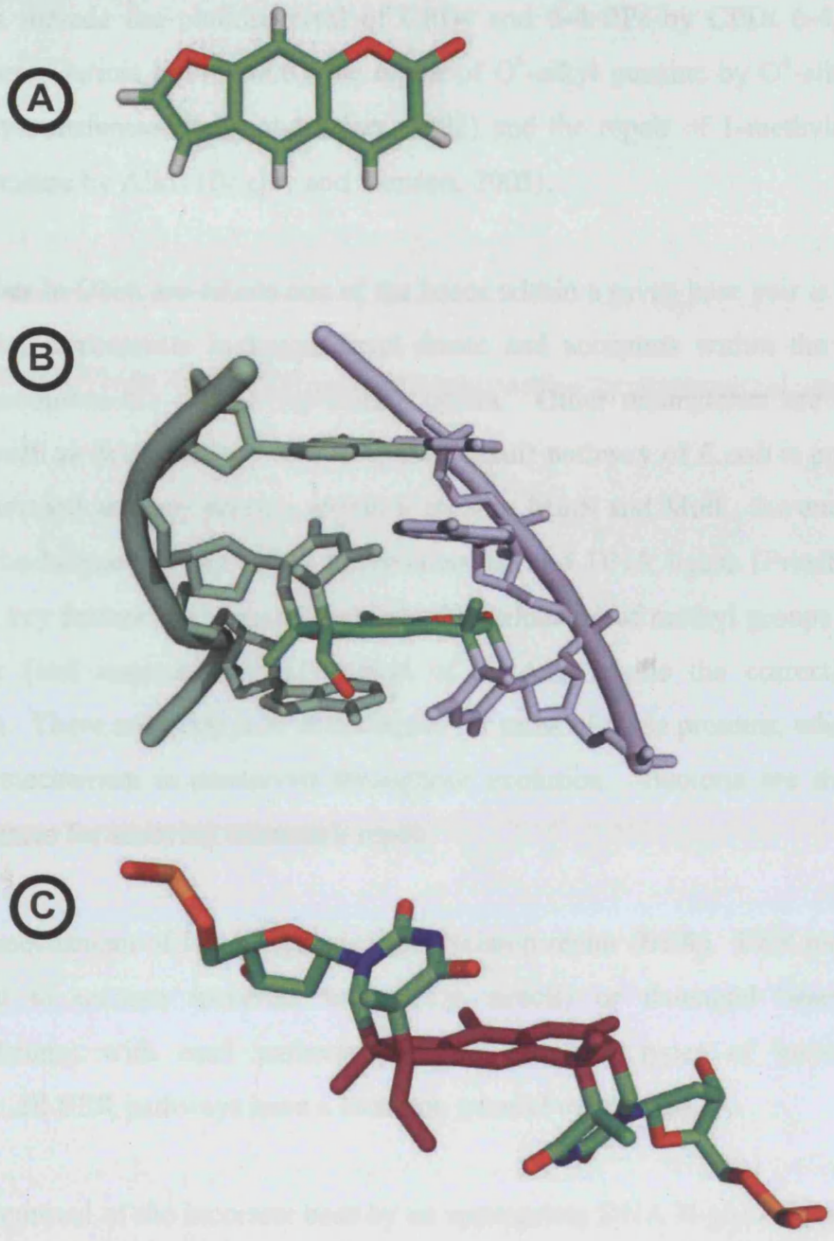

Figure 1.3 – Interstrand crosslinking by psoralen. [A] The structure of psoralen. [B] An psoralen-mediated interstrand crosslink in the context of duplex DNA (PDB accession code 1FHY (Eichman et al., 2001)). [C] Close-up of the interstrand crosslink. Psoralen is shown in maroon.

1.2 Repair of DNA damage

Although a direct reversal of DNA damage is attractive for many reasons, not least simplicity, in most cases the reverse reaction is not possible for kinetic and/or thermodynamic reasons. However, in a few cases, the reaction is reversible, and in some of these cases mechanisms have been developed to take advantage of this.

Examples include the photoreversal of CPDs and 6-4 PPs by CPD/ 6-4 PP DNA photolyases (Garinis et al., 2006), the repair of O^6-alkyl guanine by O^6-alkylguanine DNA alkyltransferase (Pegg and Byers, 1992) and the repair of 1-methyladenine/3-methylcytosine by AlkB (Begley and Samson, 2003).

Mismatches in DNA are where one of the bases within a given base pair is unsuitable for forming appropriate hydrogen bond donor and acceptors within the base pair. Most mismatches are due to replication errors. Other mismatches are caused by damage such as deamination. The mismatch repair pathway of *E.coli* is probably the best understood and key players within it include MutS and MutL, the endonuclease Mut H, the helicase UvrD, DNA polymerase III and DNA ligase (Friedberg et al., 2006). A key feature of mismatch repair is the utilisation of methyl groups present on the older (and assumed correct) strand of DNA to guide the correction of the mismatch. There are eukaryotic orthologues for most of these proteins, which implies that the mechanism is conserved throughout evolution. Bacteria are thus a good model system for studying mismatch repair.

Another mechanism of DNA repair is base excision repair (BER). BER pathways are employed to remove incorrect bases (e.g. uracil) or damaged bases (e.g. 3-methyladenine) with each pathway specific for each types of incorrect base. However, all BER pathways have a common general mechanism.

1) Removal of the incorrect base by an appropriate DNA N-glycosylase to create an apurinic/apyrimidinic (AP) site.
2) Nicking of the damaged DNA strand by AP endonuclease upstream of the AP site, thus creating a 3'-OH terminus adjacent to the AP site
3) Extension of the 3'-OH terminus by a DNA polymerase in concert with excision of the AP site.

Structures of several of the DNA N-glycosylases involved in step 1 have been solved. It is the specificity of the DNA N-glycosylase for the inappropriate nucleobase that determines whether and how BER proceeds (Friedberg et al., 2006).

1.3 Nucleotide excision repair (NER)

The repair mechanisms above can deal with DNA damage so long as there is an appropriate enzyme capable of recognising that specific damage. However, the wide range of DNA-reactive chemicals in our environment leads to a huge range of lesions that cannot be covered by specific DNA glycosylases. Evolution cannot be expected to keep developing focussed pathways for each new environmental mutagen and chemotherapy drug. To compensate for this, a more flexible, but more complex damage repair mechanism has evolved. This is called nucleotide excision repair (NER).

There are 6 major steps in eukaryotic NER involving 15-18 different polypeptides (figure 1.4; unless otherwise noted human protein names are given; (Araujo and Wood, 1999; Park and Choi, 2006)

1) Damage recognition. In NER this occurs by recognising an abnormal structure in the DNA such as that caused by CPDs and 6-4 PPs. Damage recognition is thought to involve the XPC-HR23B complex, as well as XPA and the single-stranded DNA binding protein RPA.
2) Binding of a multi-protein complex at the damage site and unwinding of the double helix around the damage. The 3' → 5' helicase XPB and the 5' → 3' helicase XPD are part of the TFIIH complex involved in this step.
3) Double incision of the damaged strand several nucleotides away from the damage site on both the 5' and 3' sides, followed by excision of this 18-25 nucleotide stretch of the damaged strand. The incisions are performed by the endonucleases XPG, which is a member of the FEN-1 family of endonucleases and ERCC1-XPF, a heterodimeric enzyme which will be discussed in more depth in section 1.5 and onwards.
4) De novo DNA synthesis filling in the gap by a repair DNA polymerase, followed by ligation.

The overall process of NER in bacteria resembles that in eukaryotes but the proteins involved are generally not homologous. The UvrABC complex performs the incision events in bacterial NER. Thus bacteria are not good model organisms for the study of

eukaryotic NER. However, yeast and Archaea provide appropriate model organisms for the study of NER.

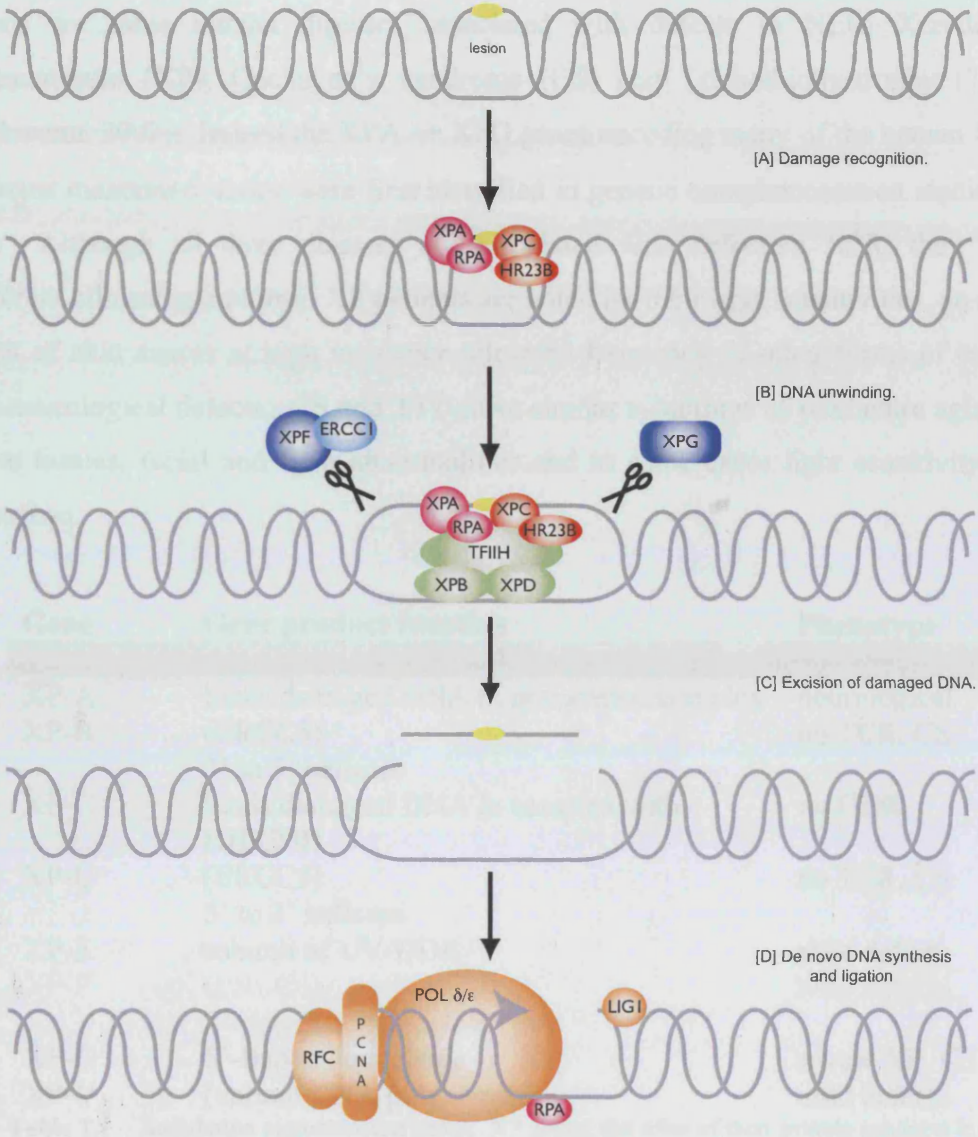

Figure 1.4 – A general scheme for eukaryotic nucleotide excision repair

How NER proceeds is dependent on how the damage is recognised. It has been observed that damage within heavily transcribed regions of the genome is usually repaired more rapidly than damage in the non-transcribed regions. Thus it is generally accepted that there are two types of NER, global genome repair (GGR) and transcription-coupled repair (TCR). It is thought that GGR is initiated by the XPC-HR23B complex "finding" damage, where as TCR is initiated by the stalling of RNA polymerase at sites of damage (van Hoffen et al., 2003). Apart from these differences

in damage recognition and initiation, the general mechanism of NER is common to both processes.

There are three human diseases associated with defects in NER; Xeroderma pigmentosum (XP), Cockayne's syndrome (CS) and Trichothiodystrophy (TTD) (Lehmann, 2003). Indeed the XPA → XPG genes encoding many of the human NER proteins mentioned above were first identified in genetic complementation studies of XP. Although all three diseases are associated with defective NER, they have different clinical symptoms. XP patients are noted for their light sensitivities, an early onset of skin cancer at high incidence, elevated frequency of other forms of cancer and neurological defects. CS and TTD have similar symptoms of premature aging of some tissues, facial and limb abnormalities and in some cases light sensitivity and dwarfism.

Gene	Gene product function	Phenotype
XP-A	binds damaged DNA in preincision complex	neurological
XP-B	(ERCC3) 3' to 5' helicase	no TCR, CS
XP-C	binds damaged DNA in complex with HHR23B	no GGR
XP-D	(ERCC5) 5' to 3' helicase	no TCR, CS
XP-E	subunit of UV-DDB	mild defects
XP-F	(ERCC4) 3'-flap endonuclease	mild defects
XP-G	5'-flap endonuclease	severe XP, CS
XP-V	(variant DNA polymerase)	mild defects

Table 1.1 – Xeroderma pigmentosum genes. XP genes, the roles of their protein products in NER and in disease.

1.4 NER in Archaea – model organisms for eukaryotic NER.

The basic repair processes discussed thus far are common to bacteria and eukaryotes, but the proteins involved are not always conserved between these two domains of life and this is particularly true for NER and double strand break repair. However, in Archaea, the third domain of life, there are marked similarities between eukaryotic and archaeal DNA replication and transcription proteins, as well DNA compaction proteins such as histones (White and Bell, 2002). Thus it is not surprising that many

archaeal organisms contain homologues of eukaryotic NER components rather than homologues of bacterial proteins such as UvrABC (White, 2003). (A small sub-group of Archaea do have UvrABC orthologues although this is thought to have occurred by a lateral gene transfer from bacteria). Table 1.2 summarises the distribution of NER proteins in sequenced archaeal genomes (adapted from White, 2003 and Kelman and White, 2005). It is therefore evident that that Archaeal DNA repair provides a model system to study eukaryotic DNA repair.

Species	Short XPF	Long XPF	XPB	XPD	PCNA	RPA	SSB	UvrABC	Optimal temperature (°C)
Aeropyrum pernix	**1**	**0**	**1**	**1**	**3**	**0**	**1**	**0**	**95**
Pyrobaculum aerophilum	**1**	**0**	**2**	**1**	**2**	**0**	**0**	**0**	**95**
Sulfolobus solfataricus	**1**	**0**	**2**	**1**	**3**	**0**	**1**	**0**	**80**
Sulfolobus tokodaii	**1**	**0**	**2**	**2**	**3**	**0**	**1**	**0**	**80**
Thermoplasma acidophilum	0	0	1	1	1	1	1	0	60
Thermoplasma volcanium	0	0	1	1	1	1	1	0	60
Methanopyrus kandleri	0	1	0	0	1	1	0	0	100
Pyrococcus furiosus	0	1	2	1	1	1	0	0	100
Pyrococcus horikoshii	0	1	2	1	1	1	0	0	100
Pyrococcus abyssi	0	1	2	1	1	1	0	0	100
Methanococcus jannaschii	0	1	0	0	1	1	0	0	85
Archeoglobus fulgidus	1	1	1	0	1	1	0	0	85
Methanothermobacter thermautotrophicus	0	1	1	0	1	1	0	1	70
Methanosarcina acetivorans	0	1	1	1	1	1	0	1	< 40
Methanosarcina mazei	0	1	1	1	1	1	0	1	< 40
Halobacterium	0	1	2	1	1	1	0	1	< 40

Table 1.2 – Distribution of selected DNA repair proteins in sequenced archaeal genomes. Crenarchaea are shown in bold, euryarchaea in normal text. Adapted from Kelman and White, 2005; White, 2003.

1.5 The XPF/Mus81 family of endonucleases

XPF-ERCC1 and Mus81-Eme1 are heterodimeric endonucleases (Boddy et al., 2001; Sijbers et al., 1996). Each endonuclease has at least a single metal-dependent catalytic site located within the XPF or Mus81 subunit, whilst their respective partners ERCC1 and Eme1 lack a catalytic function. XPF-ERCC1 and Mus81-Eme1 are structurally related but functionally distinct and have been found in eukaryotes from yeast to humans. There is evidence for a third related endonuclease in higher mammals (Meetei et al., 2005). XPF/Mus81 homologues have also been identified in

Archaea (White, 2003). Collectively these endonucleases and their non-catalytic endonuclease-like partners form a family of evolutionarily related proteins.

Members of this family have not been named systematically but have had names assigned from radiation-sensitive mutational screens or by their ability to complement certain repair-deficient cell lines. The first XPF-like endonuclease described was the *S.cerevisiae* orthologue of XPF-ERCC1, the Rad1-Rad10 complex (Tomkinson et al., 1993). Both Rad1 and Rad10 were identified by screening genomic libraries for the complementation of UV-radiation sensitive mutants. A similar screen in *S.pombe* identified another XPF orthologue Rad16 (Carr et al., 1994); whilst the ERCC1 orthologue, Swi10, was found by characterising mutants with a reduced frequency of mating-type switching (Schmidt et al., 1989).

ERCC1, the mammalian counterpart to Swi10 and Rad10, was found by its ability to complement a nucleotide excision repair (NER) repair defect in a mutant Chinese hamster ovary (CHO) cell line and was named excision repair cross-complementing gene 1 (or ERCC1) (Westerveld et al., 1984). XPF is the human orthologue of Rad1 and Rad16 and was found by degenerate PCR based on sequence similarity with the yeast orthologues (Sijbers et al., 1996). It was also independently found to complement a mutant CHO cell line and named ERCC4. The XPF gene product complements the NER defect in patients with Xeroderma pigmentosum complemenation group F (Yagi et al., 1998).

The XPF paralogue, Mus81 (methyl methansulfonate, UV sensitive clone 81), was identified independently by several groups using two-hybrid screens with Rad54 from *S.cerevisiae* and with Cds1 from *S.pombe* (Interthal and Heyer, 2000). Mus81 was also found in a synthetic lethal screen for genes that are essential in the absence of the RecQ helicase Sgs1 (Mullen et al., 2001). The non-catalytic partner of Mus81, analogous to the ERCC1 subunit of XPF-ERCC1, is called Mms4 (methyl methanesulfonate-sensitive 4) in *S.cerevisiae* (Xiao et al., 1998), and Eme1 (essential meiotic endonuclease 1) in *S.pombe*, where it was identified as a Mus81-associated protein (we refer to the *S.pombe* convention throughout) (Boddy et al., 2001).

A third human XPF/Mus81 endonuclease has recently been characterised as the product of the FANCM gene involved in the disease Fanconi anemia. Fanconi anemia patients are characterised by their genomic instability and predisposition to cancer. It has been proposed that the FANCM protein may have an active helicase domain enabling it to act as a translocase for other FANC proteins. A second ERCC1-like gene, MGC32020, is also found in the human genome and together with FANCM gene products could reconstitute an active nuclease, although this remains to be shown (Meetei et al., 2005). In a similar manner, analysis of archaeal genomes revealed XPF homologues in both the major divisions of Archaea, the Euryarchaea and the Crenarchaea (White, 2003). Many are uncharacterized open reading frames with the exception of the XPF homologue from *Pyrococcus furiosus* designated Hef (Helicase-associated endonuclease for fork-structured DNA) (Komori et al., 2002), and the smaller XPFs from *Aeropyrum pernix* (Lally et al., 2004) and *Sulfolobus solfataricus* (Roberts et al., 2003). Hef has an ATP-dependent helicase activity in addition to its endonuclease activity and was identified from a screen for factors that enhanced Holliday junction resolvase activity (Komori et al., 2002).

1.6 Domain organisation of XPF/Mus81 family members

There are five subgroups within the XPF/Mus81 family; four of which provide good evidence for a common evolutionary origin. The characteristic feature of most family members is the presence of an endonuclease or endonuclease-like domain and two helix-hairpin-helix (HhH) motifs (figure 1.5). There are three branches with catalytic endonuclease subunits and two that are non-catalytic but associate with their cognate active subunit. The ERCC1 and Eme1 proteins probably originated from a gene duplication of an active XPF subunit that subsequently diverged to produce a single catalytic centre heterodimeric endonuclease from a two catalytic centre homodimeric endonuclease. Evidence from the homodimeric *A.pernix* XPF (Newman et al., 2005) indicates that only a single catalytic centre is active at one time and could explain why eukaryotic XPF-ERCC1 or Mus81-Eme1 complexes retain only one catalytic motif within the endonuclease.

The long XPF nucleases range in size from 892 to 2048 amino acids and have an SF2 superfamily helicase domain or helicase-like domain fused to the amino-terminus of the core nuclease and tandem helix-hairpin-helix ($(HhH)_2$) domains. XPF subunits

with known or predicted helicase activity are found in Euryarchaea, for example the Hef endonuclease of *P.furiosus* (Komori et al., 2004) and the FANCM gene product in the human and mouse genomes (Meetei et al., 2005). However the conserved eukaryotic XPF subunit has a degenerate helicase domain that has accumulated mutations and not retained helicase activity.

The small XPFs have the simplest structure comprising just the conserved core nuclease and $(HhH)_2$ domains and are restricted to crenarchaea and a gene encoded within the African swine fever virus (ASFV) genome (Yanez et al., 1995). The small XPFs in crenarchaea have evolved a further elaboration through a carboxy-terminal motif that binds to a heterotrimeric PCNA processivity factor (Roberts et al., 2003). Archaeal genomes typically encode only one XPF/Mus81 gene except *A.fulgidus* that appears to have a short and a long XPF gene (White, 2003).

The Mus81 sub-group has a variant domain organization where the carboxy-terminal $(HhH)_2$ domain is replaced by single HhH motifs at each sequence extremity. The different arrangement of HhH motifs between XPF and Mus81 subunits is reminiscent of circularly permuted protein folds and it is possible that folding of Mus81 may permit contacts that reconstitute the $(HhH)_2$ domain. Mus81 proteins have two other characteristics; an insert within their nuclease domain and a small functional domain between the amino-terminal HhH domain and nuclease domain (to be discussed in Chapter 5).

The last two subgroups in the family are the ERCC1 and Eme1 proteins unique to eukaryotes. Both ERCC1 and Eme1 are structurally related to their cognate active XPF or Mus81 subunits. Both probably evolved by gene duplication of an active nuclease subunit and retained their HhH motifs but lost the catalytic function of their nuclease-like domain. ERCC1 is extremely well conserved within known orthologues suggesting a critical function that is still not well understood. The nuclease-like domain of Eme1 is poorly conserved and known orthologues cannot be reliably aligned either with ERCC1 or other XPF/Mus81 members. However secondary structure predictions support the presence of a nuclease-like domain and justify the inclusion of these proteins in the XPF/Mus81 family. Moreover the carboxy-terminal portion of Eme1 has an HhH motif that is closely related to Mus81.

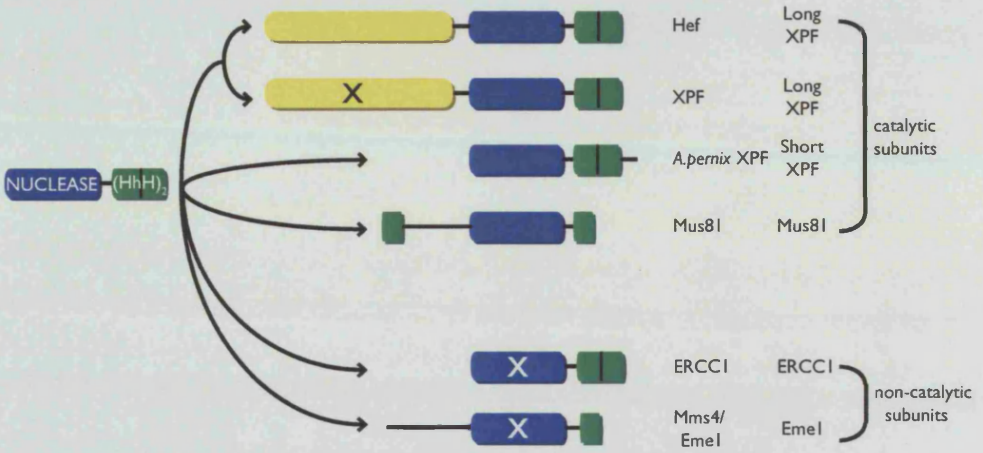

Figure 1.5 – Domain organisation of the XPF/Mus81 family of endonucleases. Helicase/helicase-like domains are in yellow, nuclease/nuclease-like domains are in blue and helix-hairpin-helix domains are in green.

1.7 Characteristic structural features

At the beginning of this thesis there were no crystal structures published from the XPF/Mus81 family. However, in recent years, several have shed light on the structures of the various domains within this family (figure 1.6).

1.7.1 Nuclease domain

The nuclease domain common to all XPF/Mus81 family members encompasses both a type II restriction endonuclease fold with a central six stranded β-sheet flanked by several α-helices and a α-helix-β-strand-α-helix dimerisation motif (Newman et al., 2005; Nishino et al., 2003)(representive structures shown in figure 1.6). The nuclease-like domain of ERCC1 has the same structural elements as the active subunits and the same interface used for dimerisation in Hef may also be used by heterodimeric XPF-ERCC1 complexes, though there is some controversy over this issue. Active subunits such as XPF and Mus81 have a conserved GDX_nERKX_3D motif within their nuclease fold indicative of a metal-dependent nuclease activity. Mutagenesis and biochemical evidence indicates that this motif and other residues within this region are required for nuclease activity but not for DNA junction binding (Enzlin and Scharer, 2002; Nishino et al., 2003). There is a good correspondence between the active site GDX_nERKX_3D motif of Hef and the $PDX_n(E/D)XK$ motif of type II restriction endonucleases (Pingoud et al., 2005). These acidic residues in the Hef active site

motif coordinate a divalent metal ion together with several main chain carbonyl groups.

The nuclease domain is required but is not sufficient for endonuclease activity. It is likely that the major DNA-binding affinity resides within the $(HhH)_2$ domain and this is required to recruit substrate to the catalytic domain. The XPF from *A.pernix* has its nuclease and $(HhH)_2$ domains connected together via a flexible linker (Newman et al., 2005). This separation of DNA-binding and catalytic functions is also found in certain modular prokaryotic restriction endonucleases such as FokI (Wah et al., 1997).

1.7.2 Helix-hairpin-helix $(HhH)_2$ domain

A distinguishing feature of the XPF/Mus81 family is their ability to recognise DNA junctions with a defined polarity and to do this they employ sequence-independent HhH DNA-binding motifs (Doherty et al., 1996)(representive structures shown in figure 1.6). The HhH motif is a ubiquitous DNA-binding motif that was first found by Thayer et al. (1995). In a more specialised variant, two HhH motifs, designated as an $(HhH)_2$ domain, form a globular domain with a fifth connecting helix (Shao and Grishin, 2000). Such domains are found in prokaryotic RuvA Holliday junction resolvases and UvrC nucleases and often have glycine residues at the first and third positions of the hairpin portion of the HhH motif (Rafferty et al., 1998; Singh et al., 2002). More reliable predictors of the motif are core hydrophobic residues typically found in the five-helix bundle and in this context the $(HhH)_2$ domain recognises the minor groove of duplex DNA (Newman et al., 2005).

As well as the nuclease-domain dimerisation motif, both Hef and *A.pernix* XPF also dimerise through their $(HhH)_2$ domains. Consequently simultaneous disruption of both interfaces is required to produce a monomeric form with a greatly diminished endonuclease activity. Association of these $(HhH)_2$ domains into distinct quaternary structures is most likely the key to understanding the XPF/Mus81 substrate specificity. For example, the $(HhH)_2$ dimer observed for *A.pernix* XPF suggests it can distort a nicked DNA duplex or 3'-flap by independently engaging both the upstream and downstream duplex regions. Recently structures of the C-terminal regions of

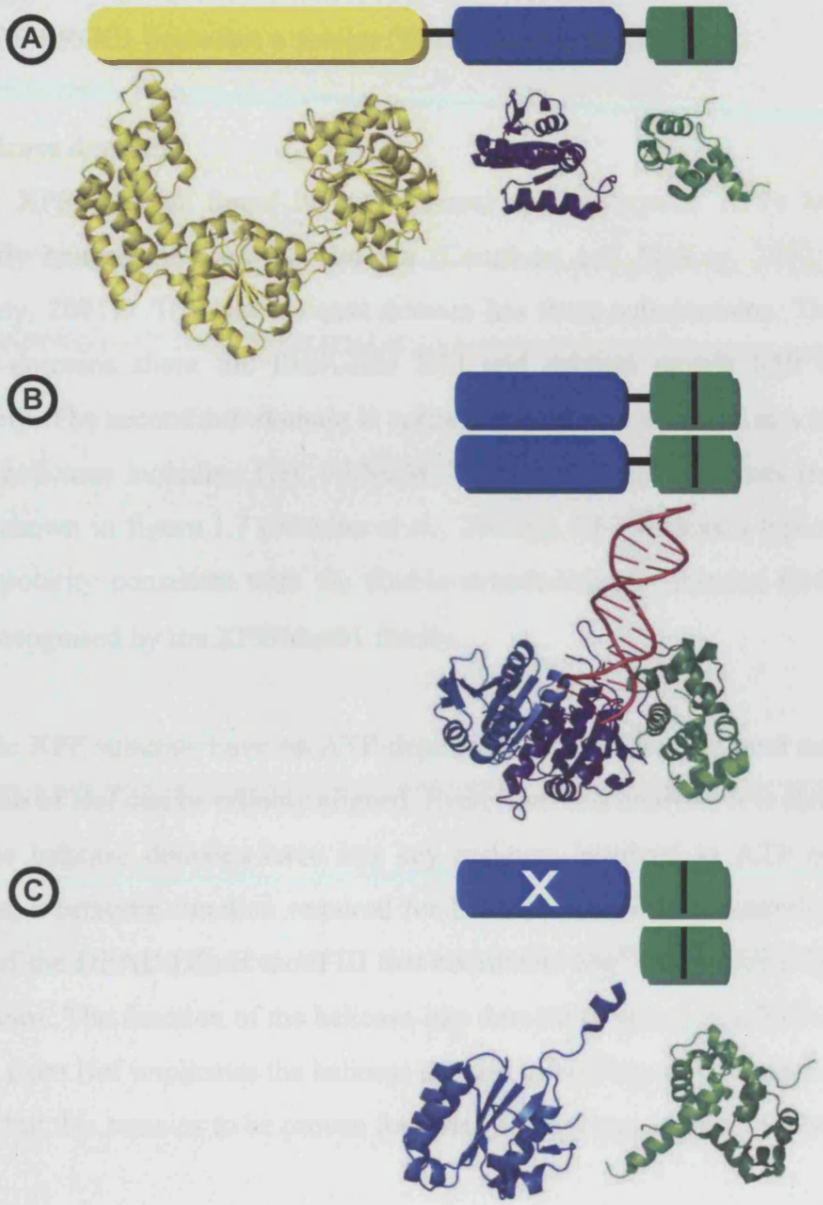

Figure 1.6 – Representative structures of domains from XPF/Mus81 family members published during this period of this thesis. [A] Crystal structures of domains from the *Pyrococcus furiosus* XPF homologue, Hef. The helicase domain (yellow; PDB accession code 1WP9 (Nishino et al., 2005a)), the nuclease (blue; PDB accession code 1J25 (Nishino et al., 2003)) and the $(HhH)_2$ domain (green; PDB accession code 1X2I (Nishino et al., 2005b)). [B] Crystal structure of the XPF homologue from *Aeropyrum pernix* with DNA-bound (PDB accession code 2BGW (Newman et al., 2005)) [C] Crystal structures of domains from human ERCC1-XPF. The helicase-like domain from human ERCC1 (blue; PDB accession code 2A1I (Tsodikov et al., 2005)) and the $(HhH)_2$ domain dimer from ERCC1-XPF (green/dark green; PDB accession code 2A1J (Tsodikov et al., 2005))

XPF and ERCC1 (Tripsianes et al., 2005; Tsodikov et al., 2005) have shown that human XPF-ERCC1 possesses a similar $(HhH)_2$ domain heterodimer.

1.7.3 Helicase domains

The long XPF proteins found in euryarchaeal and eukaryotic XPFs have a SF2 superfamily helicase/helicase-like domain (Caruthers and McKay, 2002; Soultanas and Wigley, 2001). The Hef helicase domain has three sub-domains. The first and third sub-domains share the RecA-like fold and contain motifs I-III and IV-VI respectively. The second sub-domain is entirely helical and is found in a limited sub-group of helicases including Hef, FANCM, Mph1 and Dicer helicases (representive structure shown in figure 1.7 (Nishino et al., 2005a)). SF2 helicases typically have a 3' → 5' polarity consistent with the double-stranded/single-stranded DNA junction polarity recognised by the XPF/Mus81 family.

Eukaryotic XPF subunits have no ATP-dependent helicase activity and only the first sub-domain of Hef can be reliably aligned. Even from this analysis it is clear that their degenerate helicase domains have lost key residues involved in ATP binding and hydrolysis, a catalytic function required for helicase action. For example, the acidic residues of the DEAD/DExH motif III that coordinate Mg^{2+} are lacking from eukaryl XPF proteins. The function of the helicase-like domain of eukaryotic XPFs is unclear. Evidence from Hef implicates the helicase domain in binding and recognition of DNA junctions but this remains to be proven for other XPFs (Komori et al., 2004).

1.8 Mechanism and substrate specificity

Biochemical studies on *S.cerevisiae* Rad1-Rad10 and human XPF-ERCC1 indicate that both complexes preferentially nick duplex DNA adjacent to a 3' single-stranded stretch in simple Y and stem-loop structures (Bastin-Shanower et al., 2003; de Laat et al., 1998). In contrast Mus81-Mms4 from *S.cerevisiae* nicks 3'-flaps and other model replication fork substrates, which contain 5' DNA ends not further than 3 bases downstream of the branchpoint (Bastin-Shanower et al., 2003; Fricke et al., 2005; Whitby et al., 2003)(figure 1.7). The crenarchaeal XPF from *Sulfolobus solfataricus* exhibits similar substrate preferences to Mus81-Eme1 despite their different HhH motif arrangements (Roberts and White, 2005). This has lead to the suggestion that the archaeal XPFs represent an ancestral form of XPF/Mus81 before the enzyme

diverged to select different substrates by the co-evolution of XPF-ERCC1 and Mus81-Eme1 complexes (White, 2003). These subtle differences in substrate specificity observed in vitro help ensure each endonuclease plays a distinct role in the repair and preservation of the genome.

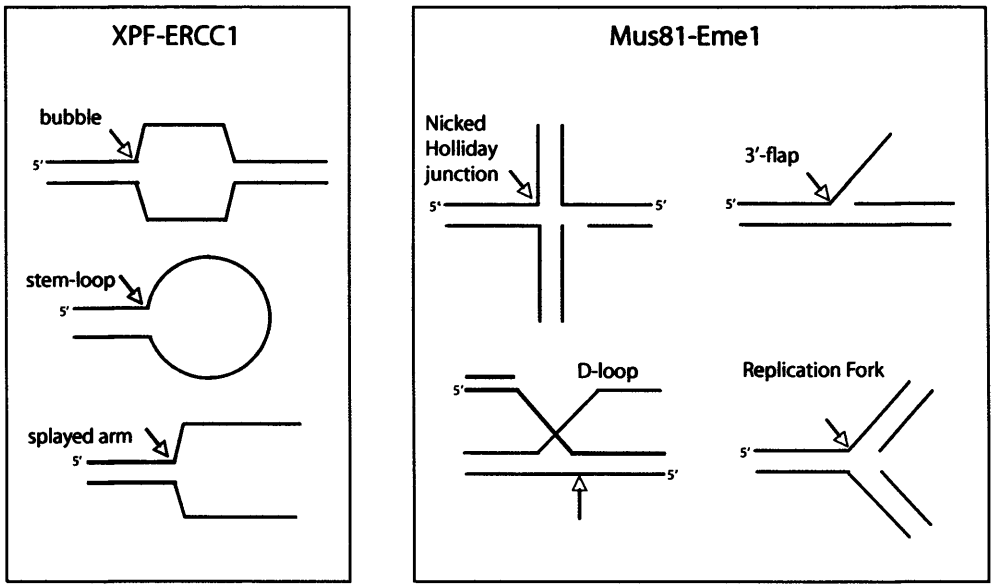

Figure 1.7 – Preferred substrates cleaved by XPF-ERCC1 and Mus81-Eme1 in vitro. Archaeal XPF homologues prefer Mus81-Eme1-type substrates rather than XPF-ERCC1-type substrates.

XPF-ERCC1 cleaves a single phosphodiester bond in one strand of the junction duplex. This activity requires Mg^{2+} (or other suitable divalent cations such as Mn^{2+}) and the active site residues coordinating this metal ion have been identified (Enzlin and Scharer, 2002). Beyond this, the catalytic mechanism of XPF/Mus81 endonucleases is not well understood though there are several general mechanistic features shared with the type II restriction enzymes. What is unique to the XPF/Mus81 family is how they recognize DNA junctions. Insights into how this is achieved have come from studies on archaeal XPFs implicating both XPF subunits in substrate recognition through a large distortion of the DNA junction (Newman et al., 2005; Nishino et al., 2005b). It remains unclear whether the duplex part of the DNA junction is bent as it enters the XPF active site as observed for EcoRV (Horton and Perona, 1998), or whether the individual strands close to the junction are locally unwound permitting only one strand to enter the XPF-ERCC1 catalytic centre.

1.9 Other cellular functions of the XPF-ERCC1 endonuclease

Although the primary cellular function of XPF-ERCC1 is in the NER-dependent repair of DNA lesions, there are other repair processes it has been implicated in. The extreme sensitivity of XPF and ERCC1 mutants to cross-linking agents as well as the severe phenotypes observed for XPF-null and ERCC1-null mice, such as growth retardation, early death and abnormalities in many organs, provides good evidence that XPF-ERCC1 and its orthologues have NER-independent functions (McWhir et al., 1993; Tian et al., 2004). XPF-ERCC1 is known to function in a recombination repair sub-pathway called single-strand annealing (Motycka et al., 2004). XPF-ERCC1 is also essential for targeted homologous recombination of exogenous DNA into mice (Niedernhofer et al., 2001). XPF-ERCC1 is also able to process blocked 3' termini that prevent priming of DNA synthesis (Guzder et al., 2004) and XPF/ERCC1 has also been shown to remove the 3' overhang from uncapped telomeres (Zhu et al., 2003).

It is well documented that XPF-ERCC1 activity is required for the repair of ICLs, independently of other NER components (De Silva et al., 2000). Evidence suggests XPF-ERCC1 is required for unhooking the ICL and resectioning the substrate for processing by homologous recombination (McHugh et al., 2001). It is not clear how XPF-ERCC1 targets ICLs or how it cooperates with recombination proteins that are also required for ICL repair.

These questions have potential medical implications since compounds such as melphalan and cisplatin introduce highly cytotoxic ICLs into DNA and are often part of chemotherapy regimens to treat cancer. Genetic and RNAi knockdown data have validated XPF-ERCC1 as an anti-cancer target and small molecular inhibitors could have important therapeutic uses. Compounds that block XPF-ERCC1 activity in the short term and thereby impede DNA repair capacity could enhance the toxicity of ICL-generating compounds. Indeed it been found that testis tumour cell lines are defective in NER and it is thought that this is a major reason why cisplatin-based chemotherapy is so effective against testicular cancer. Thus inhibition of NER by inhibition of XPF-ERCC1 could sensitize a wider range of tumours to such treatments and reduce acquired drug resistance.

1.10 Cellular function(s) of Mus81-Eme1

In keeping with its preferred substrates in vitro, the major cellular role of Mus81-Eme1 is though to be in the repair of replication forks and double-strand breaks by a mechanism distinct from that of classical bacterial homologous recombination. A model has been suggested where by Mus81-Eme1 is involved in double-strand break repair without the formation of a double HJ intermediate (figure 1.8, (Whitby, 2005)). In this model Mus81-Eme1 cleaves a D-loop on one side of the DSB and a half-junction formed by second strand capture on the other side of the break. This cleavage would leave gaps and/or 5'-flaps of ~ 5 nucleotides. Such flaps could isomerise and be cut by Mus81-Eme1 or could be dealt with by a Rad2/XPG/Fen-1 type endonuclease. DNA synthesis and ligation of the cleaved strands results in a repaired molecule, which has been crossed-over without formation of a double HJ. This model for Mus81-Eme1 action is attractive as it ties together biochemical and genetic evidence nicely (Hollingsworth and Brill, 2004).

Mus81-null mice, unlike XPF-ERCC1-null mice, are viable but are highly predisposed to lymphomas and other cancers, pointing to a role in maintaining genomic stability. Both Mus81-null and Eme1-null mice are hypersensitive to mitomycin C and cisplatin, both agents which cause ICLs. (Abraham et al., 2003; McPherson et al., 2004). These mice are however, not sensitive to other DNA damaging agents such as infrared, UV or hydroxyurea. However double strand break formation in these mice is normal, which suggests that if Mus81-Eme1 is involved in ICL repair, it is at a late stage (Dendouga et al., 2005). This is consistent with the model of action in figure 1.8.

The cellular function of archaeal XPF/Mus81 homlogues is not yet understood. In vitro data for the *Pyrococcus furiosus* XPF homologue, Hef, demonstrating it can cleave nicked, flapped and fork-structures, is consistent with a role in resolving stalled replication forks similar to Mus81-Eme1 (Komori et al., 2002).

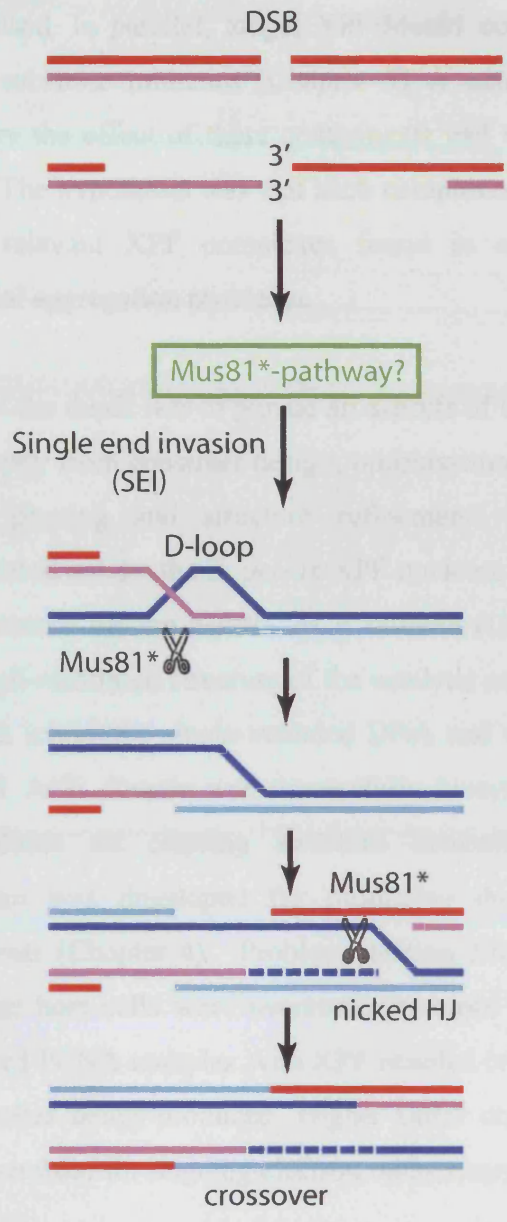

Figure 1.8 – A scheme for double-strand break repair involving Mus81.

1.11 Thesis outline

The aim of this thesis was to understand how XPF/Mus81 endonucleases recognise their substrates and the catalytic mechanism underlying the cleavage reaction.

XPF and Mus81 endonucleases are large, multi-domain proteins that pose challenging problems for X-ray crystallography analysis. To develop a tractable crystallographic project, a reductionist approach was taken to study individual domains from XPF and Mus81 and thereby remove conformational flexibility. In this way the nuclease domain of *A.pernix* XPF (Chapter 3) and the ACE domain of Mus81 (Chapter 5) were

prepared and crystallised. In parallel, longer XPF/Mus81 complexes were prepared that included DNA substrate mimetics (Chapter 3) or activating PCNA subunits (Chapter 4) to explore the effect of these components had on the crystallisation of resulting complexes. The hypothesis was that such complexes evidently better reflect the physiologically relevant XPF complexes found in cells and could reduce flexibility and potential aggregation problems.

The main objective of the thesis was to pursue all aspects of a structure determination by X-ray crystallography from construct design, optimisation and expression through to data collection, phasing and structure refinement. A successful structure determination was carried out for the *A. pernix* XPF nuclease domain (Chapter 3) and a near successful outcome for the Mus81 ACE domain (Chapter 5). The nuclease domain provides a high-resolution structure of the catalytic centre of an XPF and may allow complexes with inhibitors, single-stranded DNA and various metal ions to be prepared. The Mus81 ACE domain was successfully biosynthetically labelled with nitrogen-15 to facilitate an ongoing structure determination by NMR. An overexpression system was developed for producing the heterotrimeric PCNA complex from *A. pernix* (Chapter 4). Problems arising from toxicity of individual PCNA subunits to the host cells were overcome by more stringent transcriptional control. A reconstituted PCNA complex with XPF enabled crystal screens to be setup leading to small needles being produced. Higher order complexes with the XPB helicase were also assembled for ongoing electron microscopy experiments.

Structure-function analysis on XPF-ERCC1 required a functional assay for catalytic activity to allow selected point mutations to be prepared and characterised. Efforts to increase yields of human XPF-ERCC1 by optimisation of expression constructs and an improved purification protocol had a major impact on the human XPF-ERCC1 project (Chapter 2). Benefits included allowed establishment of activity assay, permitting screening of a large chemical library for XPF-ERCC1 inhibitors and allowing selected point mutations of XPF-ERCC1 to be prepared and assayed. Functional assays were also established for *A.pernix* XPF and for Mus81-Eme1 complexes.

2 Preparation and characterisation of the human XPF-ERCC1 endonuclease

The role of XPF-ERCC1 as a structure-dependent endonuclease belonging to the XPF/Mus81 family of endonuclease is outlined in chapter 1. This chapter describes experiments relating to structural studies of the human XPF-ERCC1 complex, as well as functional assays and efforts to enhance yields of XPF-ERCC1 A flowchart describing these experiments can be found below (Figure 2.1).

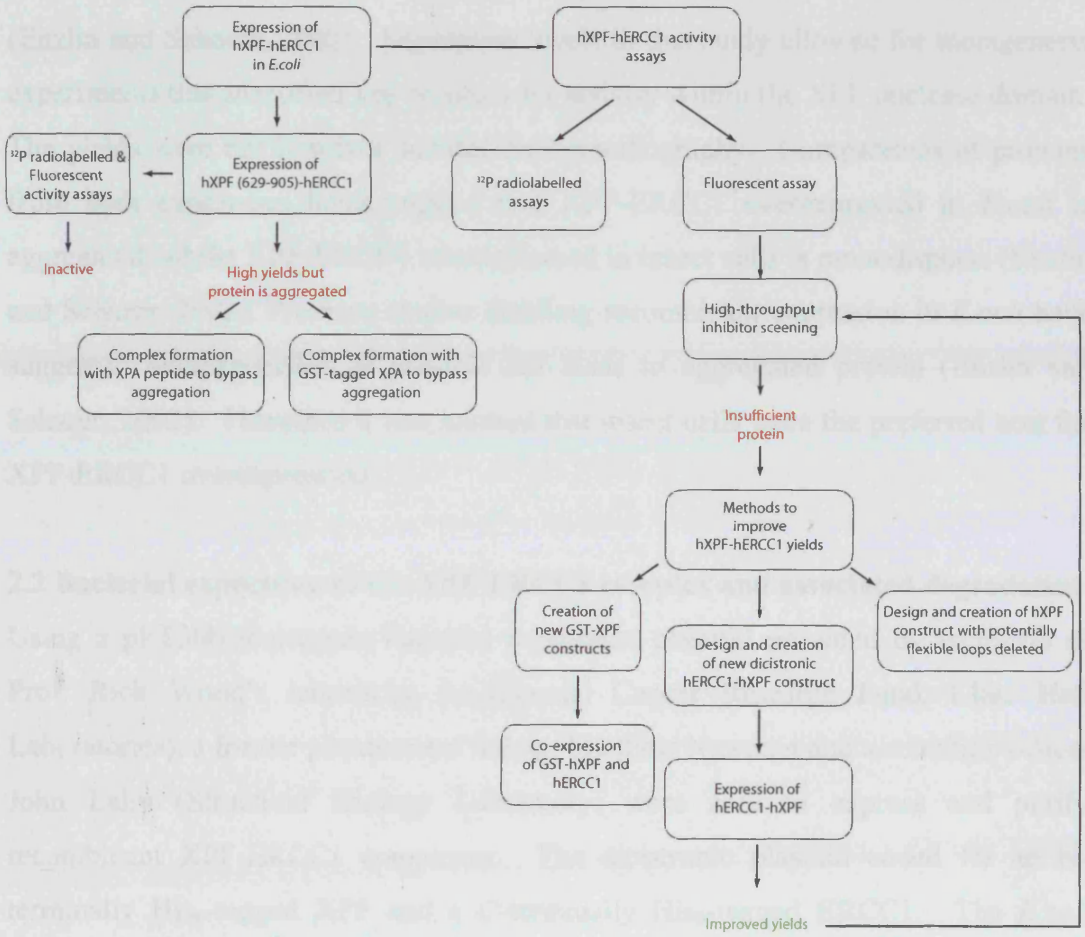

Figure 2.1 – Flowchart of the experiments described in chapter 2

2.1 General considerations

The human XPF-ERCC1 endonuclease is a large protein complex with a predicted molecular mass of 136 kDa. Given its role in nucleotide excision repair and its status as a validated drug target, a structure-function analysis of this enzyme may yield valuable information. However, large protein complexes from mammalian sources,

such as XPF-ERCC1, are rarely amenable to structural studies. XPF-ERCC1 is too large for three-dimensional structure determination by nuclear magnetic resonance measurements but too small for study by electron microscopy. Crystallography is thus the only avenue to pursue. However, limited protein expression and conformational flexibility are often hurdles in structure determination by this method.

Earlier studies have shown that it is possible to produce active XPF-ERCC1 in *E.coli* (Gaillard and Wood, 2001). Subsequent studies have shown that it is possible to produce active XPF-ERCC1 recombinantly in baculovirus-infected insect cells (Enzlin and Scharer, 2002). Expression levels in this study allowed for mutagenesis experiments that identified key residues for activity within the XPF nuclease domain. The yields were not however suitable for crystallography. Comparisons of proteins from both expression hosts suggest that XPF-ERCC1 overexpressed in *E.coli* is aggregated, whilst XPF-ERCC1 overexpressed in insect cells is monodisperse (Enzlin and Scharer, 2002). Previous studies detailing recombinant expression in *E.coli* have suggested overexpression is possible but leads to aggregated protein (Enzlin and Scharer, 2002). Therefore it was implied that insect cells were the preferred host for XPF-ERCC1 overexpression.

2.2 Bacterial expression of the XPF-ERCC1 complex and associated degradation

Using a pET30b dicistronic bacterial expression plasmid provided by members of Prof. Rick Wood's laboratory (ex-Imperial Cancer Research Fund, Clare Hall Laboratories), a former postdoctoral fellow, Matthew Newman and a scientific officer John Lally (Structural Biology Laboratory) were able to express and purify recombinant XPF-ERCC1 complexes. The dicistronic plasmid coded for an N-terminally His_6-tagged XPF and a C-terminally His_6-tagged ERCC1. The *E.coli* strain used was FB810 (BL21(DE3) recA$^-$) and affinity chromatography on nickel-NTA was used. In addition, non-aggregated XPF-ERCC1 has been produced. The key to this success appears to be use of the non-denaturing zwitterionic detergent 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS). CHAPS was added to all extraction, purification and storage buffers. The success of this expression was tempered by a significant proportion of the XPF subunit being proteolytically degraded. This occurred at two major sites within the XPF enzyme. The first yielded a 70 kDa XPF fragment and the second a 40 kDa XPF fragment.

Mass spectrometry revealed that the 70 kDa fragment was caused by an amino-terminal degradation and the first peptide identified after tryptic digest started at residue 361. The 40 kDa fragment was shown to compose of residues 629-905 of XPF. There did not appear to be any degradation of ERCC1.

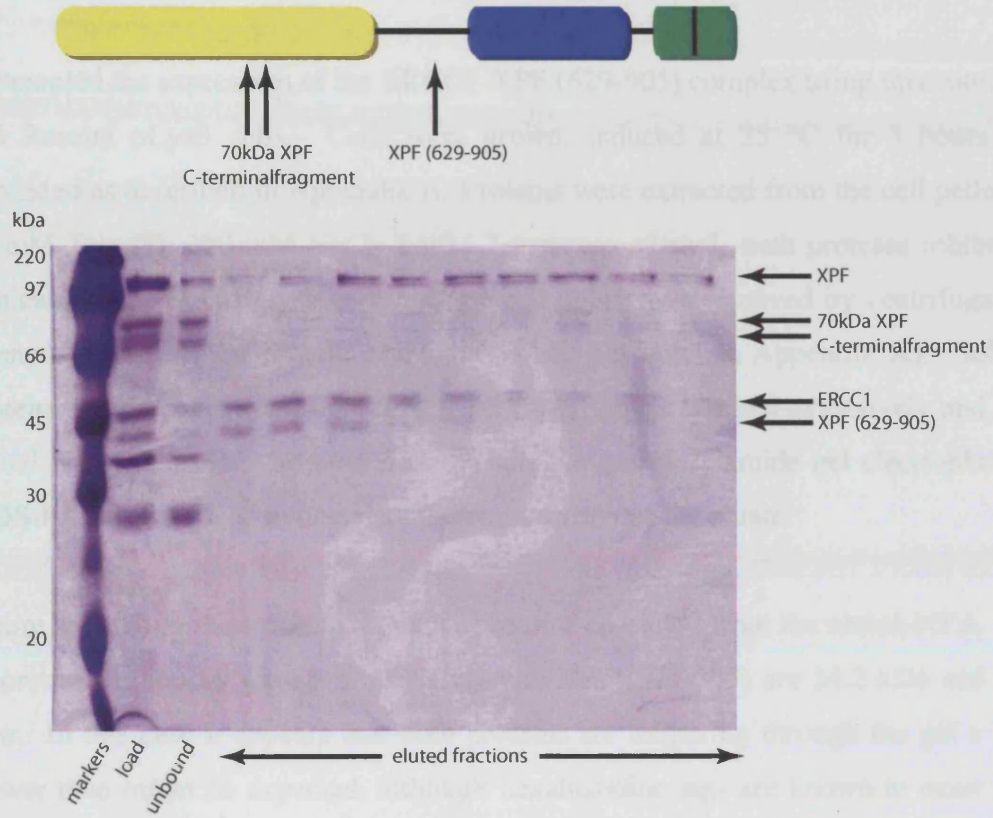

Figure 2.2 – Bacterial overexpression of the XPF-ERCC1 complex using a dicistronic expression plasmid. The SDS-PAGE gel shows protein extracted and purified by Ni^{2+}-NTA chromatography before loading onto a single-stranded DNA-agarose affinity column and elution with NaCl. The full-length XPF-ERCC1 complex is eluted from the column at high NaCl concentrations. A complex of ERCC1 and a 40 kDa XPF fragment is eluted from the column at low NaCl concentration. A significant 70 kDa XPF fragment does not bind to the DNA-agarose column. The sites of proteolysis of the XPF polypeptide with respect to its known domain organisation are shown in the schematic.

2.3 Expression, purification and characterisation of an ERCC1-XPF (629-905) complex.

As the schematic in figure 2.2 suggests, the fragment of XPF consisting of residues (629-905) encodes a little more than the core nuclease and $(HhH)_2$ domains. However, since the proteolysis in the XPF-ERCC1 preparation highlighted this

sequence as a potentially stable polypeptide, Matthew Newman (Structural Biology Laboratory) created a pET3b XPF (629-905). Using a pET30b ERCC1 construct (a gift from Prof. Rick Wood), Matthew Newman also created a dicistronic pET30b ERCC1-XPF (629-905) construct. This plasmid encodes a carboxy-terminally hexahistidine-tagged ERCC1 protein and an untagged XPF fragment.

I attempted the expression of the ERCC1-XPF (629-905) complex using this construct and Rosetta pLysS cells. Cells were grown, induced at 25 °C for 3 hours and harvested as described in Appendix A. Proteins were extracted from the cell pellets in 50 mM Tris (8), 300 mM NaCl, 2 mM 2-mercaptoethanol, with protease inhibitors. Sonication was used to lyse cells and the cell debris was removed by centrifugation. (General protocols for protein preparations can be found in Appendix A). Soluble proteins were bound to nickel-NTA, the affinity matrix washed extensively and then eluted with imidazole. Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was used to elucidate the composition of the eluate.

Figure 2.3 clearly shows that two major proteins co-eluted from the nickel-NTA. The theoretical molecular masses for ERCC1 and XPF (629-905) are 34.2 kDa and 30.6 kDa. In this case it appears that both proteins are migrating through the gel a little slower than might be expected, although hexahistidine tags are known to exert such an effect on proteins during SDS-PAGE analysis. The conclusion was that the correct complex had been formed, with a respectable expression level of 1-2 mg per litre of bacterial culture. The purity of the complex after this purification step was good. The eluate was dialysed to remove the imidazole, and concentrated to 10 mg ml^{-1} for crystallisation trials.

Hanging drop vapour diffusion and microbatch experiments were set up using commercial sparse matrix screens. However, no crystallisation hits were produced and many drops had precipitate in them. Parallel crystallisation trials with simple synthetic duplex DNA (11, 13, 15, 17 base-pairs in length) were also unsuccessful.

At this stage further biochemical characterisation of the protein was attempted. Gel filtration of ERCC1-XPF (629-905) protein was attempted in order to identify whether a homogeneous, monodisperse sample was being used for crystallisation.

After running the sample on both Sephacryl-100 and Superose-12 columns the sample did not elute as a single peak, indeed there was a lot of complex lost on the columns (data not shown). This indicated that perhaps the protein had a sticky patch that was interacting with the column. Detergents (CHAPS) and glycerol were added to alleviate these problems but with little effect.

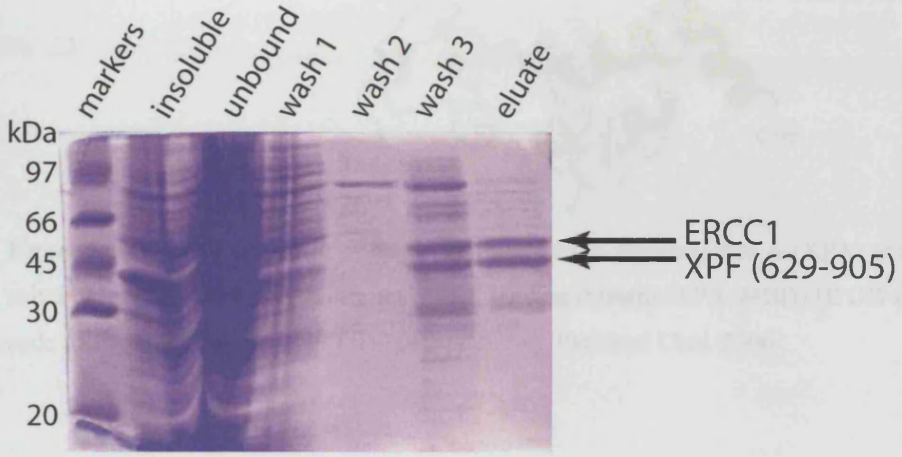

Figure 2.3 – Expression of ERCC1-XPF (629-905). SDS-PAGE gel showing fractions taken during protein extraction and purification on Ni^{2+}-NTA.

Such sticky hydrophobic patches on protein surfaces are not uncommon. In this case one might base this conclusion on the fact that the XPF (629-905) is engineered and in vivo would be part of a larger polypeptide. Thus a formerly buried surface could be accessible due to an inappropriate truncation. However, as we know that the full length XPF polypeptide is degraded options are limited in this direction. Another possibility lies in the fact that the XPF-ERCC1 endonuclease is part of the NER machinery and is specifically recruited to the site of damage by protein-protein interactions. A well-documented interaction is that with XPA. XPA is a 36 kDa zinc metalloprotein that is known for its interactions with RPA, TFIIH and ERCC1, as well as binding single-stranded DNA (figure 2.4). A linear epitope has been identified within the XPA amino acid sequence that defines the ERCC1 binding site. It was hoped that the addition of XPA as a binding partner might improve the stability of the ERCC1-XPF (629-905) complex.

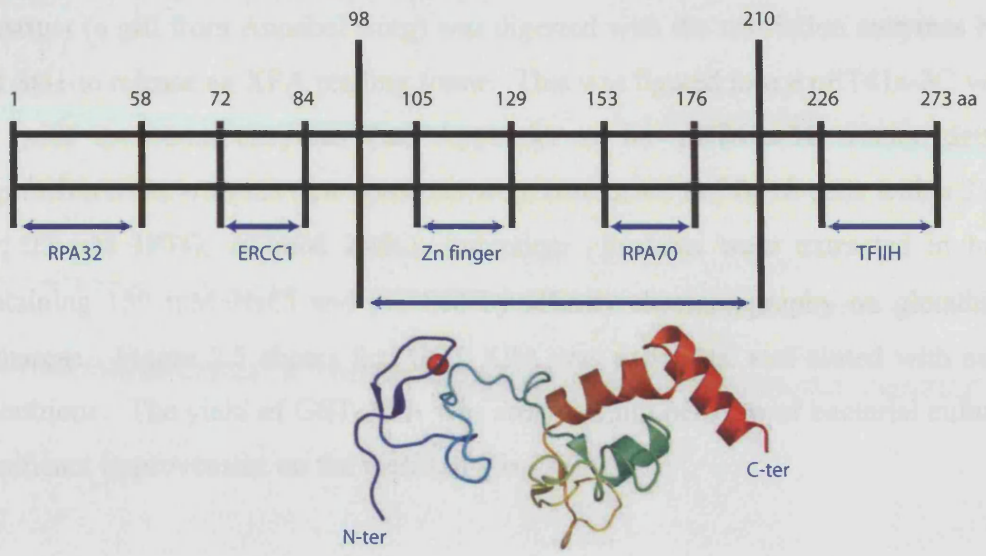

Figure 2.4 –Domain structure of the human xeroderma pigmentosum A (XPA) protein and solution structure of the XPA minimal DNA-binding domain (XPA–MBD) (IPDB accession code 1XPA (Ikegami et al., 1998)). Adapted from Park and Choi, 2006.

2.4 Preparation of an ERCC1-XPF (629-905) – XPA complex

The Peptide Synthesis Laboratory at Cancer Research UK synthesised the appropriate XPA peptide. The peptide sequence is GGKIIDTGGGFILEEEEEEE. This peptide is based on studies of the XPA-ERCC1 interaction that identify these residues as the minimal XPA fragment required for interaction with ERCC1 (Li et al., 1994; Li et al., 1995). The run of seven glutamic acid residues at the end gives this peptide an acidic nature (pI 3.88). This acidity caused solubility problems when trying to dissolve it in similar buffers to that of the ERCC1-XPF (629-905) protein. Adding a five-fold molar excess of this peptide to the purified ERCC1-XPF (629-905) protein did not improve its monodispersity on gel filtration columns. Crystallisation trials with a 2-fold molar excess of peptide resulted in the limited solubility of the peptide being explored.

However, with the ERCC1-XPA interaction possibly being of crucial importance, a study involving full length XPA was undertaken. His_6-tagged XPA had been produced in modest amounts by Annabel Borg (Protein Isolation Laboratory, Cancer Research UK). However, as the ERCC1-XPF (629-905) complex is also His_6-tagged, this precluded a simple pull-down experiment. To circumvent this problem, a new GST-tagged XPA construct was made. The original pET15b-XPA expression

construct (a gift from Annabel Borg) was digested with the restriction enzymes NcoI and SalI to release an XPA reading frame. This was ligated into a pET41a-3C vector cut with the same enzymes (see Appendix A for pET41a-3C vector details). Expression trials with the new construct were conducted in FB810 cells with a 25 °C, 3hr, 25 μM IPTG, 10 mM $ZnSO_4$ induction. Proteins were extracted in buffer containing 150 mM NaCl and purified by affinity chromatography on glutathione-sepharose. Figure 2.5 shows that GST-XPA was expressed and eluted with excess glutathione. The yield of GST-XPA was around 2 mg per litre of bacterial culture, a significant improvement on the yields of His_6-XPA.

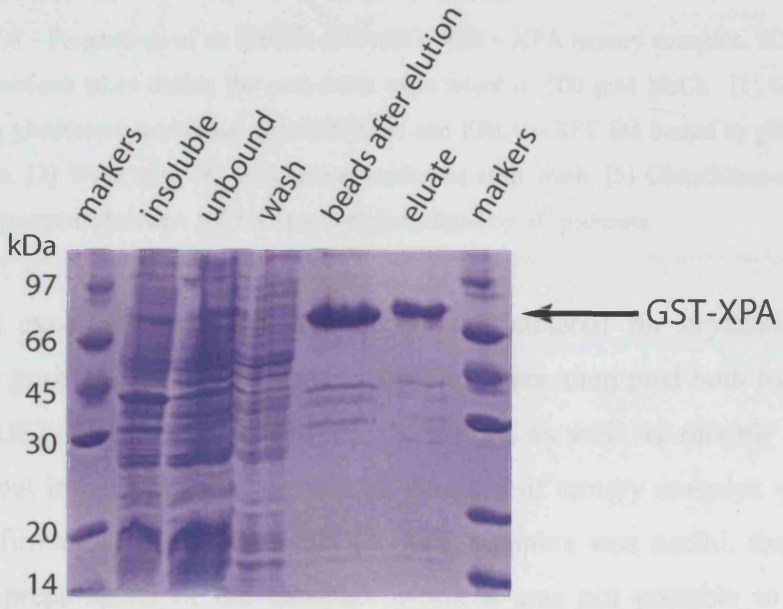

Figure 2.5 – Expression of GST-tagged XPA. SDS-PAGE gel showing fractions taken during protein extraction and purification on glutathione sepharose

The eluted GST-XPA was dialysed free of glutathione and rebound to a fresh batch of glutathione-sepharose. Purified ERCC1-XPF (629-905) was then passed over these beads in the hope of XPA-ERCC1-XPF (629-905) ternary complex formation. Two salt concentrations were used for these experiments 100 mM and 500 mM NaCl. As figure 2.6 shows ternary complex formation occurred and the GST-XPA pulled down the ERCC1-XPF (629-905). This complex was then eluted from the affinity column by cleavage with 3C protease. The complex formation was not efficient and some ERCC1-XPF (629-905) was not pulled down by the XPA.

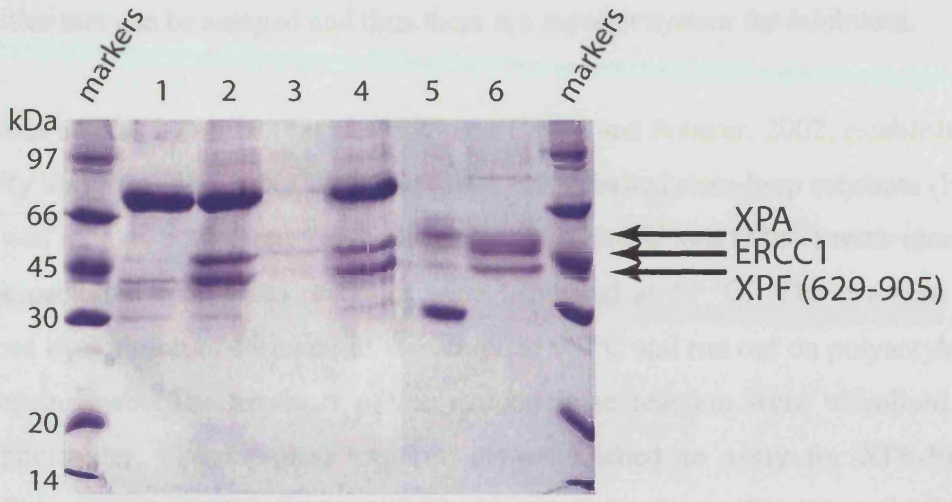

Figure 2.6 – Preparation of an ERCC1-XPF (629-905) – XPA ternary complex. SDS-PAGE gel of fractions taken during the pull-down experiment in 500 mM NaCl. [1] GST-XPA bound to glutathione-sepharose. [2] GST-XPA and ERCC1-XPF D3 bound to glutathione-sepharose. [3] Wash step. [4] Glutathione-sepharose after wash. [5] Glutathione-sepharose after 3C protease cleavage. [6] Ternary complex eluted by 3C protease.

This initial experiment did not provide enough material for crystallisation trials. Large scale purifications of this ternary complex were attempted both by mixing cell pellets of GST-XPA and ERCC1-XPF (629-905), as well as mixing lysates after sonication but in both cases no significant amounts of ternary complex were formed. Whilst confirmation of the XPA-ERCC1-XPF complex was useful, the inability to "scale-up" preparations of the complex meant it was not possible to exploit this interaction to aid our structural goals. A recent study investigating the recognition of helical kinks by XPA also suggests that XPA binds ERCC1-XPF more readily when associated with a distorted DNA substrate (Camenisch et al., 2006). It may be possible to exploit this enhanced interaction in future experiments.

2.5 Activity of the ERCC1-XPF (629-905) complex

As has been mentioned, ERCC1-XPF's critical role in NER makes it an attractive drug target. Inhibition of XPF activity would stop NER. NER is responsive to chemotherapy agents such as cis-platin that mediate DNA damage and subsequent apoptosis. Therefore, inhibiting NER could enhance and render many tumour types sensitive to cis-platin based drugs. If this inhibition could be targeted to specific cells such as tumour cells, then such an inhibitor drug could be very desirable. XPF-

ERCC1 and XPG are particularly good targets because, as endonucleases, they have activities that can be assayed and thus there is a reporter system for inhibition.

Previous studies by de Laat et al., 1998 and Enzlin and Scharer, 2002, established an activity assay for XPF-ERCC1 activity. A ^{32}P 5'-labelled stem-loop substrate (Figure 2.8) was used as a synthetic NER substrate. Manganese ions (Mn^{2+}) were identified as the preferred cation and reactions were incubated at 28 °C. The reactions were stopped by addition of formamide, denatured at 95 °C and run out on polyacrylamide sequencing gels. The products of the endonuclease reaction were visualised on a Phosphoimager. Whilst these experiments established an assay for XPF-ERCC1 activity and showed that the recombinant enzyme is active; this type of gel-based assay is not suitable for high-throughput inhibitor screening. The number of gels that would have to be run would be prohibitive and the use of radiolabelled phosphorus on a large scale is not desirable. Also, despite the fact that XPF-ERCC1 can be expressed in bacteria, the yields are certainly not great enough to provide enough protein for a comprehensive high-throughput inhibitor screen.

The ERCC1-XPF (629-905) protein described earlier contains a whole nuclease domain and thus may be expected to contain all the residues required for endonuclease activity. An attempt to develop a high-throughput assay centred around the ERCC1-XPF (629-905) protein was initiated in collaboration with Tim Hammonds' group at Cancer Research Technology. The basic principle behind the high-throughput assay proposed is depicted in figure 2.7. A stem-loop substrate is retained with a 5' fluorophore (carboxyfluorescein, FAM) and a quencher (dabcyl) at the 3' end. When the cleavage occurs the stem becomes unstable and the labelled 5' strand is lost. The delocalisation of the fluorophore from the quencher results in an increase in fluorescence at the specific emission wavelength.

In order to verify that any perceived activity was due to the enzyme and not a contaminant in the preparation, I engineered a catalytically dead form of ERCC1-XPF (629-905). Amino acid residue 720 of XPF was mutated from aspartic acid to asparagine using the Stratagene QuikChange protocol. This mutation was based on the previous mutagenesis study on human XPF (Enzlin and Scharer, 2002) that targetted this aspartic acid residue. Its role was postulated to be as a general base in

activating a water molecule as a nucleophile required for phosphodiester bond hydrolysis. Hence, this D720N mutation was predicted to eliminate activity but not to affect the intrinsic protein fold.

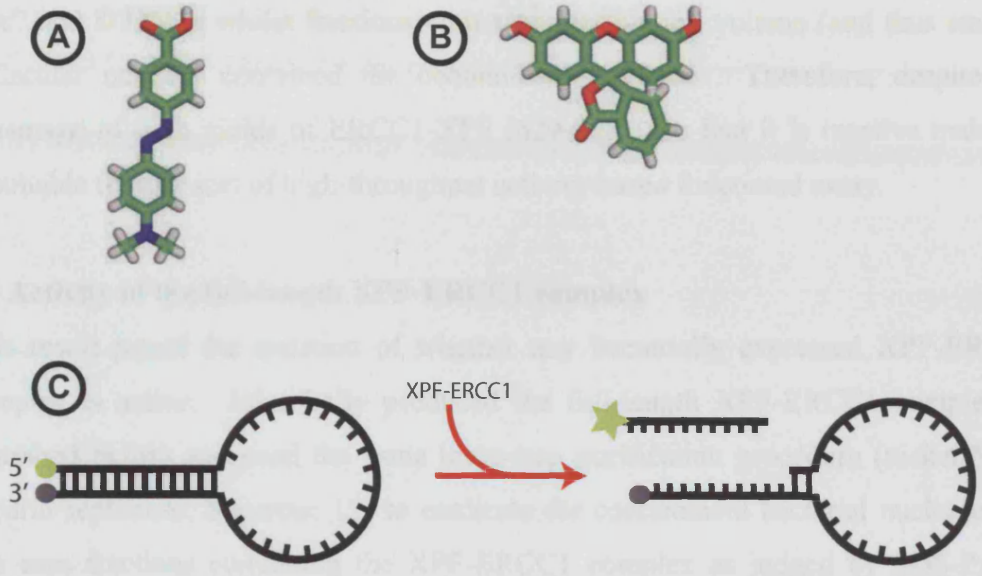

Figure 2.7 – Schematic diagram of the high-throughput assay for XPF-ERCC1 activity. [A] Structure of carboxyfluorescein. [B] Structure of dabecyl. [C] The stem-loop substrate has a 5'-fluorophore (yellow) and a 3'-quencher (grey). Upon cleavage of the substrate by ERCC1-XPF, the stem of the substrate is no long stable and the 5'-strand dissociates from the rest of the substrate. The delocalisation of the fluorophore from the quencher leads to an increase in fluorescence at 517 nm when excited with light at 492 nm.

Although some nuclease activity was present in the ERCC1-XPF (629-905) preparations, the D720N mutant proved that this was due to a contaminating *E.coli* nuclease (data not shown). This *E.coli* nuclease preferred Mg^{2+} to Mn^{2+} (the study of de Laat et al. showed XPF-ERCC1 prefers Mn^{2+} in vitro). Using radiolabelled stem/loop substrates and a gel-based assay (akin to those of de Laat et al., 1998) and the ERCC1-XPF (629-905) proteins, a scientific officer, Maureen Biggerstaff (Structural Biology Laboratory) showed that the substrate was not being cut cleanly near the double-strand/single-strand junction but rather indiscriminately along its length.

Extra purification steps were undertaken in order to remove the contaminating *E.coli* nuclease. As well as the initial purification step on nickel-NTA, John Lally (Structural Biology Laboratory) used affinity chromatography on Hi-Trap heparin-

sepharose columns and gel filtration on a Superose 12 column to improve the purity of the ERCC1-XPF (629-905) complex. By assaying various fractions from the Superose 12 columns and using the gel-based assay, Maureen Biggerstaff was able to show that fractions containing ERCC1-XPF (629-905) were inactive (both "wild-type" and D720N), whilst fractions with a greater elution volume (and thus smaller molecular masses) contained the contaminant nuclease. Therefore, despite the advantage of high yields of ERCC1-XPF (629-905), the fact it is inactive makes it unsuitable for any sort of high-throughput activity-based functional assay.

2.6 Activity of the full-length XPF-ERCC1 complex

This result posed the question of whether any bacterially expressed XPF-ERCC1 complex is active. John Lally produced the full-length XPF-ERCC1 complex as described before and used the same three-step purification procedure (nickel-NTA, heparin-sepharose, Superose 12) to eradicate the contaminant bacterial nuclease. In this case fractions containing the XPF-ERCC1 complex as judged by SDS-PAGE were active and cleaved specifically near the double-strand/single-strand junction, whilst later fractions contained the contaminant nuclease. In order to confirm that the observed structure-specific activity was due to XPF-ERCC1, I used the QuikChange mutagenesis strategy to make the D720N mutation in the full-length XPF-ERCC1 construct. As expected the D720N mutant showed no activity and was catalytically dead (figure 2.8).

With the evidence of real structure-specific activity, Maureen Biggerstaff successfully transferred the principles behind the radioactive gel-based assay to a fluorescence-based high-throughput assay. A stem-loop substrate was again used, with a fluorophore and a quencher. Mn^{2+} ions were used in the reaction buffer. A key consideration in this assay is the length of the substrate's stem. The stem is ideally short to increase the oligonucleotide yield. However, the stem has to be long enough to be recognised by the enzyme. On top of this, the stem has to be long enough to be annealed at the assay temperature; yet short enough to be destabilised after XPF-ERCC1 has cleaved it. The salt concentration of the reaction is critical in these considerations. In the end, a stem 10/loop 22 substrate was chosen with a reaction temperature of 25 °C in a buffer containing 20 mM NaCl.

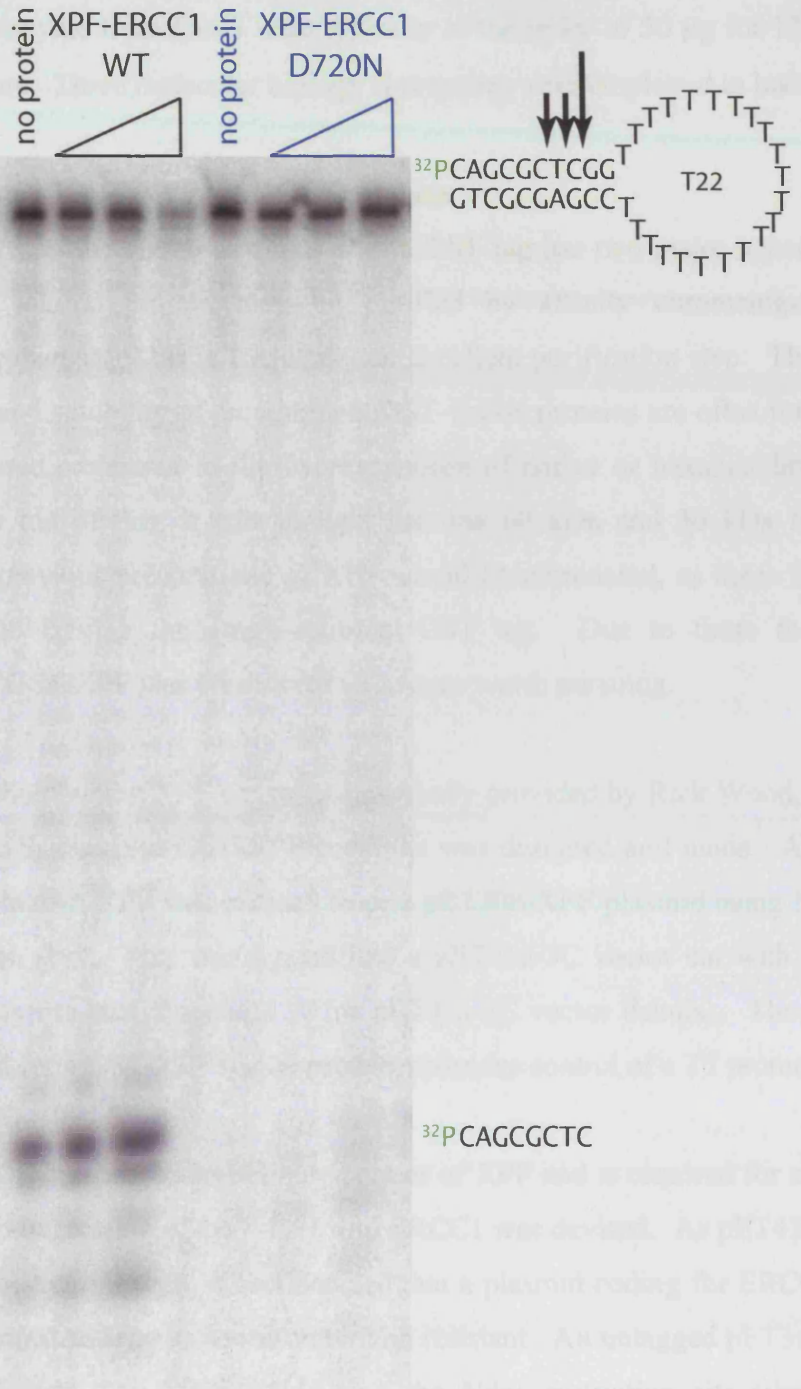

Figure 2.8 – XPF-ERCC1 activity. An autoradiogram showing the activity of increasing amounts of XPF-ERCC1 against a stem 10 – loop 22 substrate. Note the significant depletion of the stem-loop substrate at higher wild-type XPF-ERCC1 concentrations. Also note that the D720N mutant complex shows no significant activity.

Another key element in this assay is of course the XPF-ERCC1 enzyme. In order to screen tens of thousands of compounds for inhibition, it was necessary to produce milligrams of XPF-ERCC1. Clearly this was a problem, as the existing yields of the

recombinant enzyme from *E.coli* were typically of the order of 50 μg for 12 litres of bacterial culture. Three molecular biology approaches were employed to boost yields.

2.7 Expression of GST-XPF – ERCC1 complex

Tagging protein at the amino-terminus with a GST tag has two major benefits. The first is that soluble protein may be purified by affinity chromatography on glutathione-sepharose. This is frequently an excellent purification step. The second is that yields and solubility of recombinant GST-fusion proteins are often found to be greatly increased compared to the overexpression of native or hexahistidine tagged proteins. On top of this, it was thought that the 60 kDa and 30 kDa fragments generated in previous preparations of XPF would be eliminated, as these fragments would also be lacking the amino-terminal GST tag. Due to these factors, an expression of GST-XPF was considered an avenue worth pursuing.

Using an existing human XPF construct (originally provided by Rick Wood, Imperial Cancer Research Fund), a GST-XPF construct was designed and made. An coding sequence for human XPF was excised from a pET30b-XPF plasmid using NcoI and SalI restriction sites. This was ligated into a pET41a-3C vector cut with the same restriction enzymes (see Appendix A for pET41a-3C vector details). The resultant plasmid coded for a GST-XPF fusion protein under the control of a T7 promoter.

Since ERCC1 is considered an obligate partner of XPF and is required for activity, a strategy for co-expression of GST-XPF and ERCC1 was devised. As pET41a-3C is a kanamycin-resistant plasmid, it was decided that a plasmid coding for ERCC1 could be co-transformed so long as it was ampicillin resistant. An untagged pET3a ERCC1 expression plasmid was constructed using the NdeI restriction site (the ERCC1 reading frame was taken from pET3/4b ERCC1, a plasmid generated by Matthew Newman). The orientation of the ERCC1 reading frame within the NdeI site was checked by DNA sequencing. Both pET41a-3C XPF and pET3a ERCC1 were co-transformed in FB810 cells, selected for with both kanamycin and ampicillin, and expression trials were conducted.

Bacteria were grown at 37 °C and induced at 18 °C. Protein was extracted as for the previous XPF-ERCC1 complex preparations. Protein was purified by affinity

chromatography on glutathione-sepharose. Proteins were eluted with buffer containing 20 mM reduced glutathione (pH 8). SDS-PAGE analysis of the preparations (figure 2.9) showed that the majority of GST-XPF protein is not eluted from the glutathione sepharose but rather is stuck to the affinity matrix. In addition, little ERCC1 appears to be made at all. Since ERCC1 is an obligate partner to XPF, and is required for activity, this method was not considered a viable route to enhanced XPF yields.

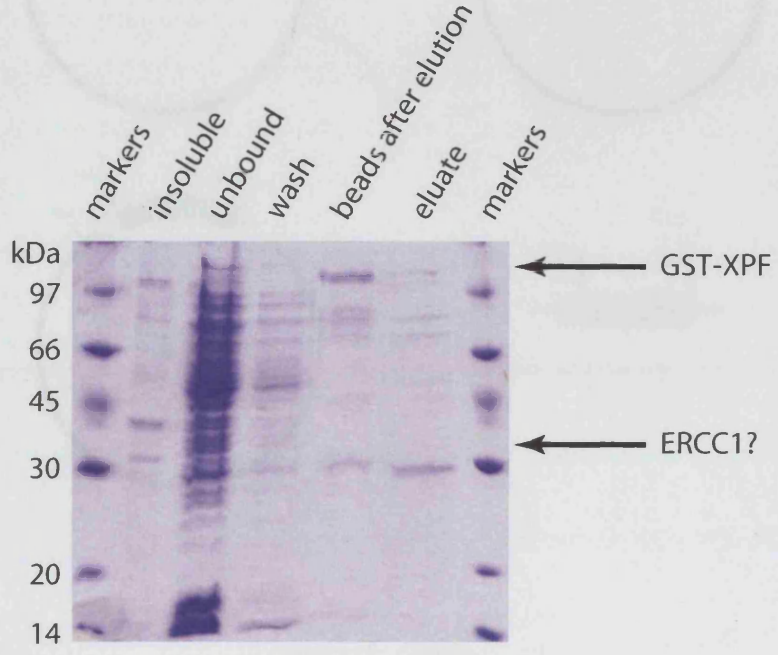

Figure 2.9 – Expression of GST-XPF – ERCC1. SDS-PAGE gel showing fractions taken during protein extraction and purification on glutathione sepharose

2.8 Expression using a redesigned ERCC1-XPF dicistronic vector

The respectable yield of the ERCC1-XPF (629-905) construct is interesting for many reasons. The enhanced yield of the XPF (629-905) fragment relative to that of full-length XPF is not surprising as it is a shorter and probably structurally more compact polypeptide. The enhanced expression of ERCC1 when co-expressed with XPF (629-905) compared to when expressed with full-length XPF is slightly puzzling, as both ERCC1 coding sequences are identical. One possible explanation for this observation is that for the ERCC1-XPF (629-905) construct, the ERCC1 is transcribed first from within the dicistronic plasmid. In the full-length XPF-ERCC1 construct the ERCC1

is transcribed after the XPF. It was felt that the order in which the ERCC1 and XPF coding sequences were placed within the dicistronic plasmid could be significant and was worth investigating.

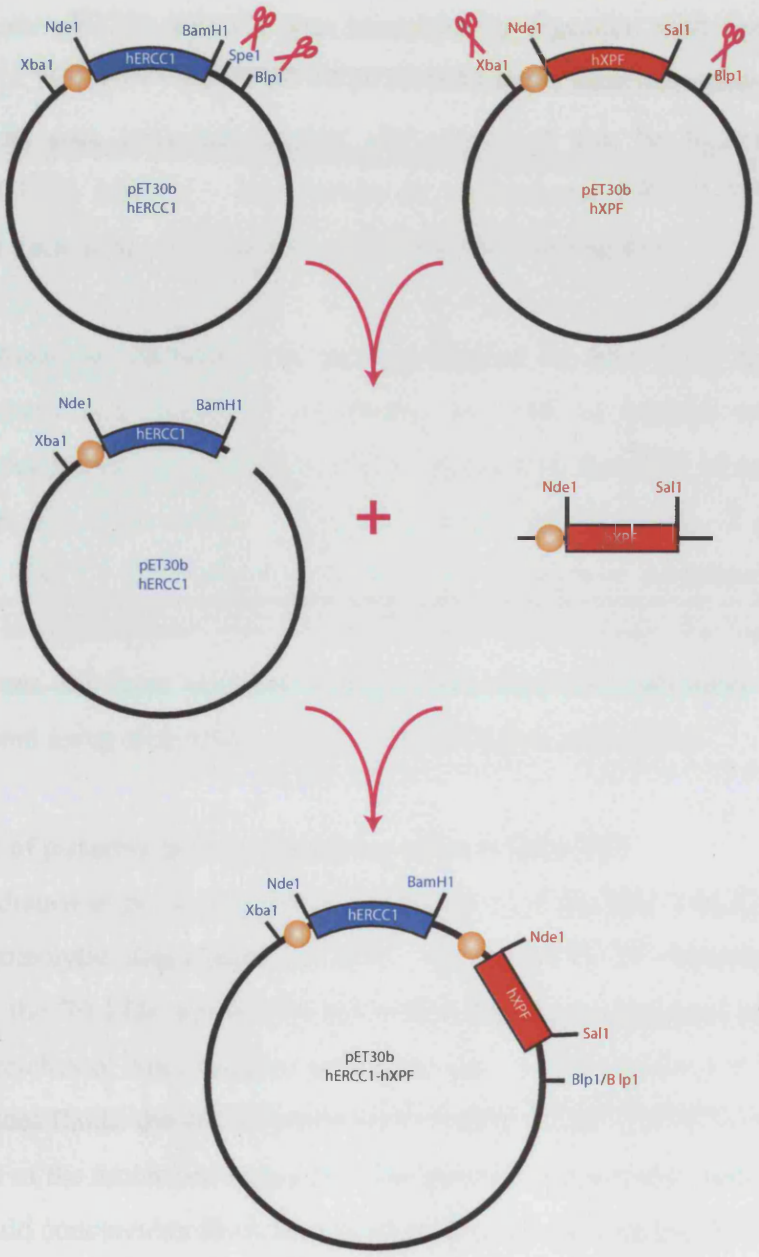

Figure 2.10 – Construction of a new ERCC1-XPF dicistronic vector. This strategy relies on the presence of a unique Xba1 site before the ribosome binding site in pET vectors and the fact that Xba1 and Spe1 sites have identical cohesive overhangs after cleavage. Orange circles indicate ribosome-binding sites.

To this end, a new ERCC1-XPF dicistronic plasmid was created. The method of creation of the dicistronic plasmid is shown in figure 2.10. The method exploits the

fact that pET vectors have a unique Xba1 site upstream of the ribosome-binding site, and a unique Blp1 site downstream of the multiple cloning site. Thus by cutting a pET30b XPF construct (a gift from Prof. Rick Wood) with Xba1 and Blp1, the coding sequence for XPF complete with its own ribosome-binding site was excised. The recipient vector pET30b ERCC1 was linearised by digestion with Spe1 and Blp1. Since Spe1 (A↓CTAGT) and Xba1 (T↓CTAGA) leave identical cohesive ends, the XPF (with its own ribosome-binding site) sequence can be ligated easily into linearised pET30b ERCC1. This results in a functional ERCC1-XPF dicistronic plasmid with each gene possessing its own ribosome-binding site.

Expression from this dicistronic vector was trialled by John Lally using identical bacterial growth and induction conditions, as well as protein extraction and purification conditions. In general, a 200 % increase in the yield of useful ERCC1-XPF complex has been found. This has allowed, albeit slowly, 2 milligrams of recombinant ERCC1-XPF required for the high-throughput inhibition assay to be stockpiled. In collaboration with Cancer Research Technology, the high-throughput inhibitor screen has been performed and potential specific inhibitors are currently being evaluated using material produced with this expression vector.

2.9 Deletion of putative protease sensitive loops within XPF

A major hindrance in producing milligram quantities of the XPF-ERCC1 enzyme has been the proteolytic degradation of XPF. Up to 50 % of expressed protein is degraded to the 70 kDa and 40kDa (629-905) fragments discussed in section 2.2. There are stretches of basic residues within the amino acid sequence of XPF. One of these sequences flanks the site of proteolysis causing the 40 kDa (629-905) fragment. Examination of the amino acid sequence also points to a particular basic stretch (339-360) that could conceivably flank the site of proteolysis causing the 70 kDa fragments (figure 2.11). Since the polar nature of these stretches make them unlikely to be contained in the core fold of the protein, it was hypothesised that these stretches of basic residues form flexible loops that make the XPF polypeptide sensitive to degradation.

With this in mind, a strategy was devised to delete the basic stretches from within the XPF polypeptide (figure 2.12). Overlap PCR was employed (Ho et al., 1989).

Figure 2.11 – Sequence alignment of selected XPF orthologues. Sequence numbers for each alignment panel are indicated. Magenta boxes indicate the positions of the basic stretches referred to in the text that were deleted. The orange line demarcates the helicase domain; the blue line, the nuclease domain; and the green line, the $(HhH)_2$ domain. The positions of human XPF residues 629 and 720 are shown by red stars.

Complementary primers and PCR are used to generate primary PCR products having overlapping ends. These primary products are combined in a subsequent reaction in which the overlapping ends anneal, allowing the 3'-overlap of each strand to serve as a primer for the 3'-extension of the complementary strand. Flanking primers can be used to amplify this product by standard PCR. This process can be repeated depending on the strategy and number of changes desired. In this case, the primers anneal near the site of the desired loop deletions.

The basic stretches in XPF that were targeted were as follows (and figure 2.11). The sequence coding for amino acid residues 339 to 360 was changed from KMSKKEKISEKMEIKEGEETKK to TGS. A short replacement linker was used in this case, as this putative loop is entirely absent in many other eukaryotic XPF homologues (figure 2.11). The sequence coding for residues 475 to 480 was changed from KKKKRK to TGSTGS. Finally the sequence coding for residues 613 to 626 was changed from RKEKEAFEKLIREK to TGSSEAFEKLISES. It was felt necessary to introduce some residues to replace these putative loops. The sequence $TGS_{(n)}$ has been used previously as a flexible linker to couple two weakly interacting proteins into a single polypeptide (Pellegrini et al., 2002).

The overlap PCR strategy worked and a coding sequence for the XPFΔloops was produced. This sequence was cloned into pET14b-3C using NcoI and NdeI restriction sites. After this, an ERCC1- XPFΔloops dicistronic plasmid was constructed in the manner explained previously in section 2.8 and figure 2.10. (An XbaI-BlpI fragment from pET14b-3C XPFΔloops was ligated into pET30b-ERCC1 linearised with SpeI and BlpI).

Expression trials from this dicistronic plasmid were conducted. Bacteria were grown at 37 °C and then induced at 18 °C. Proteins were extracted as for the wild-type ERCC1-XPF complex in the presence of CHAPS. The protein was purified by affinity chromatography on Ni2+-NTA. However, SDS-PAGE analysis (figure 2.13) after this step revealed no significant ERCC1-XPFΔloops expression. Subsequent trials with other induction conditions yielded similarly disappointing results.

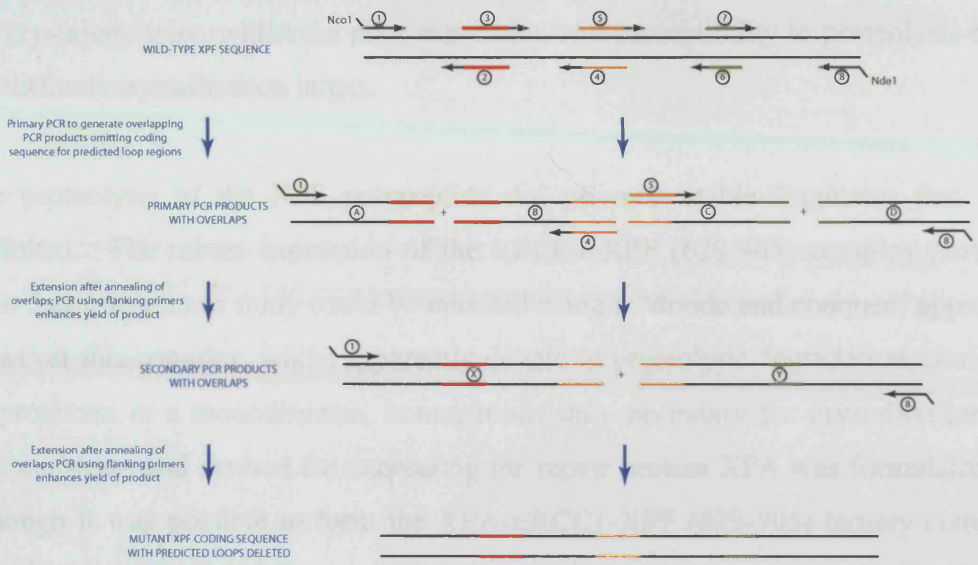

Figure 2.12 – Strategy for generating XPFΔloops coding sequence. The red lines represent the deletion of residues 339 to 360. The orange lines represent the mutation of residues 475 to 480. The green lines represent the mutation of residues 613 to 626.

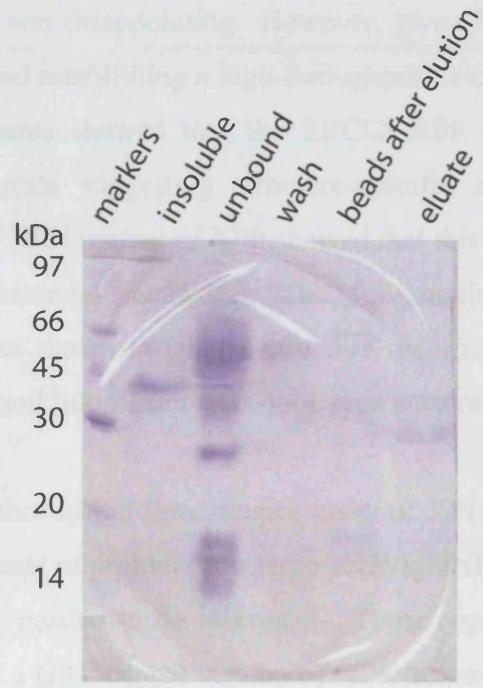

Figure 2.13 – Expression of ERCC1-XPFΔloops. SDS-PAGE gel showing fractions taken during protein extraction and purification on Ni^{2+}-NTA.

2.10 Summary

Structural characterisation of the full-length XPF-ERCC1 complex proved to be unfeasible at this time. Its size prevents study by structural techniques other than X-

ray crystallography, whilst the poor expression and susceptibility to proteolysis make it a difficult crystallisation target.

The proteolysis of the XPF polypeptide did generate stable fragments that were exploited. The robust expression of the ERCC1-XPF (629-905) complex provided hope that a structural study could be initiated using a "divide and conquer" approach. However this complex, whilst apparently devoid of proteolytic degradation, could not be produced in a monodisperse, homogenous state necessary for crystallisation. A new construct and method for expressing the repair protein XPA was formulated and although it was possible to form the XPA-ERCC1-XPF (629-905) ternary complex; quantities of such a complex necessary for crystallography were not obtainable.

The lack of success in these initial efforts to obtain structural information for the XPF-ERCC1 complex was disappointing. However, given XPF-ERCC1's status as a drug target, pursuing and establishing a high-throughput functional assay was of equal importance. Experiments showed that the ERCC1-XPF (629-905) was inactive, despite early experiments suggesting structure-specific activity. However, the expression of a dead D720N mutant of XPF showed that this activity was actually due to a contaminating bacterial nuclease. The full-length XPF-ERCC1 complex expressed in *E.coli* was shown to exhibit structure-specific activity in assays using both ^{32}P-radiolabelled and fluorescent stem-loop type substrates.

A protocol for a high-throughput fluorescence assay of XPF was devised. However, to obtain suitable amounts of protein for a large-scale inhibitor screen, the expression levels of XPF-ERCC1 needed to be increased. Three approaches were employed. The first was to create a GST-tagged version of XPF that would hopefully eliminate the 70 and 40 kDa proteolysis fragments from the enzyme preparations as well as increasing the yield of the enzymes. Although expression of GST-XPF was observed, there appeared to be no ERCC1 co-purified and the GST-XPF was not eluted from a glutathione-sepharose affinity matrix. Another approach was to delete putative protease sensitive loops from the XPF sequence. An overlap PCR approach was employed to achieve this and a dicistronic plasmid of ERCC1-XPFΔloops was constructed. However, protein expression experiments using this construct proved

fruitless. The most successful approach to enhance XPF-ERCC1 yields involved creating a new dicistronic plasmid where the respective positions of XPF and ERCC1 were changed. In the redesigned vector, the coding sequence for ERCC1 is transcribed first, followed by that of XPF. Such a rearrangement of coding sequences has enhanced the yield of XPF-ERCC1 by 200 %. This has allowed the successful screening of 55,000 compound library for potential inhibitors of XPF-ERCC1 activity.

3 Structural analysis of *Aeropyrum pernix* XPF domains and complexes with DNA substrates.

The presence of archaeal homologues of XPF has been discussed in Chapter 1. This chapter describes the structure determination of the nuclease domain of the XPF homologue from *Aeropyrum pernix* and attempts to crystallise this enzyme with different DNA substrates. A description of the experiments carried out is illustrated in the flowchart in Figure 3.1.

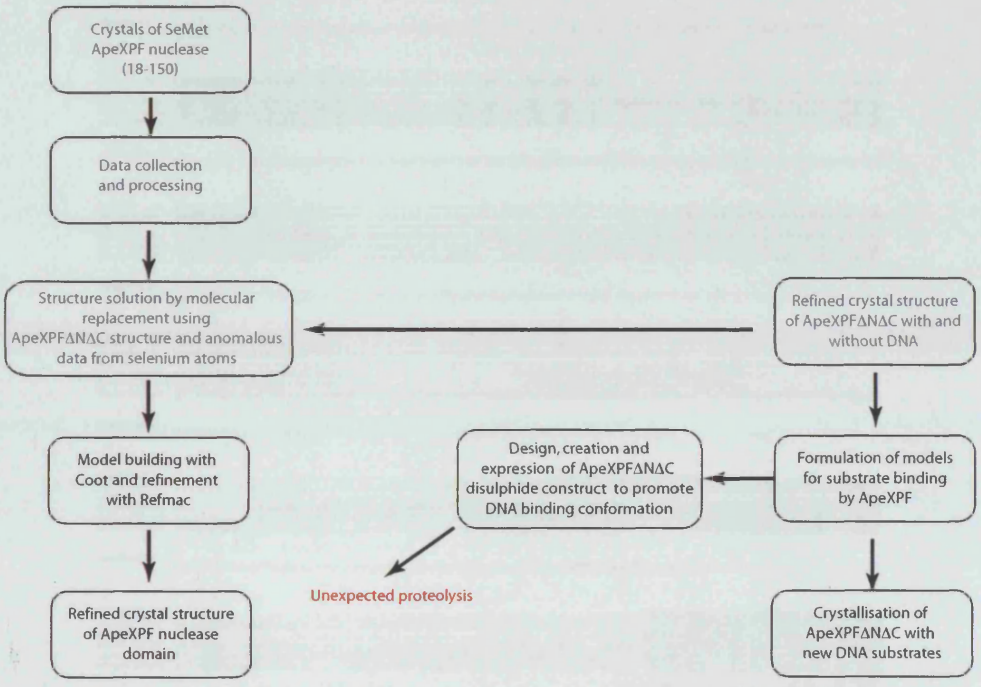

Figure 3.1 – Flowchart of the experiments described in chapter 3. Experiments in grey were largely performed by others.

3.1 Expression, purification, crystallisation and structure solution of the XPF homologue from *Aeropyrum pernix*

The use of model organisms in the study of DNA replication, recombination and repair is well documented. Virtually every aspect of DNA biology was first studied in *E.coli* and yeast before analogies were drawn to mammalian pathways.

Although there are XPF homologues in lower eukaryotes such as *S. cerevisiae* and *S. pombe*, they remain large proteins that are difficult to express recombinantly and thus

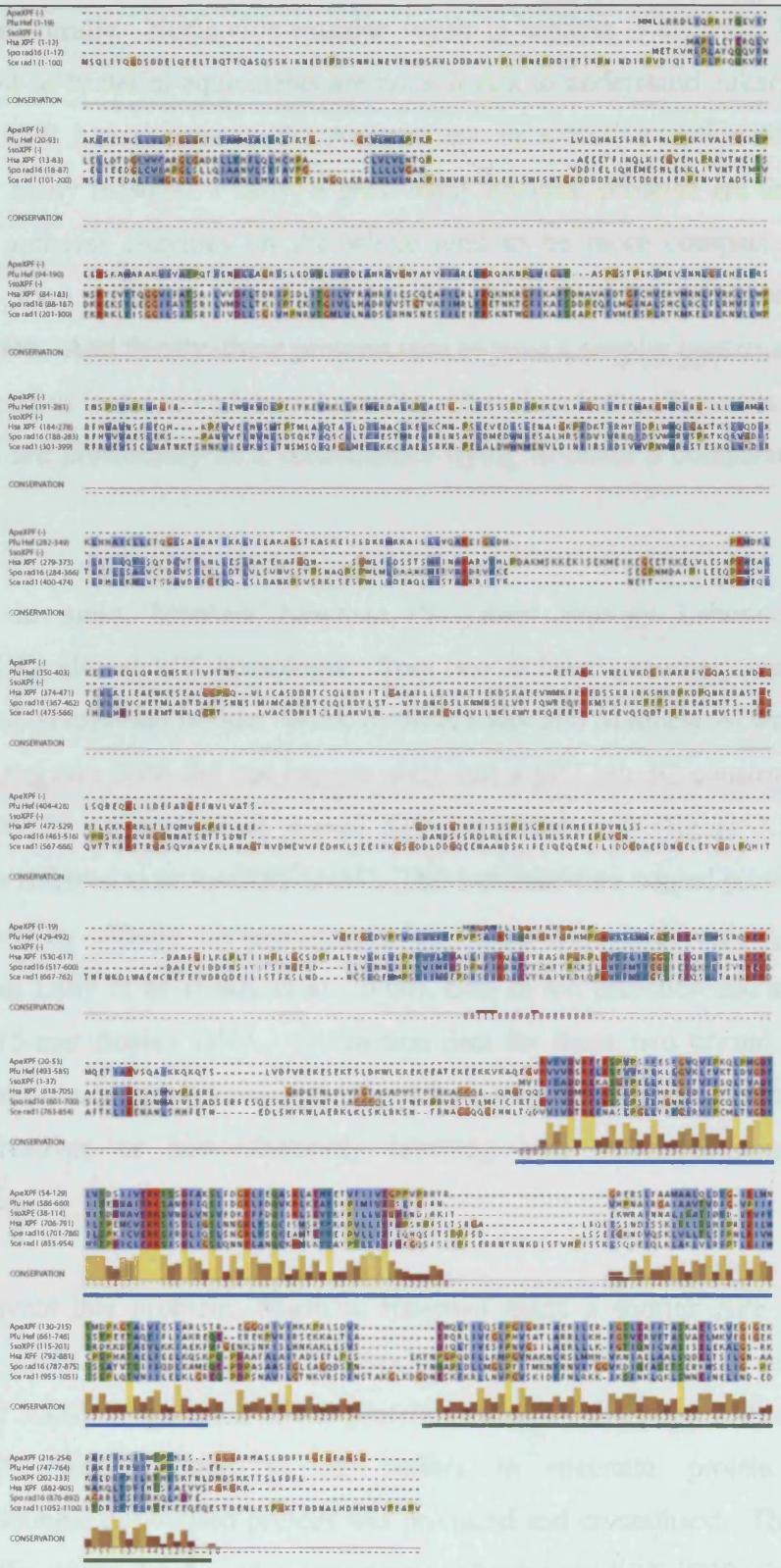

Figure 3.2 –Sequence alignment of selected archaeal and eukaryotic XPF homologues. Sequence numbers for each alignment panel are indicated. Ape = *Aeropyrum pernix*, Pfu = *Pyrococcus furiosus*, Sso = *Sulfolobus solfataricus*, Hsa = *Homo sapiens*, Spo = *Schizosaccharomyces pombe* and Sce = *Saccharomyces cerevisiae*. The blue line demarcates the nuclease domain; the green line, the $(HhH)_2$ domain.

to study structurally. Nucleotide excision repair in bacteria is achieved by different proteins and so bacterial equivalents are not relevant to understand eukaryotic XPFs. However, XPF homologues in the Archaea are an attractive option for structural studies for many reasons. Firstly, a great many archaeal proteins are thermostable. Secondly, archaeal enzymes on the whole tend to be more compact with fewer disordered loop regions. This often makes these proteins more amenable to crystallisation. And thirdly, these proteins tend to have a simpler quaternary structure as they are often found to be homooligomers, rather than heterooligomers. These last two points are presumably as a consequence trying to retain a comparatively small genome.

With this in mind, Matthew Newman (Structural Biology Laboratory, Cancer Research UK) cloned XPF homologues from two archaeal genomes; *Archaeoglobus fulgidus* and *Aeropyrum pernix*. Work by John Lally and Matthew Newman showed that the *A.fulgidus* XPF did not express well, but a pET14b-3C construct encoding residues 18-231 of *Aeropyrum pernix* XPF expressed well (figure 3.3 a). This construct is referred to as ApeXPFΔNΔC. This hexahistidine-tagged protein construct was purified by affinity chromatography on nickel-NTA and was crystallised as described in Lally et al. (Lally et al., 2004), both in the presence and absence of a synthetic 15-mer duplex DNA. Diffraction data for these two crystal forms were collected but obtaining experimental phases by either extensive screening for heavy atom derivatives or biosynthetically labelling with selenomethionine proved problematic.

To circumvent this problem, Matthew Newman made a shorter XPF pET14b-3C construct, which corresponded to the ApeXPF nuclease domain only (residues 18-150, figure 3.3 b). Expression of this protein construct was again good, although 0.5 % CHAPS was required in all buffers to maintain protein solubility. Selenomethionine derivatised protein was produced and crystallised. These triclinic crystals diffracted to 2.5 Å, and a dataset was collected at ESRF ID29 (figure 3.4 and table 3.1).

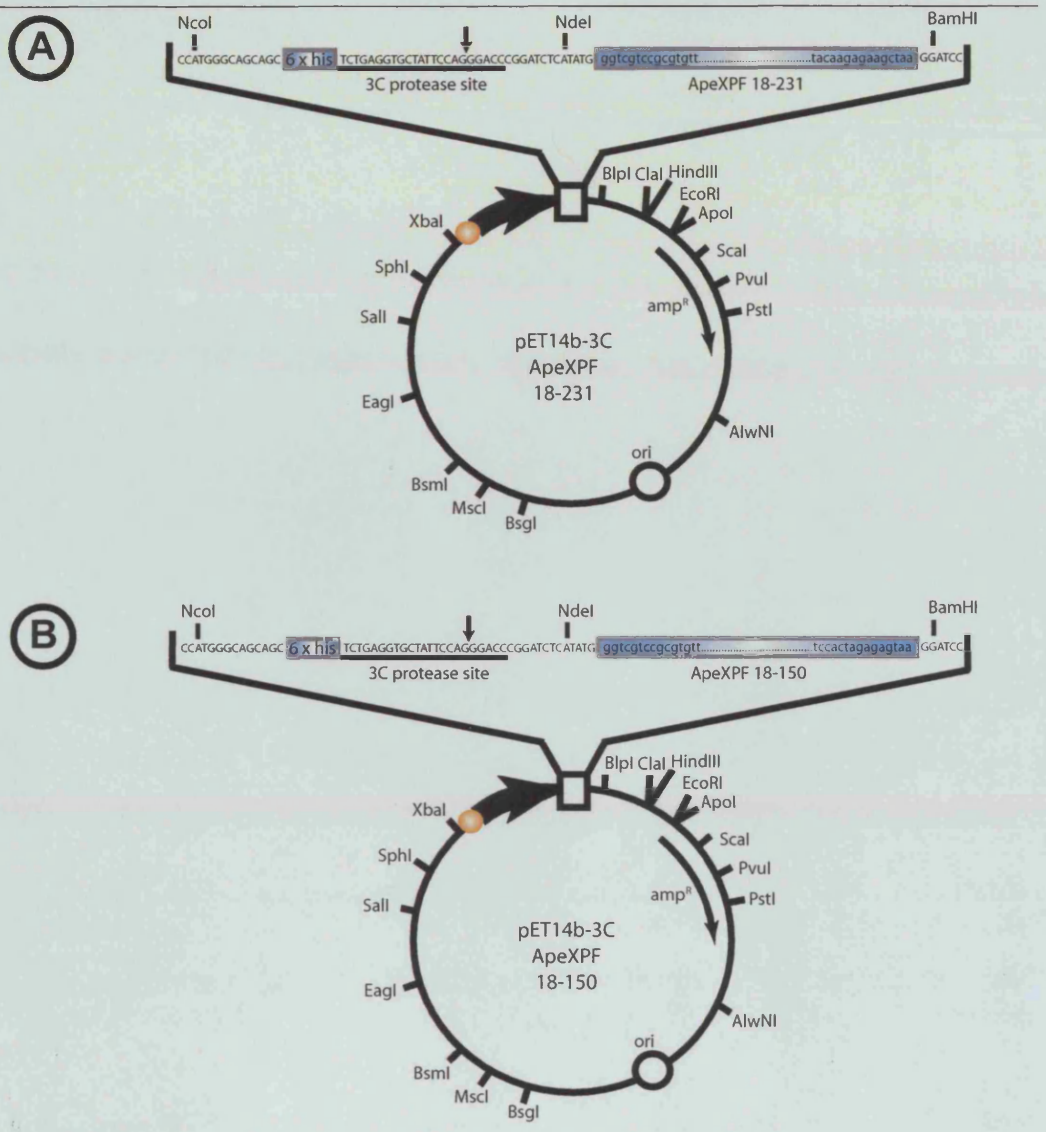

Figure 3.3 – Schematic diagram of ApeXPF expression plasmids used in chapter 3.

This dataset was processed and scaled, and the structure of this ApeXPF nuclease domain was solved by Judith Murray-Rust (Structural Biology Laboratory). This was achieved by molecular replacement using the structure of the *Pyrococcus furiosus* Hef nuclease domain that had been recently published at the time (Nishino et al., 2003)PDB code 1J23). An anomalous difference Fourier map confirmed the molecular replacement solution by indicating the position of the selenium sites. Judith Murray-Rust used a partially refined (Rfact =35.2 %; Rfree = 41.5 %) ApeXPF nuclease model to solve the larger ApeXPF (18-231) structure and the ApeXPF-DNA complex as mentioned in (Newman et al., 2005). This structure has already been addressed briefly in Chapter 1.

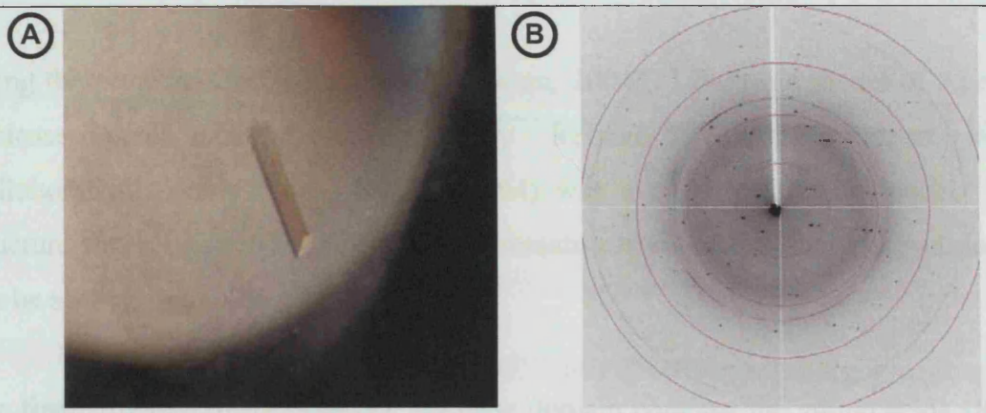

Figure 3.4 – Crystals and diffraction for ApeXPF nuclease domain. [A] Single ApeXPF nuclease crystal grown in 10 % PEG 400, 50 mM sodium acetate (pH 5). [B] Diffraction data image from ESRF BM14. Circles are at 2.1, 2.8, 4.2 and 8.4 Å.

3.2 Building, refinement and validation of ApeXPF nuclease domain model.

The structure of the ApeXPF nuclease on its own remained unrefined. Since the data collected extended to a higher resolution than observed for the ApeXPF (18-231) DNA complex, it was thought that full refinement would yield interesting information on the active site and the water structure around the nuclease domain. Towards this end, the structure of the ApeXPF nuclease domain was re-solved by Judith Murray-Rust by molecular replacement. Molrep was implemented in CCP4 (Collaborative Computational Project, 1994) with a validated refined ApeXPF nuclease structure from the ApeXPFΔNΔC-DNA complex. This yielded better electron density maps than the initial molecular replacement with the Hef nuclease domain.

Data collection site	ESRF ID29
Space group	P1
Cell a, b, c (Å)	33.65, 38.68, 55.33
Cell α, β, γ (°)	89.04, 102.90, 115.66
Number of protomers in a.u.	2
Resolution range (Å)	34.731-2.10 (2.21-2.10)
Wavelength (Å)	0.979127
Total number of reflections	50529
Unique Reflections	13094
$\langle I/\sigma I \rangle$	8.6 (5.2)
Completeness (%)	91.4 (91.8)
R_{sym}	0.071 (0.105)

Table 3.1 – Data processing statistics for ApeXPF nuclease domain crystals. Figures in parentheses refer to the highest resolution shell. Information provided by Judith-Murray Rust.

Using the program Coot (Emsley and Cowtan, 2004)., I built the model of ApeXPF nuclease domain into the electron density. Refmac 5 ((Murshudov et al., 1997); Colloborative Computational Project, 1994) was used to refine the model. The structure was refined to 2.1 Å and a representative view of the final electron density can be seen in figure 3.5.

The final structure of the ApeXPF nuclease domain contains the amino acids 18-148 with the remaining two amino acids at the carboxy-termini of the polypeptides disordered and not visible in the electron density. The refinement of the structure was concluded when the R_{fact} and R_{free} converged at 20.4 % and 25.9 % respectively. The final model has acceptable stereochemistry. Root mean square deviation (RMSD) values from target bond lengths and target bond angles are reasonable at 0.015 Å and 1.498 ° respectively (table 3.2). The two protomers present in the asymmetric unit are very similar: least squares superposition of the two molecules gave a RMSD of 0.288 Å.

During refinement if became clear that the side chains for several residues were disordered and thus there was insufficient electron density to be modelled into. Details of these residues are given in table 3.3.

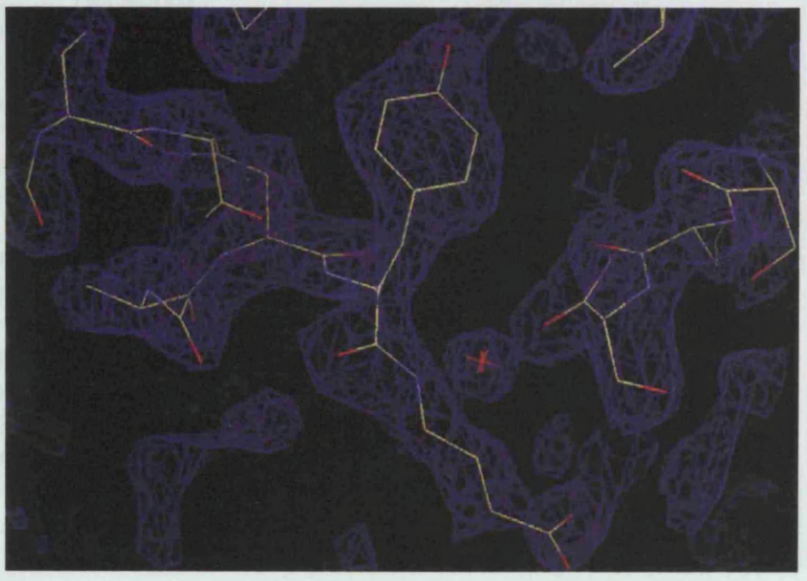

Figure 3.5 – Final weighted electron density map (2Fo-Fc) contoured at 1σ with superposed ApeXPF nuclease domain model. Screenshot from Coot (Emsley and Cowtan, 2004).

Using the program PROCHECK (Laskowski et al., 1993), it is possible to generate Ramachandran plots to evaluate how the main chain dihedral angles ϕ and φ compare to theoretically allowed limits. Figure 3.6 shows the Ramachandran plot. Excluding glycine and proline residues, 97.3 % of residues fall within the most favoured regions of the plot and 2.7 % in additionally allowed regions. There are no residues that fall within disallowed regions of the plot.

Model	
Protein atoms in final model	2060
Solvent atoms	24
Protein atoms missing from the structure	138
Diffraction Agreement	
Resolution limits (Å)	34.731- 2.100
Total reflections (test for R_{free})	12381 (649)
R_{fact} (%)	20.5
R_{free} (%)	25.9
Stereochemistry	
RMSD bond lengths (Å)	0.015
RMSD bond angles (°)	1.498

Table 3.2 – Statistics for final ApeXPF nuclease model from REFMAC 5.0.

Residue	Chain A	Chain B
Arg 20	Cγ-Cδ-Nε-Cζ-Nη1 and Nη2	Cγ-Cδ-Nε-Cζ-Nη1 and Nη2
Arg 26	Cδ-Nε-Cζ-Nη1 and Nη2	Cγ-Cδ-Nε-Cζ-Nη1 and Nη2
Glu 27	Cβ- Cγ-Cδ-Oε1 and Oε2	-
Glu 28	Oε1 and Oε2	Cδ-Oε1 and Oε2
Arg 29	Cβ-Cγ-Cδ-Nε-Cζ-Nη1 and Nη2	-
Glu 37	Cβ-Cγ-Cδ-Oε1 and Oε2	Cγ-Cδ-Oε1 and Oε2
Lys 46	Cε-Nζ	Cδ-Cε-Nζ
Gln 47	Cγ-Cδ-Oε and Nε	Cγ-Cδ-Oε and Nε
Asp 57	Oδ1 and Oδ2	Oδ2
Glu 62	Cγ-Cδ-Oε1 and Oε2	
Lys 71	Cγ-Cδ-Cε-Nζ	Cδ-Cε-Nζ
Arg 77	Cγ-Cδ-Nε-Cζ-Nη1 and Nη2	Cδ-Nε-Cζ-Nη1 and Nη2
Arg 84	Nε-Cζ-Nη1 and Nη2	Cδ-Nε-Cζ-Nη1 and Nη2
Glu 87	Cγ-Cδ-Oε1 and Oε2	Cγ-Cδ-Oε1 and Oε2
Glu 90	Cβ-Cγ-Cδ-Oε1 and Oε2	
Arg 103	-	Cγ-Cδ-Nε-Cζ-Nη1 and Nη2
Arg 104	Cγ-Cδ-Nε-Cζ-Nη1 and Nη2	Cγ-Cδ-Nε-Cζ-Nη1 and Nη2
Arg 106	Cγ-Cδ-Nε-Cζ-Nη1 and Nη2	-
Lys 134	Cε-Nζ	Cγ-Cδ-Cε-Nζ

Table 3.3 – Table listing omitted and truncated amino acids.

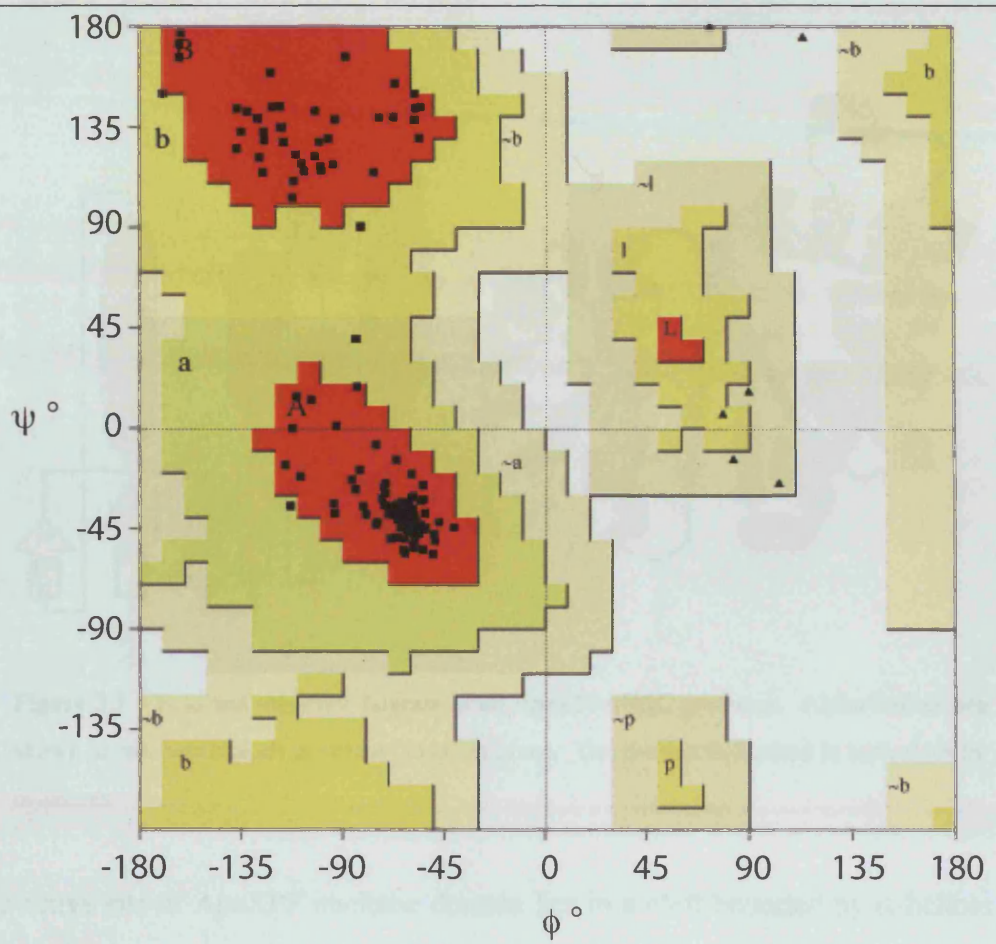

Figure 3.6 – Ramachandran plot for the ApeXPF nuclease domain model. The plot was calculated using the program PROCHECK (Laskowski et al., 1993) using the final refined co-ordinates of the ApeXPF nuclease structure. Glycine residues are represented as black triangles.

3.3 Discussion of the structure of ApeXPF nuclease domain.

The structure of the ApeXPF nuclease domain consists of an α/β fold with a six-stranded β-sheet flanked on both sides by α-helices. Five of the six β-strands run parallel to one another. A topology diagram is shown in figure 3.7.

The structure of the ApeXPF nuclease domain is closely related to the *Pyrococcus furiosus* Hef nuclease domain. Like Hef, the nuclease domain of ApeXPF adopts the type II restriction endonuclease fold found in the majority of restriction enzymes, as well as Hjc, an archaeal Holliday junction resolvase from *Pyrococcus furiosus* (Nishino et al., 2003).

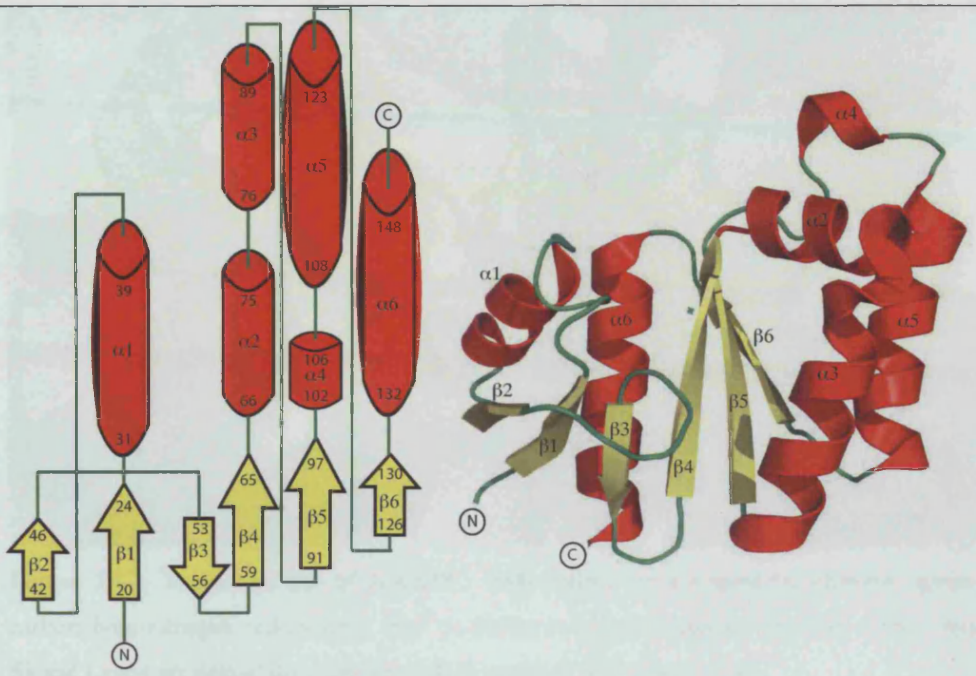

Figure 3.7 – Fold and topology diagram of an ApeXPFΔNΔC protomer. Alpha helices are shown in red, beta strands in yellow, loops in green. The β-sheet is flanked in both sides by α-helices.

The active site of ApeXPF nuclease domain lies in a cleft bounded by α-helices α1' (α1 from the other protomer), α2 and α3 and the loop linking strands β2 and β3. This cleft is lined with many residues conserved in XPF homologues, including those involved in metal binding and catalysis (figure 3.8). Residues within this cleft include those from the GDX_nERKX_3D motif discussed in Chapter 1. The first aspartic acid residue in this motif, Asp52, lies on a loop connecting β2 and β3. Glu62, Arg63, and Lys64 all lie on strand β4. Asp68 is not directly around the Mg2+ ion identified but is involved in co-ordination of a water molecule. Gln81 reaches into the active site from α-helix α4. The active site is shown in figure 3.9.

Although the active site contains only one Mg^{2+} ion, this does not necessarily mean that catalysis is achieved with a one metal ion mechanism. It is often the case that an enzyme's full complement of divalent ions is only observed when substrate is bound. The majority of DNA processing enzymes use two metal ion mechanisms, but it is possible in this case that the presence of Arg63 in the active site stabilises an activated nucleophilic water, thus negating the need for a second metal ion. Future studies with substrate-bound would aid analysis of the catalytic mechanism.

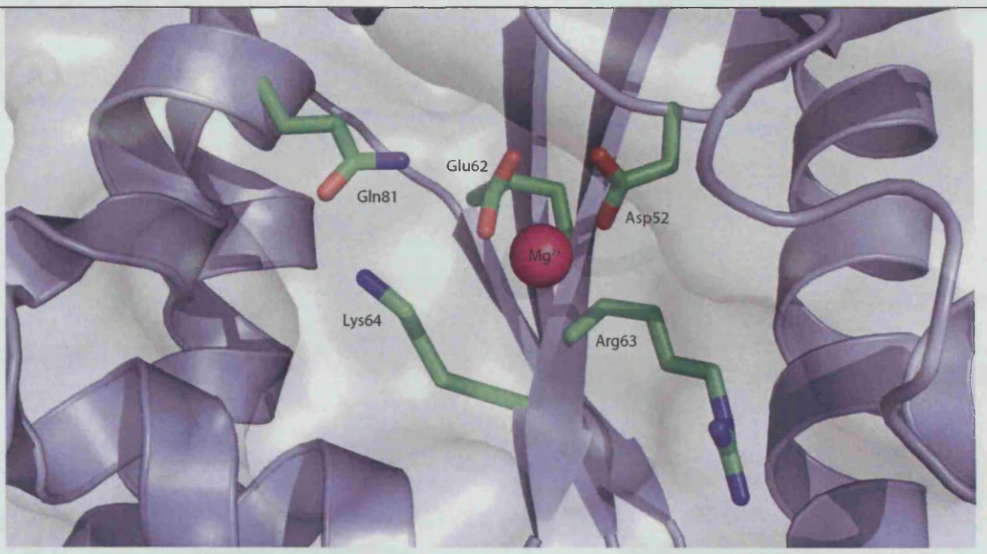

Figure 3.8 – The active site of ApeXPF. Side chains are coloured by element; green-carbon, blue-nitrogen, red-oxygen. Mg^{2+} is shown as magenta sphere. Residues Glu62, Arg 63 and Lys64 are part of the signature E-R-K motif of XPF homologues.

The crystal structure, as expected, shows that the nuclease domain of ApeXPF forms a homodimer. The dimer interface of the nuclease domain of the ApeXPF nuclease domain is again similar to that of the Hef nuclease domain. The dimer interface consists mainly of C-terminal residues contributed by two α-helices (α6 and α7) and a β-sheet (β6). Many residues involved in the dimer interfaces are hydrophobic in nature (Y113, A117, L121, M128, M131, L138, L146) but there are residues involved in polar interactions as well (Q120, R126, N129, D132, K134, E141, S142, R145). Figure 3.9 (a and b) shows the extent of the ApeXPF nuclease dimer interface.

There is however, a notable difference between the dimer interfaces of ApeXPF and Hef. The difference is at the position occupied by R126 in ApeXPF. Sequence alignments of ApeXPF and other XPF homologues (figure 3.9 c) show that Hef does not have a conserved arginine at this site but has a proline residue instead. The arginine is notable as sequence alignments predict this residue to be equivalent to R788 in human XPF. A patient with a mild case of XP with late onset carcinomas and neurological disease was shown to be homozygous for a point mutation that changed R788 to tryptophan (Sijbers et al., 1998). The substitution of arginine for a bulky tryptophan residue in the XPF-ERCC1 dimer interface may explain the phenotype of this patient.

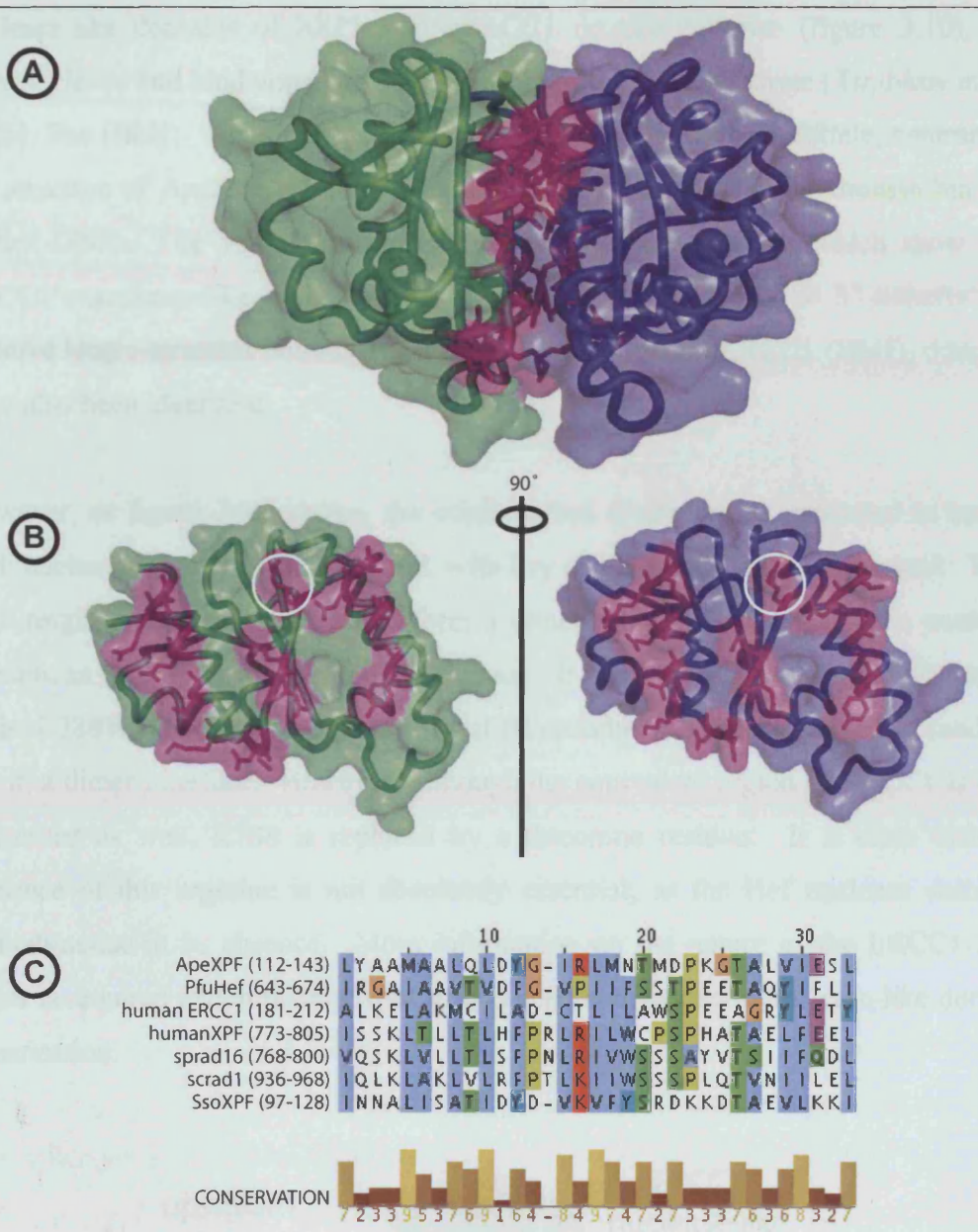

Figure 3.9 - The ApeXPF (18-150) nuclease dimer interface. [A] Surface and cartoon representation of the ApeXPF nuclease domain dimer. Residues contributing to the dimer interface are coloured magenta. [B] The dimer interface. Protomers have been rotated through 90 °. Again residues contributing to the dimer interface are coloured magenta. These residues are found in a "S"-shape made by α6-β6-α7. Residue R126 in each protomer is circled [C] Alignment of XPF homologues in the region of the ApeXPF dimer interface. Note that ApeXPF and the eukaryotic XPF homologues share the same conserved arginine residue discussed in the text.

This observation has a bearing on current opinion as to how XPF/Mus81 endonucleases engage their substrates. Tom Ellenberger's group at Harvard University have proposed a model for human XPF action where the nuclease and

nuclease-like domains of hXPF and hERCC1 do not dimerise (figure 3.10), but instead cleave and bind opposing single-strands of a bubble substrate (Tsodikov et al., 2005). The $(HhH)_2$ domains bind opposing single-strands of the substrate, contrary to the structure of ApeXPF with DNA-bound which shows an $(HhH)_2$ domain binding duplex DNA. The model is supported by DNA-binding studies which show that ERCC1's nuclease-like domain binds DNA of the suggested 3' to 5' polarity and putative single-stranded binding surfaces on the hXPF and hERCC1 $(HhH)_2$ domains have also been identified.

However, as figure 3.9C shows, the conservation at the region predicted to be the XPF nuclease dimer interface is good, with key residues such as R788 present. Thus XPF might indeed be predicted to form a dimer with ERCC1 through its nuclease domain, as well as through its $(HhH)_2$ domain. It is difficult to explain the phenotype of the R788W mutation discussed above, if this residue is neither involved in catalysis nor in a dimer interface. However, although the equivalent region of ERCC1 is well conserved as well, R788 is replaced by a threonine residue. It is clear that the presence of this arginine is not absolutely essential, as the Hef nuclease domains homodimerise in its absence. More information on the nature of the ERCC1-XPF dimer is required to completely resolve the question of nuclease/nuclease-like domain dimerisation.

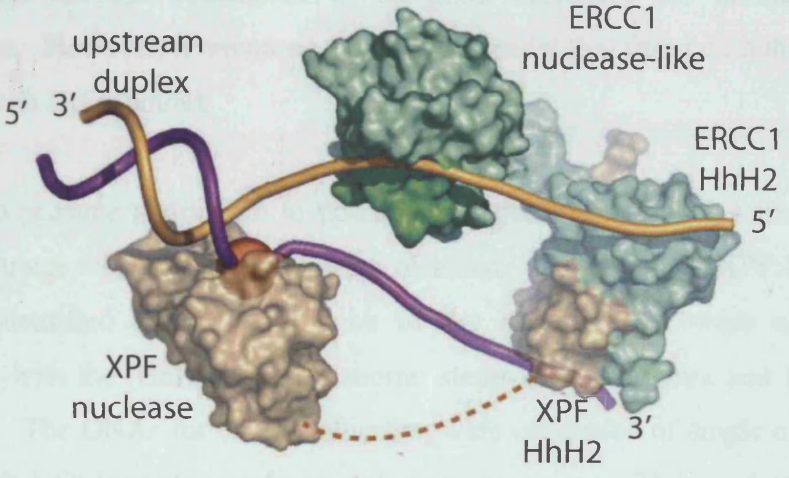

Figure 3.10 – A model suggested for the action of human XPF/ERCC1. (adapted from Tsodikov et al., 2005).

3.4 Investigating the mechanism of ApeXPF action: Model 1

The structure of ApeXPF complexed with 15-mer duplex DNA (Newman et al., 2005), gave some insights into DNA binding and substrate recognition by ApeXPF. The binding of DNA via a minor groove engaged both HhH motifs, each contacting a phosphodiester backbone on a different strand. This suggested how the enzyme may position a substrate the correct distance away from its double-strand/single-strand junction. However, the structure did not contain DNA in the active site and models about how DNA extended into the active site were proposed for validation.

The simplest model involved a single-stranded overhang threading its way into the active site, as depicted in figure 3.11. In this scheme, the duplex DNA observed in the crystal structure would act as the upstream duplex of a substrate. At a double-strand/single-strand junction, the single-strand would thread its way into the active site. Any downstream DNA would presumably pass around the enzyme. This model will be referred to as "Model 1". There are some obvious caveats to this model. Firstly, one of the $(HhH)_2$ domain is not used in substrate recognition for no obvious reason. Indeed footprinting experiments on Hef, have showed that both $(HhH)_2$ domains are thought to bind duplex DNA (Nishino et al., 2005). Also, the model does not fit with with kinetic data from Malcolm White's group in St. Andrews. They showed that the preferred substrates for another crenarchael XPF from *Sulfolobus solfataricus*, SsoXPF, were nicked duplexes and 3'-flaps (Roberts and White, 2005). 3'-overhangs are not considered to be good substrates for crenarchaeal XPF homologues. However, it remained the simplest model that fitted with the structure of ApeXPF with DNA-bound.

In order to produce a structure to potentially validate this model, various synthetic DNA substrates with 3'-overhangs were obtained. Since the ApeXPF:15-mer DNA structure identified only 8 base pairs of the duplex DNA were necessary for interaction with the $(HhH)_2$ domain, shorter stems of 8 base pairs and 10 base pairs were used. The DNAs for co-crystallisation were composed of single oligonucletide strands with hairpins between the complementary regions. The overhangs consisted of 2, 4, 6 and 8 thymine bases (figure 3.13).

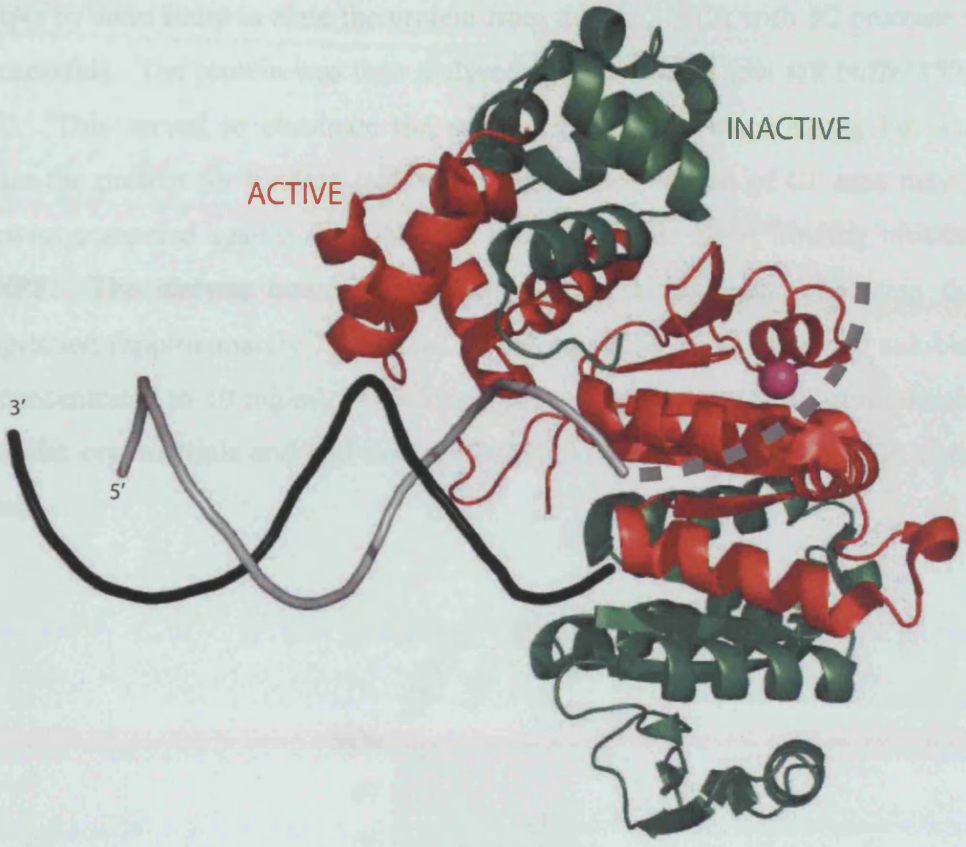

Figure 3.11 – Model 1. The structure of ApeXPFΔNΔC complexed with 15-mer DNA is shown. The active protomer is coloured red, the inactive protomer, green. The possible path of a single stranded 3'-overhang to the active site is shown as a dashed line. The magenta sphere is the Mg^{2+} in the active site. Adapted from PDB file 2BGW.

3.5 Expression of ApeXPFΔNΔC protein

The pET14b-3C ApeXPFΔNΔC (18-231) construct created by Matthew Newman was transformed into rosetta pLysS cells. The protocol used is described in Lally et al. (Lally et al., 2004). Bacteria were grown, induced and harvested according to the methods outlined in Appendix A. Protein was extracted in a Tris buffer containing 500 mM NaCl. The high salt concentration was necessary for the extraction of a high yield of soluble protein. A dependence on high salt buffers is not uncommon amongst DNA-binding proteins (Yang; 2004) and there are examples of proteins such as MarA (Jair et al., 1995) that require 1 M NaCl for solubility in the absence of DNA. The ApeXPFΔNΔC protein was purified by affinity chromatography on Ni^{2+}-NTA and eluted from the resin using imidazole. After overnight dialysis to remove

the imidazole, 3C protease was used to remove the His_6-tag (figure 3.12; previous attempts by John Lally to elute the protein from the Ni^{2+}-NTA with 3C protease were unsuccessful). The protein was then dialysed again against a low salt buffer (50 mM NaCl). This served to eliminate the small peptide with the His_6-tag but also to prepare the protein for binding to DNA. High concentration of Cl^- ions may have otherwise competed against the DNA for access to basic DNA binding residues on ApeXPF. The dialysis however resulted in a substantial loss of protein due to precipitation (approximately 75 % loss). The remaining 25 % remained soluble and was concentrated to 10 mg/ml. This loss, whilst disappointing, had been tolerated in the earlier crystal trials and had not prohibited the growth of protein:DNA complex crystals.

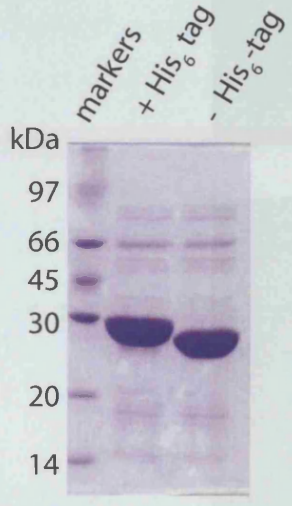

Figure 3.12 – ApeXPFΔNΔC protein expression and purification. Purification followed published methods (Lally et al., 2004). The figure shows the removal of the His_6-tag by 3C protease.

3.6 Co-crystallisation of ApeXPFΔNΔC protein with model 1 substrates

Crystallisation trials were set-up using ApeXPFΔNΔC protein and the new 3'-overhang substrates with a protein to DNA ratio of 1:1.2. Vapour diffusion and microbatch drops were set-up using commercial sparse matrix and grid screens.

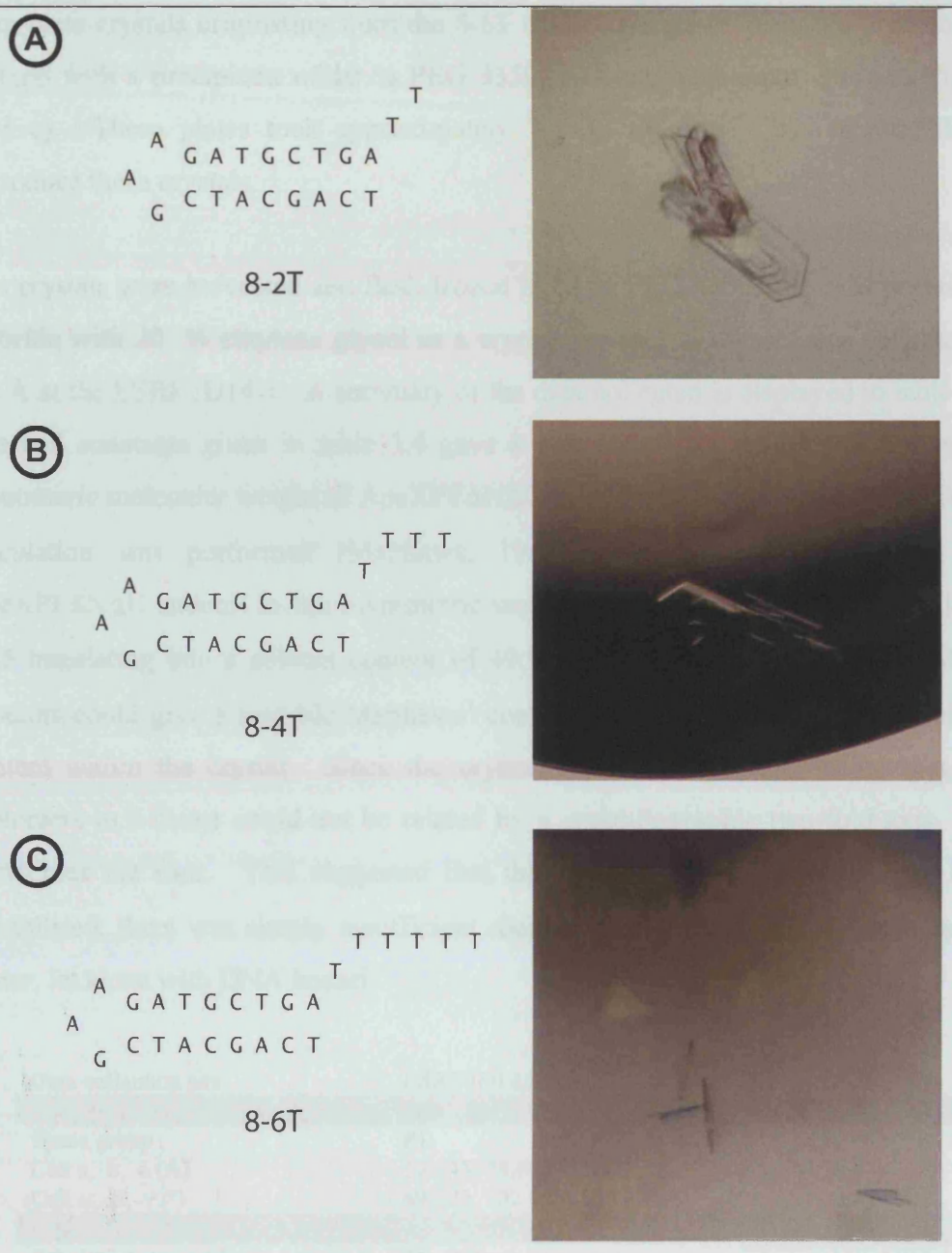

Figure 3.13 – Synthetic DNA oligonucleotides with overhangs and crystals produced when co-crystallised with ApeXPFΔNΔC. [A] 8-2T. [B] 8-4T. [C] 8-6T. Only the crystals with the 8-6T DNA were reproducible.

Crystals were obtained for three of the four substrates. Small crystals were obtained in hanging drops with the 8-2T and 8-4T substrates. The precipitant in both these cases was 200 mM magnesium chloride, 0.1 M Tris (pH 8.5) and 20 % PEG 8000. As figure 3.13 shows these were not single crystals but rather plates grown on top of one another. These crystals proved to be difficult to reproduce.

Thin plate crystals originating from the 8-6T trials were grown using the microbatch method with a precipitant of 10 % PEG 3350, 100 mM potassium chloride (figure 3.13 c). These plates took approximately 7 days to grow. It was possible to reproduce these crystals.

The crystals were harvested and flash-frozen in 12 % PEG 3350, 100 mM potassium chloride with 20 % ethylene glycol as a cryoprotectant. A dataset was collected to 2.2 Å at the ESRF ID14-1. A summary of the data collected is displayed in table 3.4. The cell constants given in table 3.4 gave a cell volume of 62989 $Å^3$. Using the monomeric molecular weight of ApeXPFΔNΔC as 25631 Da, a Matthews' coefficient calculation was performed (Matthews, 1968). This indicated that for one ApeXPFΔNΔC subunit in the asymmetric unit, the Matthews' coefficient would be 2.45 translating into a solvent content of 49.9 %. No other number of ApeΔNΔC subunits could give a sensible Matthews' coefficient and thus a permissible solvent content within the crystal. Since the crystals were triclinic, this meant that two protomers in a dimer could not be related by a crystallographic two-fold axis, as is sometimes the case. This suggested that the ApeXPFΔNΔC dimer had not been crystallised; there was simply insufficient room within the unit cell to house such a dimer, let alone with DNA bound.

Data collection site	ESRF ID14 EH1
Space group	P1
Cell a, b, c (Å)	33.643, 38.683, 55.321
Cell α, β, γ (°)	89.073, 102.900, 115.685
Resolution range (Å)	34 - 2.2
Wavelength (Å)	0.9791
Total number of reflections	41 926
Unique Reflections	11 581
<I/σI>	4.5 (1.5*)
Completeness (%)	96.4 (97.1*)
R_{sym}	0.103 (0.283*)
Mosaic spread (°)	0.7

Table 3.4 – Data processing statistics for crystals of ApeXPFΔNΔC co-crystallised with 8-6T DNA. Figures in parentheses refer to the highest resolution shell.

At this point, a similarity between the cell parameters of this dataset and the ApeXPF nuclease domain dataset discussed in section 3.2 became obvious. The equivalent cell constants and space group suggested that perhaps these crystals were in fact ApeXPF

nuclease domain crystals again. After integration, scaling and conversion to structure factors using Mosflm, Scala and Truncate respectively (Collaborative Computational Project, 1994; Leslie, 1992), molecular replacement using Molrep was attempted using the refined structure of the ApeXPF nuclease domain. The molecular replacement was successful and the electron density housed the ApeXPF nuclease domain. No extra density for DNA was evident.

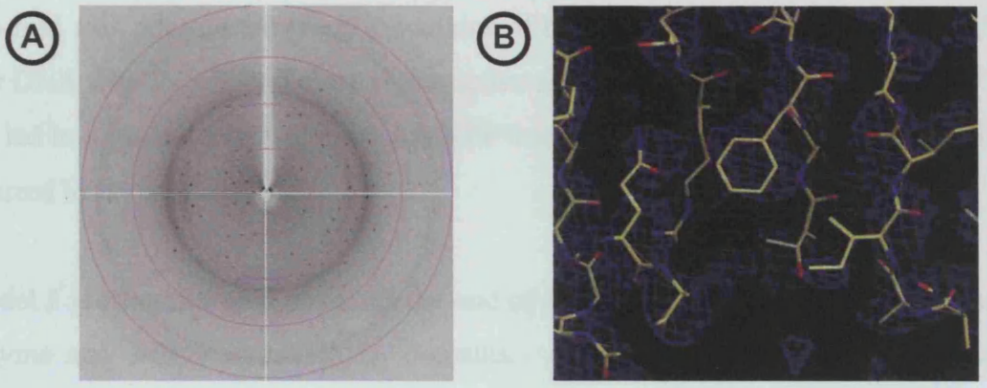

Figure 3.14 – ApeXPFΔNΔC co-crystallised with 8-6T DNA. [A] Diffraction from ESRF ID14-1. Circles are at 2.1, 2.8, 4.1 and 8.3 Å. [B] Initial electron density showing that the molecular replacement solution using the refined ApeXPF nuclease domain appears to be correct.

Excess crystals were harvested, washed, dissolved in SDS-PAGE sample buffer and ran on a SDS-PAGE gel. From the gel it was clear that proteolysis had occurred in the drop, meaning the prevalent protein species was the ApeXPF nuclease domain. The exact cause of the proteolysis remains unclear. Slightly puzzling is that fact that the preparation of ApeXPFΔNΔC for the solved complex (with 15-mer DNA) was essentially identical, yet the protein within those crystals was not degraded. One obvious difference is in the crystallisation method. For the ApeXPFΔNΔC:15-mer DNA complex, sitting-drop vapour diffusion was used. Also the precipitant conditions are very different. The condition found previously used was 100 mM sodium acetate (pH 4.6), 200 mM $(NH_4)_2SO_4$, 20 % PEG MME 2000. The low pH in this drop may have prohibited proteolysis in these drops. The neutral pH provided by 20 % PEG 3350, 0.1 M KCl, is likely to fall within the optimal range of a great number of proteases.

3.7 Investigating the mechanism of ApeXPF action: Model 2

The absence of DNA in these crystals was a setback. However, further refinement of the ApeXPFΔNΔC structure by Judith Murray-Rust suggested a new approach should be taken. The refinement identified sulphate ions near the DNA-free $(HhH)_2$ domain at positions equivalent to the phosphate backbone contacts with the DNA-binding $(HhH)_2$ domain (figure 3.15 a). It is possible that these sulphate ions mimic the phosphodiester backbone of DNA. Rotation of the observed 15-mer DNA about the two-fold axis relating the $(HhH)_2$ domains of both protomers generates a second 15-mer DNA with its a 3' end close to the active site of the enzyme (figure 3.15 b). This has led to a second model of how ApeXPF interacts with its substrates, which will be referred to as "Model 2".

Model 2 is attractive as it positions the end of the DNA closer to the active site of the enzyme and utilises both $(HhH)_2$ domains. An obvious caveat is that the second $(HhH)_2$ domain does not bind DNA in the crystal structure. This can be explained by analysis of crystal packing in which it is evident that a second 15-mer DNA would have clashed with a crystallographic axis, sterically inhibiting crystal growth. This model was also in good agreement with kinetic data from Malcolm White's group in St. Andrews. They showed that the preferred substrates for SsoXPF, were nicked duplexes and 3'-flaps both of which contained upstream and downstream duplexes (Roberts and White, 2005). Model 2 accounts for binding of both upstream and downstream duplexes in a way Model 1 did not. Model 2 is also in agreement with a model later proposed by Kosuke Morikawa's group in Osaka. They also proposed a model for *P.furiosus* Hef action in which both $(HhH)_2$ domains bind duplex DNA (backed up by foot-printing data) and although the nuclease domains independently dimerise, only one nuclease domain is active (fig 3.14 c). Attempts to crystallise ApeXPFΔNΔC with synthetic DNA substrates mimicking 3'-flaps according to Model 2 were therefore undertaken.

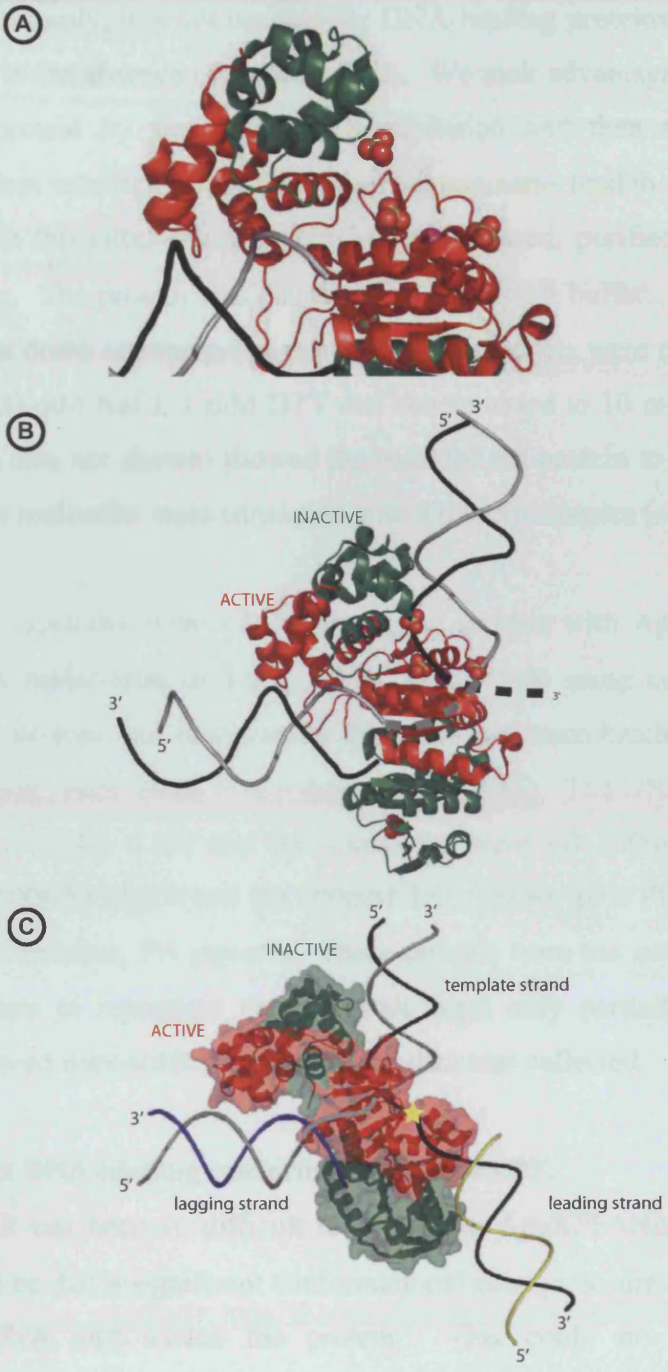

Figure 3.15 – Model 2. [A] Further refinement of the ApeXPFΔNΔC-DNA structure showed sulphate ions at key positions near the $(HhH)_2$ domain and near the active site. It is likely that these sulphate ions mimic a phosphodiester DNA backbone. [B] Model 2. Again the active subunit is coloured red, the inactive subunit, green. An upstream duplex is generated by rotating the observed duplex 90 ° around the two-fold axis relating the $(HhH)_2$ domains. Possible path for a continuous strand and the cleaved strand are shown as dashed lines. Adapted from PDB file 2BGW. [C] Model for *P.furiosus* Hef action. Although a stalled replication fork is modelled, the basic features are similar. Adapted from Nishino et al. 2005b).

As discussed previously, it is not unusual for DNA-binding proteins to precipitate in low salt buffers in the absence of cognate DNA. We took advantage of this property to purify the protein by first selective precipitation and then resolublising the precipitated protein in a high salt buffer. Most contaminants tend to remain soluble in the first step. In this case, ApeXPFΔNΔC was expressed, purified, eluted and de-tagged as before. The protein was dialysed into a low-salt buffer. The precipitated protein was spun down and the pellet retained. These pellets were resolublised in 20 mM Tris (8), 300 mM NaCl, 1 mM DTT and concentrated to 10 mg ml^{-1}. Dynamic light scattering (data not shown) showed the resolublised protein to be monodisperse with an apparent molecular mass consistent with a dimeric species (~ 52 kDa).

Different DNA substrates were trialled in crystal screens with ApeXPFΔNΔC at a protein to DNA molar ratio of 1:1.2. Trials were set-up using commercial sparse matrix and grid screens and both vapour diffusion and microbatch methods. After three weeks crystals were found in two drops (figure 3.16). The DNA substrate was a 3'-flap substrate (figure 3.16) and the conditions were (a) 100mM MES (pH 6), 100mM NaCl, 100mM $MgCl_2$ and (b) 100mM Tris (pH 8), 10% PEG 8000, 100mM $MgCl_2$, 2mM spermidine, 5% glycerol. These crystals were too small for diffraction studies. Attempts to reproduce these crystals were only partially successful and optimisation proved impossible. No diffraction data was collected.

3.7 Promoting a DNA binding conformation of ApeXPF.

A reason why it has been so difficult to crystallise ApeXPFΔNΔC with synthetic substrates could be that a significant conformational change occurs on DNA binding, both in the DNA and indeed the protein. This could be why crystals of ApeXPFΔNΔC with DNA-bound took over 3 months to grow. Comparisons of the apo-ApeXPFΔNΔC structure with the structure of ApeXPFΔNΔC once 15-mer DNA was bound showed a large movement (~ 30 Å) of the $(HhH)_2$ domains towards the nuclease domains on DNA-binding (figure 3.17). Given ApeXPF is optimised for working at high temperatures (~95 °C), this conformational change might be easier to promote at such temperatures than at room temperature. The role of the sliding clamp PCNA in promoting the activation of ApeXPF could promote a conformation competent to engage the substrate fully. Malcolm White's group have suggested that

PCNA affects the kinetic constant k_{cat} rather than K_m when stimulating XPF activity (Malcolm White, unpublished result). This is consistent with an activation event as opposed to enhancement of ApeXPF's affinity for its substrate. This idea is explored further in Chapter 4. However, at this point a way of tethering the nuclease domain and $(HhH)_2$ domain together in a DNA-binding conformation was considered.

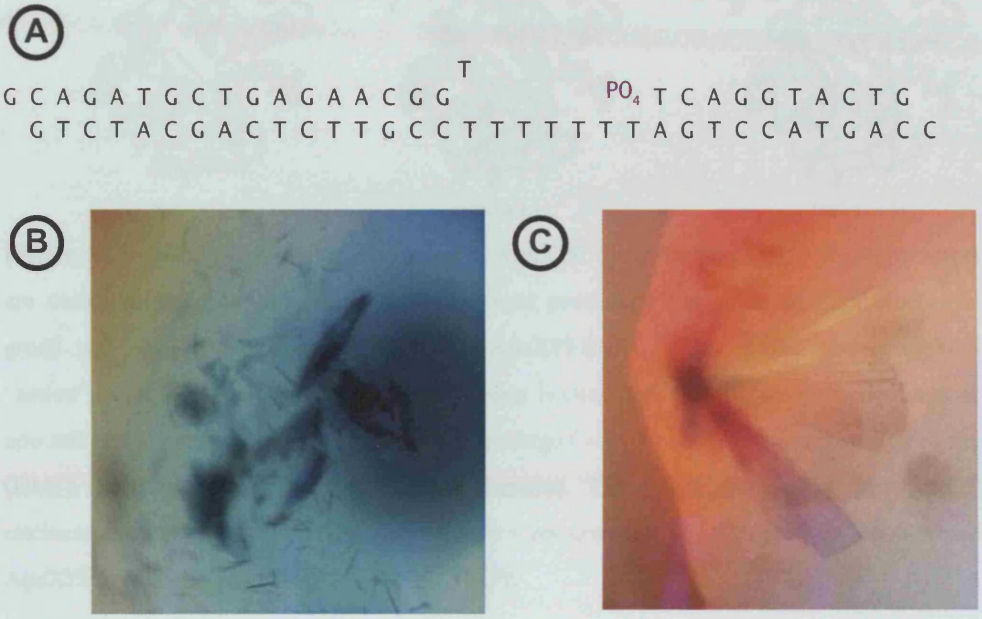

Figure 3.16 – Synthetic DNA oligonucleotide with upstream and downstream duplex and crystals produced when co-crystallised with ApeXPFΔNΔC. [A] The oligonucleotide that yielded crystals. Initial crystallisation hits. [B] Precipitant was 100mM MES (pH 6), 100mM NaCl, 100mM $MgCl_2$. [C] Precipitant was 100mM Tris (pH 8), 10% PEG 8000, 100mM $MgCl_2$, 2 mM spermidine, 5% glycerol.

In the apo structure of ApeXPFΔNΔC, amino acid residues lysine 46 (in the nuclease domain) and aspartic acid 165 (in the $(HhH)_2$ domain) of the "active" protomer are far apart (~ 33 Å). In the DNA bound form of ApeXPFΔNΔC, lysine 46 and aspartic acid 165 are close together and form a salt bridge (figure 3.18). Assuming that the interdomain contacts seen in the ApeXPFΔNΔC-DNA structure are complete, it was hoped that if this salt bridge could turned into a permanent covalent bond such as a disulphide bridge, then the ($(HhH)_2$ domain) could be tethered to the nuclease domain to mimic an activated conformation.

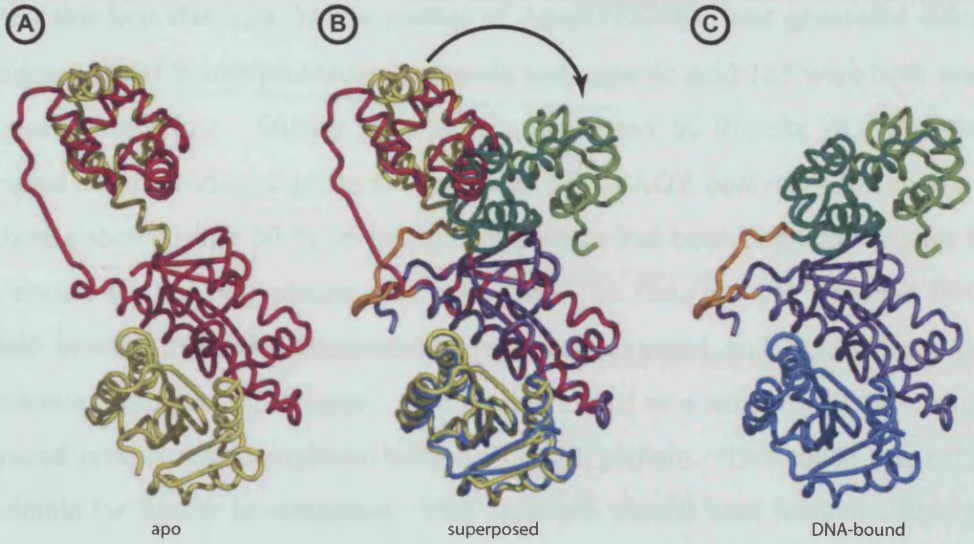

Figure 3.17 – The conformational change in ApeXPF upon binding DNA. Apo-protomers are coloured magenta and yellow. DNA-bound protomers are coloured dark-blue/dark-green and light-blue/light-green. [A] apo-ApeXPFΔNΔC. The linker connecting the "active" nuclease domain and its $(HhH)_2$ domain is extended. [B] Superposed structures of apo and DNA-bound ApeXPFΔNΔC. Note the large (~ 30 Å) conformational change of the $(HhH)_2$ domains relative to the nuclease domains. The linker connecting the "active" nuclease domain and its $(HhH)_2$ domain is now more compact. [C] Structure of DNA-bound ApeXPFΔNΔC. Adapted from 2BHN and 2BGW.

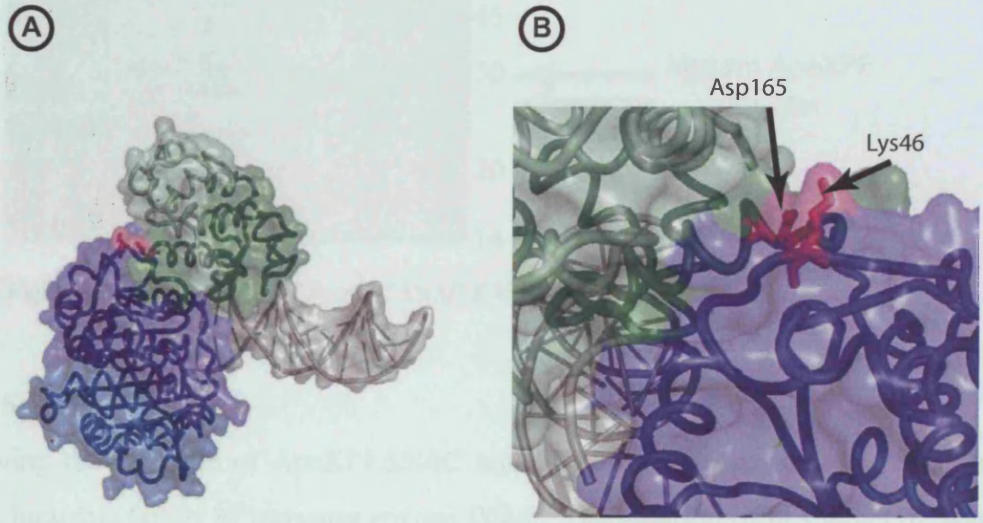

Figure 3.18 – Salt bridge across the nuclease – $(HhH)_2$ interface in the ApeXPFΔNΔC – DNA complex. [A] Nuclease domains are shown in blue and cyan, the $(HhH)_2$ domains in dark green and green, DNA in is grey. Residues K46 and D165 are shown in magenta. [B] Close-up of residues K46 and D165. Adapted from PDB file 2BGW.

To test this hypothesis, a double mutant of ApeXPFΔNΔC was generated using the Stratagene QuikChange protocol. Lysine 46 and aspartic acid 165 were both mutated to cysteine residues. Mutant protein was expressed in Rosetta pLysS cells and extracted as for wild-type protein. However, SDS-PAGE analysis after elution with imidazole showed that 50 % of the mutant protein had been degraded (figure 3.19). The reason for this degradation was not clear. The His_6-tag was removed from the mutant protein using 3C protease and was concentrated and loaded onto a pre-equilibrated Superdex 75 column. The protein eluted as a single peak indicating that degraded protein was complexed with full-length protein. This made this approach unsuitable for further investigation. This approach should have worked, although the effects of protein engineering experiments are notoriously difficult to predict.

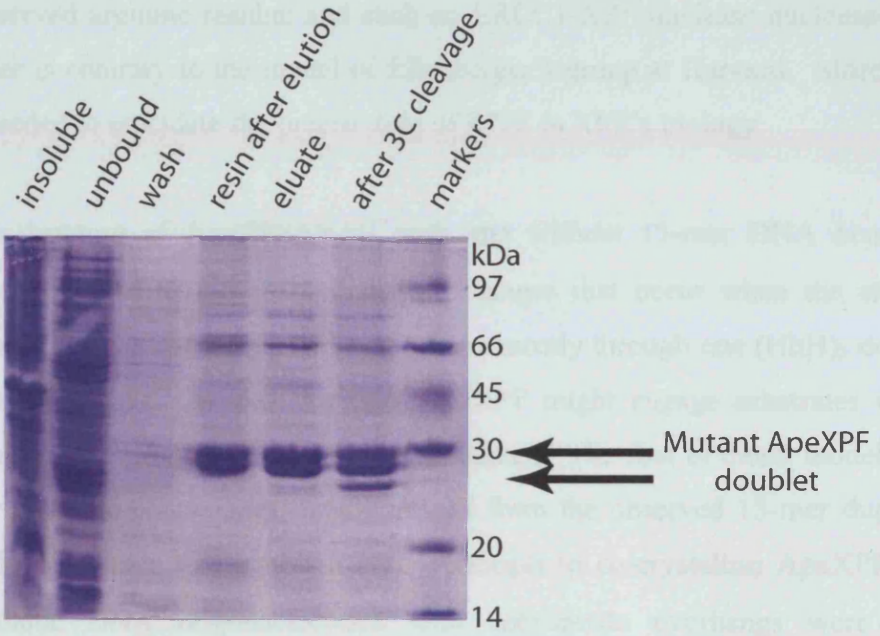

Figure 3.19 – Expression of ApeXPFΔNΔC K46C D165C mutant protein.

3.8 Summary

Solving the structure of ApeXPFΔNΔC with DNA bound provided the first insight into how this family of enzymes engage DNA. The solution of this structure was not easy and required a combination of phases from heavy atom derivatives and molecular replacement using data from ApeXPF nuclease domain solution. Using the data for the ApeXPF nuclease domain was a useful "training exercise" in building and

refinement. It also had the additional benefits of being at higher resolution. The final model has acceptable R-factors and geometry.

The overall structure of the ApeXPF nuclease is closely related to that of the *Pyrococcus furiosus* Hef nuclease domain. The overall fold and the dimer interface are all closely related. The residues of the GDX_nERKX_3D motif common to all XPF/Mus81 homologues are mostly found in the active site. However, there is a key difference at the dimer interface. Residue R126 is not conserved in Hef but it is present in eukaryotic XPF orthologues. The equivalent residue in human XPF, R788 was mutated to tryptophan in a patient with late-onset XP. This suggests that this patient's phenotype was caused by distortions in the dimer interface between XPF and ERCC1's nuclease and nuclease-like domains. However, ERCC1 does not have this conserved arginine residue and such an ERCC1-XPF nuclease-nuclease-like domain dimer is contrary to the model of Ellenberger's group at Harvard. More information is needed to elucidate the precise role of R788 in XPF's biology.

The structure of ApeXPFΔNΔC with and without 15-mer DNA bound provided information as to the conformational changes that occur when the enzyme binds DNA. The actual binding of DNA occurs mostly through one $(HhH)_2$ domain across a minor groove. Models for how ApeXPF might engage substrates with double-strand/single-strand junctions were formulated. The first of these, model 1, proposed that a simple 3'-overhang would extend from the observed 15-mer duplex into the nuclease domain to the active site. Attempts to co-crystallise ApeXPFΔNΔC with synthetic DNA oligonucleotides with appropriate overhangs were undertaken. Crystals grown using the microbatch method with PEG 3350 as a precipitant were promising. However, the processing of diffraction data collected at the ESRF showed that these crystals were of the ApeXPF nuclease domain with no DNA bound.

A second, more probable model, model 2, was proposed. Sulphate ions near the second (free) $(HhH)_2$ domain were thought to mimic a phosphodiester backbone. The modelling of an equivalent 15-mer duplex to the second $(HhH)_2$ domain provided a means by which XPF engages both upstream and downstream duplexes, generating a

~ 90 ° kink in the DNA substrate. This is consistent with the preferred substrates of SsoXPF (another crenarchaeal XPF), namely 3'-flaps and nicked duplexes.

Attempts to crystallise ApeXPFΔNΔC with more sophisticated DNA substrates designed with model 2 in mind have been largely unsuccessful. Despite two promising leads it has not been possible to produce crystals suitable for diffraction studies. An attempt to tether ApeXPFΔNΔC in a DNA-binding conformation by introducing a disulphide bond between nuclease and $(HhH)_2$ domain residues has been blighted by proteolysis. Experiments investigating the effect of ApePCNA on stimulating ApeXPF are detailed in Chapter 4.

4 Preparation and characterisation of *Aeropyrum pernix* XPF-PCNA complexes.

This chapter describes experiments relating to the preparation and characterisation of *Aeropyrum pernix* XPF-PCNA complexes. A flowchart describing these experiments can be found below (Figure 4.1).

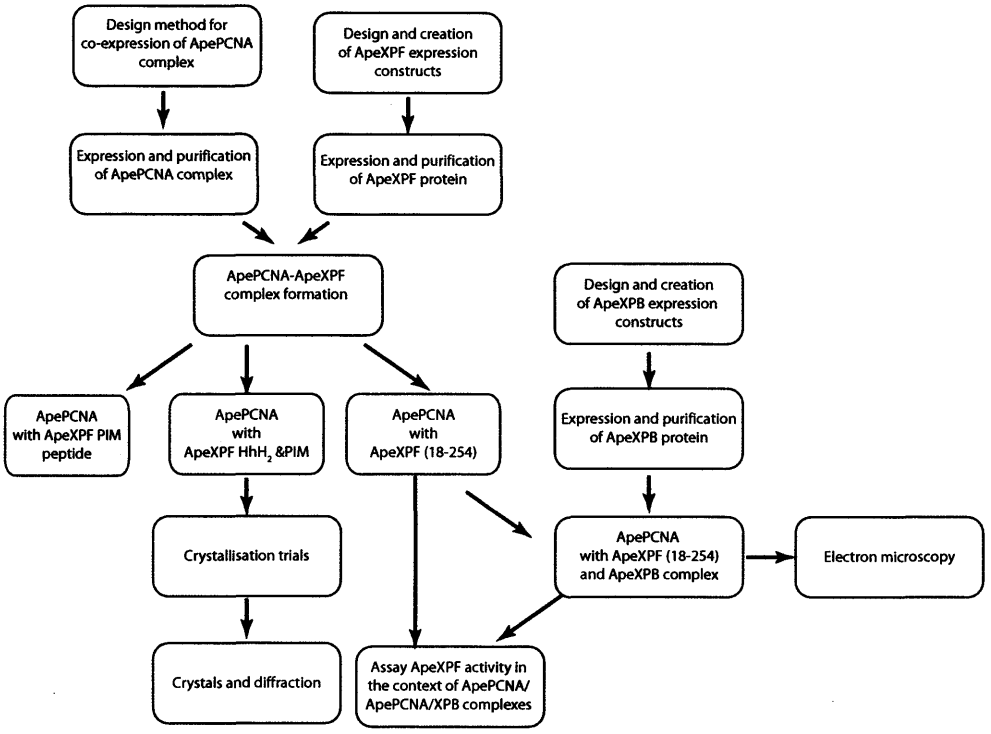

Figure 4.1 – Flowchart of the experiments described in chapter 4.

4.1 General considerations

The models of archaeal XPF action proposed by Newman et al. 2005 and Nishino et al. 2005 discussed in chapter 3 are different to that proposed by Tsodikov et al., 2005. for eukaryotic XPF action. One of the limitations of such models is that the structural evidence that they are based on is only a snapshot of a small part of a more complex machine. The effects of the other parts of this machine cannot be "second-guessed".

With respect to crenarchael XPFs such as the XPF homologues from *Aeropyrum pernix* and *Sulfolobus solfataricus*, it has been shown that the presence of PCNA is essential for efficient XPF activity (Roberts et al., 2003). The presence of XPB and XPD homologues in these genomes (Kelman and White, 2005), suggests that a

unified model of NER/replication fork repair in crenarchaea would need to incorporate the roles of all these proteins.

It has been noted from the kinetics of SsoXPF that the heterotrimeric SsoPCNA is not just responsible for guiding the XPF to the appropriate site in damaged DNA but actually appears to activate the SsoXPF. This manifests itself as a change in the kinetic constant k_{cat}, rather than K_m (Malcolm White, unpublished observation). The structural basis of such an activation event is of some interest and was the question we sought to address by these studies.

The proliferating cell nuclear antigen (PCNA) is a ring-shaped sliding clamp protein. Such sliding clamps are found in all organisms (Bruck and O'Donnell, 2001). In bacteria, they are referred to as β-clamps, which is in reference to their role as the β-subunit of DNA polymerase III. Both PCNA and β-clamp form a ring structure around DNA, composed of six domains. In PCNA the ring is compose of three two-domain subunits. In bacterial β-clamps the ring is composed of two three-domain subunits. Structures of bacterial β-clamps, eukaryotic and euryarchaeal PCNAs are very similar (figure 4.2).

The three PCNA homologues in *Aeropyrum pernix* had previously been identified and characterised by their interaction with DNA polymerases (Daimon et al., 2002). This study considered that the three subunits might each form homotrimeric processivity factors for the DNA polymerases in *Aeropyrum pernix* cells under different conditions. The study by Malcolm White's group in 2003 (Roberts et al., 2003) reporting that the XPF from *Sulfolobus solfataricus* is dependent on a heterotrimeric PCNA means it is likely that when considering NER, the three PCNA homologues present in *Aeropyrum pernix* could also associate into a heterotrimer. In the study of Roberts et al., it was found that SsoXPF interacts with two SsoPCNA subunits but not the third. It seems reasonable to assume that this would be case for an equivalent *Aeropyrum pernix* complex. However, phylogenetic comparisons of SsoPCNA and ApePCNA subunits do not readily identify which PCNA subunits are functional orthologues. Instead they show that PCNA subunits from within each genome are closely related to one another.

Proteins that bind PCNA are usually involved in various aspects of DNA replication and processing, using PCNA's sliding-clamp properties. It has become clear that a great many PCNA binding proteins contain a conserved PCNA-binding motif (PIM) (Warbrick, 2000). The interactions of some of the proteins listed in Table 4.1 have been extensively studied and the interaction of PIM and PCNA examined by structural methods. These studies have led to the establishment of consensus PIM sequences. Perhaps the most essential feature of a PIM is the presence of two aromatic amino acid residues adjacent in the polypeptide to one another.

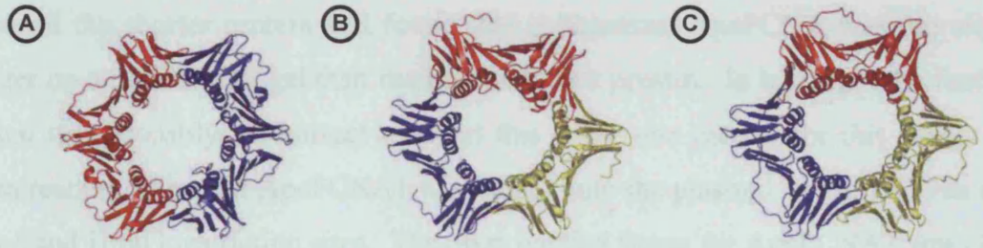

Figure 4.2 – Structures of sliding clamps. [A] *E.coli* β-clamp (PDB code 2POL (Kong et al., 1992)). [B] *Pyrococcus furiosus* PCNA (PDB code 1GE8 (Matsumiya et al., 2001)). [C] Human PCNA (PDB code 1AXC (Gulbis et al., 1996)).

Protein	Function	PCNA-binding consensus
p21	CDK regulatory protein	RQ--MTDFY---RR---
FEN-1	Structure-specific endonuclease	-QGRLD-FF---S---
Cdc27	DNA polymerase δ subunit	-Q--I-SFF-K
DNA ligase 1	Replication specific DNA ligase	-Q--I--FF---K--K
RFC p140	Large subunit of PCNA-loading complex	-I--FFG------K
MSH3	Mismatch repair protein	-Q--LSRFF--------
MSH6	Mismatch repair protein	-Q--L-SFF-K------
XPG	Nucleotide excision repair endonuclease	TQ-RI--FF--------
MCMT	5' cytosine methyltransferase	-Q-TI--HF-----KRK
UNG2	Nuclear form of UNG uracil DNA glycosylase	-Q-TL--FF--------
WRN	Helicase required for genomic stability	DQWKL--DF-KL----
POGO	Type II transposases	-Q--L--FY
ApeXPF	Crenarchaeal NER endonuclease	RHASLDDFY
SsoXPF	Crenarchaeal NER endonuclease	TTSLF-DFL
SsoXPB	Crenarchaeal NER helicase	MV-RL-RYF

Table 4.1 – Examples of PCNA Interaction Motifs (PIM). Note the highlighted double aromatic motif. Adapted from (Warbrick, 2000).

4.2 Molecular cloning of ApePCNA subunits.

A system for overexpression of SsoPCNA was devised by Stephen Bell's group at Cambridge (Dionne et al., 2003) and was adapted by Malcolm White's group at St. Andrews to study SsoXPF activity. An agreement was made that Malcolm's White's group would continue to work on SsoXPF-PCNA complexes, whilst our group would

continue to work on the homologues from *Aeropyrum pernix*. In order to attempt to elucidate the structural basis for XPF activation by PCNA in *Aeropyrum pernix*, several new plasmid constructs were designed and made. Firstly, open reading frames for the three ApePCNA subunits were amplified using appropriate PCR primers and genomic DNA from *Aeropyrum pernix* (a gift from Dale Wigley, Molecular Enzymology Laboratory, Cancer Research UK). The ApePCNA subunits were named ApePCNA1, ApePCNA2 and ApePCNA3 according to the earlier *A.pernix* PCNA study (Daimon et al., 2002). Within the study of Daimon et al., two possible start codons for ApePCNA2 were identified. The authors chose the start that encoded the shorter protein and found that endogenous ApePCNA was ran slightly bigger on a SDS-PAGE gel than their recombinant protein. In hindsight the first start codon was probably the correct one and this is the one chosen for this study. The open reading frame for ApePCNA1 was cloned into the plasmid vector pET30a using Ndel and Hind3 restriction sites. The open reading frame for ApePCNA2 was cloned into pET3a, using Ndel and BamH1 restriction sites. The open reading frame for ApePCNA3 was cloned into the plasmid vector pET30a using Ndel and BamH1 restriction sites (figure 4.3). All three constructs coded for untagged proteins.

4.3 Molecular cloning of new ApeXPF constructs

In addition to the ApePCNA constructs required, two new ApeXPF constructs were also required. The major interaction between ApeXPF and ApePCNA is thought to utilise a C-terminal PCNA interaction motif (PIM) on XPF. The main ApeXPF construct described in Chapter 3 encodes residues 18-231. This construct is truncated at both the amino and carboxy termini, and lacks the PIM. Thus two new constructs were designed and made. Open reading frames encoding full length ApeXPF (1-254) and just the $(HhH)_2$ domain and the C-terminal PIM (164-254) were amplified using appropriate primers and *Aeropyrum pernix* genomic DNA. These open reading frames were cloned into pET14b-3C using Ndel and BamH1 restriction sites (figure 4.4 (a) and (b)).

4.4 Expression and purification of ApePCNA

Expression trials of the three individual ApePCNA subunits were undertaken using the study of Daimon et al. (2002) as a guide. Trials were hindered by the inability to

transform BL21(DE3) competent cells with plasmids coding for ApePCNA2. The inference was that leaky expression of ApePCNA2 in the cells was toxic, possibly due to interference in the host cells' DNA dynamics. The use of minimal media to eliminate any β-galactosides in the media causing leaky expression did not circumvent the problem. Nor did the use of Rosetta pLysS, an *E.coli* BL21(DE3) strain which expresses T7 lysozyme, a known inhibitor of T7 RNA polymerase. A solution was found using BL21-AI cells, in which the expression of T7 RNA polymerase is under the control of the arabinose inducible araBAD promoter. Leaky expression is not an issue in these cells. Figure 4.5 shows schematically how the expression of T7 RNA polymerase is regulated in these cells. To enhance yields, a strain of BL21-AI containing a chloramphenicol resistant Codon Plus plasmid (extracted from BL21(DE3) Codon Plus cells (Stratagene)) was generated in order to supplement tRNAs for rare codons. This strain was named BL21-AI +.

For the sake of simplicity BL21-AI + cells were used for the expression of all three ApePCNAs. The transformed strains were grown as described in Appendix A. Expression of T7 polymerase within the cells was induced with L-arabinose and IPTG was also added to release the lac repressor protein from the lac operator sequence on the expression plasmid. Cells were induced at 37 °C for 5 hours before harvesting, and storage at -80 °C.

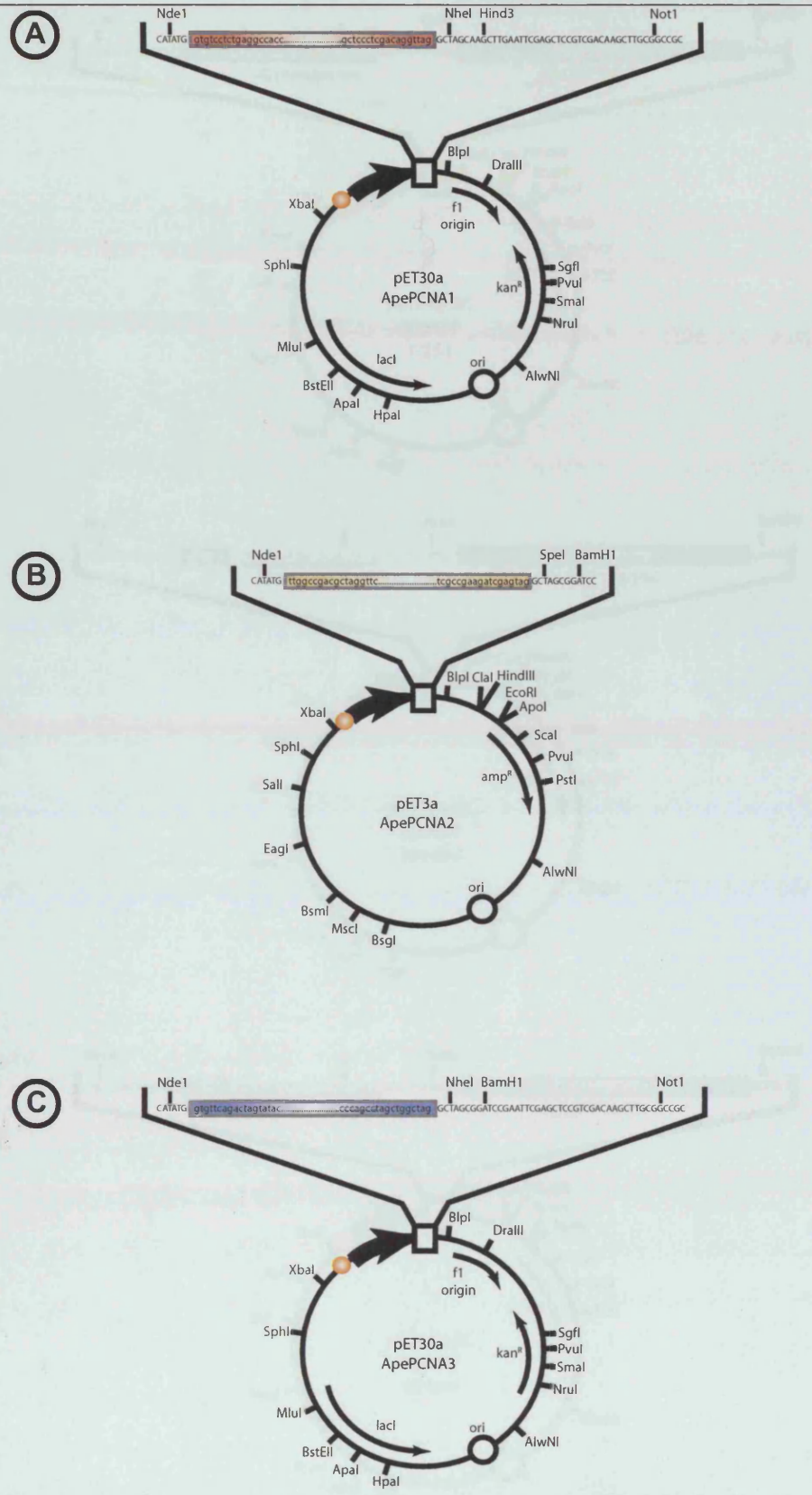

Figure 4.3 – Schematic diagram of individual ApePCNA expression plasmids. [A] pET30a ApePCNA1. [B] pET30a ApePCNA2 and pET3a ApePCNA2. [3] pET30a ApePCNA3.

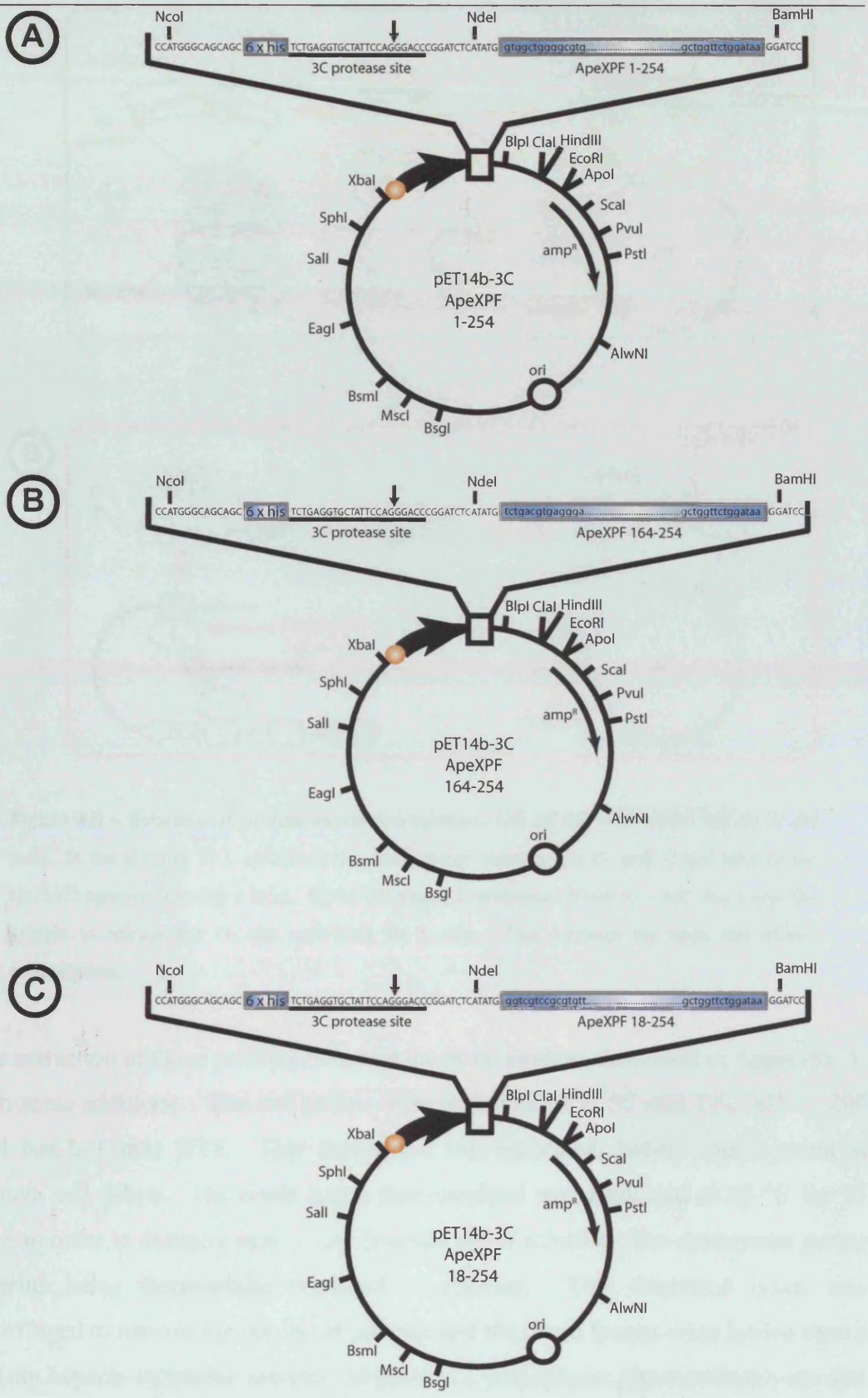

Figure 4.4 – Schematic diagram of ApeXPF expression plasmids used in this chapter. [A] Full length ApeXPF (1-254). [B] ApeXPF HhH_2 & PIM (164-254). [C] ApeXPF (18-254).

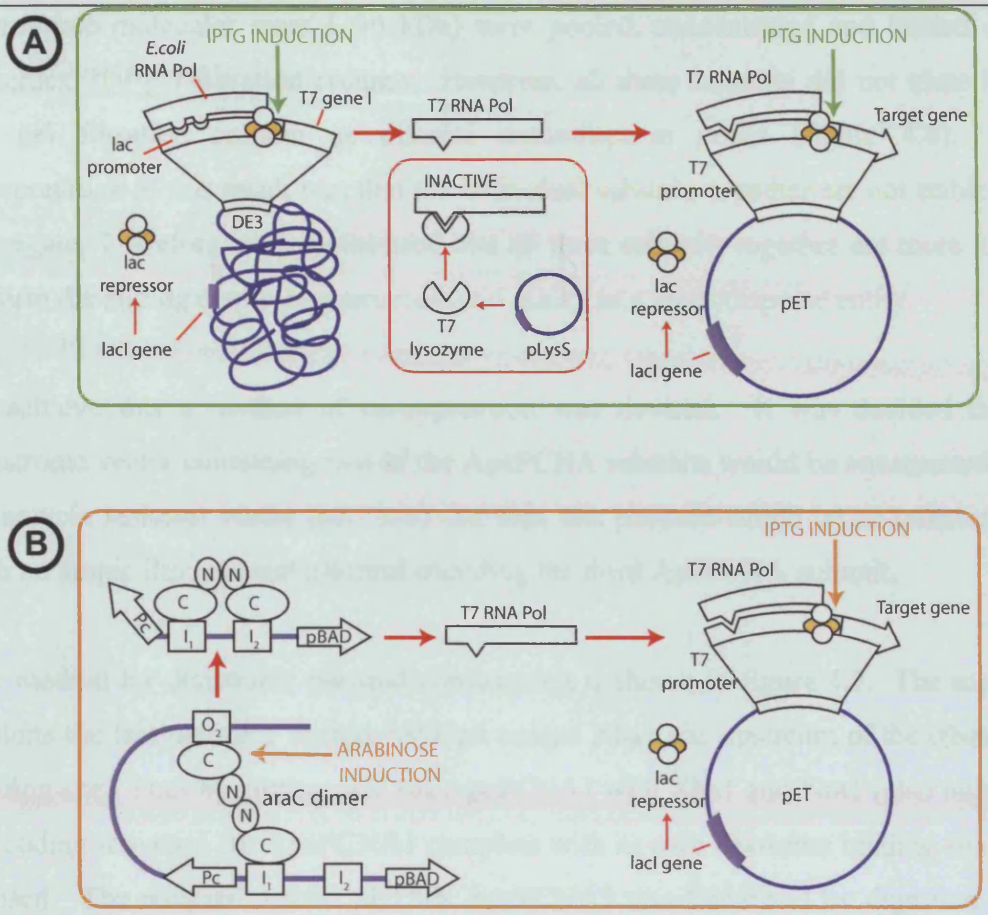

Figure 4.5 – Schemes of protein expression systems. [A] BL21(DE3) cells. [B] BL21-AI cells. In the absence of L-arabinose the AraC dimer contacts the O_2 and I_1 half sites of the araBAD operon, forming a loop. Upon induction, L-arabinose binds to AraC and cause the protein to release the O_2 site and bind the I_2 site. This releases the loop and allows transcription.

The extraction of these proteins followed the basic strategy described in Appendix A, with some additions. The cell pellets were resuspended in 50 mM Tris (pH 8), 100 mM NaCl, 1 mM DTT. This suspension was sonicated, before centifugation to remove cell debris. The crude lysate that remained was incubated at 65 °C for 15 mins in order to denature most *E.coli* proteins out of solution. The *Aeropyrum pernix* proteins being thermophilic remained in solution. This denatured lysate was centrifuged to remove the denatured proteins and the crude lysates were loaded onto a HiTrap heparin-sepharose column. Heparin is a polyanionic glycosaminoglycan and is often used in the purification of DNA-binding proteins. Proteins were eluted off the heparin column with NaCl and analysed by SDS-PAGE. All three ApePCNA subunits appeared to be expressed and those fractions containing proteins of an

appropriate molecular mass (~30 kDa) were pooled, concentrated and loaded on a Superdex 200 gel filtration column. However, all three subunits did not elute from the gel filtration column as discrete monodisperse peaks (figure 4.6). An interpretation of this result was that the individual subunits together are not stable and aggregate. Therefore we hypothesised that all three subunits together are more likely to form the sliding clamp ring structure and purify as a monodisperse entity.

To achieve this a method of co-expression was devised. It was decided that a dicistronic vector containing two of the ApePCNA subunits would be constructed in a kanamycin resistant vector (pET30a) and then this plasmid would be co-transformed with an ampicillin resistant plasmid encoding the third ApePCNA subunit.

The method for dicistronic plasmid construction is shown in figure 4.7. The method exploits the fact that pET vectors have an unique Xba1 site upstream of the ribosome binding site. Thus by cutting pET30a ApePCNA1 with Xba1 and Not1 (also unique), the coding sequence for ApePCNA1 complete with its own ribosome binding site was excised. The recipient vector pET30a ApePCNA3 was linearised by digestion with Nhe1 and Not1. Since Nhe1 (G↓CTAGC) and Xba1 (T↓CTAGA) leave identical cohesive ends, the ApePCNA1 (with its own ribosome-binding site) sequence can be ligated easily into linearised pET30a ApePCNA3. This results in a functional ApePCNA3-ApePCNA1 dicistronic plasmid with each gene possessing its own ribosome binding site.

This newly transformed strain was grown up, induced, harvested and the proteins extracted as described for the individual ApePCNAs. Purification of the proteins on Hi Trap heparin-sepharose yielded three proteins of the expected molecular mass that co-eluted together. These fractions were pooled, concentrated and loaded onto a Superdex 200 column. In this case the three ApePCNA subunits were eluted off the column together as a monodisperse peak (figure 4.8). The bands on the gel were sent for identification by mass spectrometry and peptides for all three ApePCNAs were identified. Thus a clean monodisperse ApePCNA123 sample had been achieved.

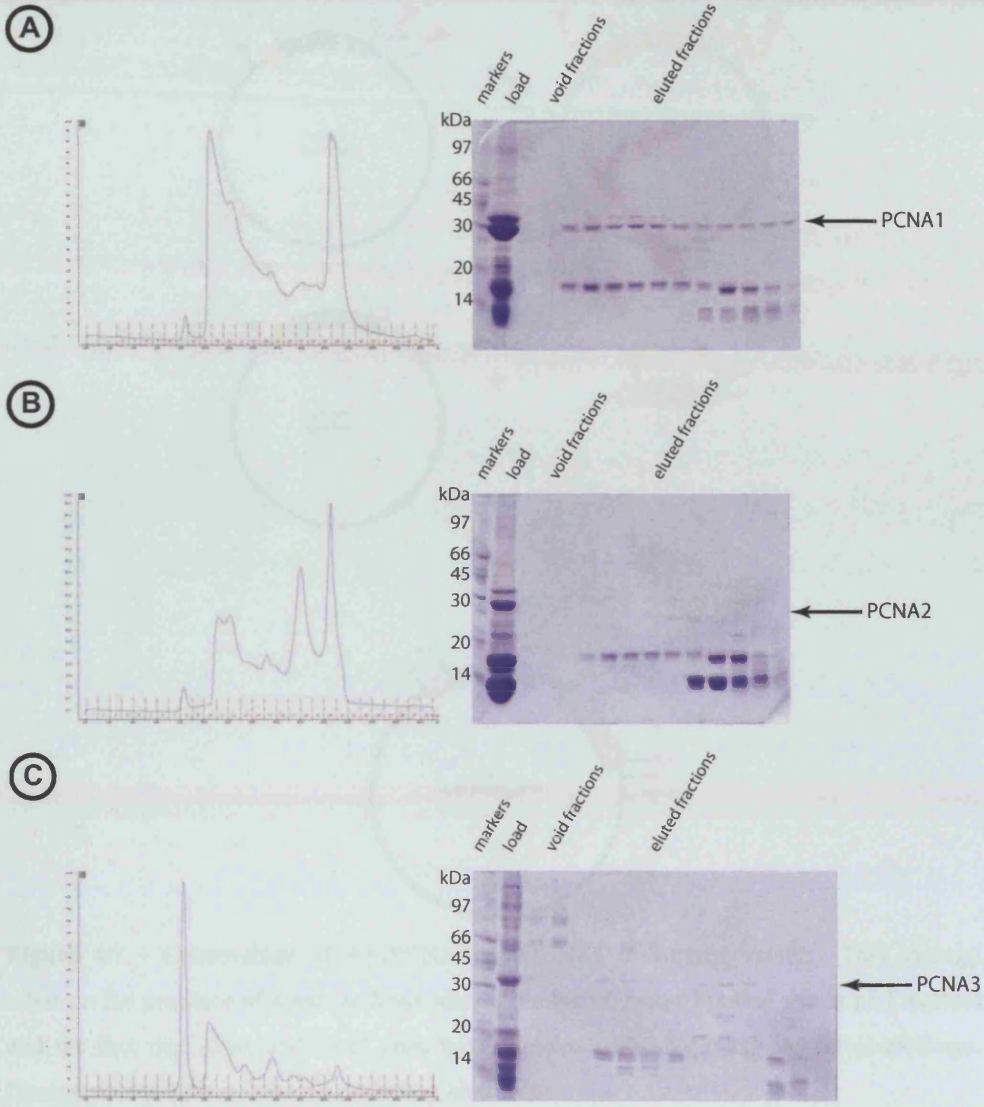

Figure 4.6 – Expression of ApePCNA subunits. Gel filtration profiles of ApePCNA subunits. [A] ApePCNA1. [B] ApePCNA2 [C] ApePCNA3. As all three profiles show, the individual ApePCNA subunits are not eluted as single monodisperse peaks.

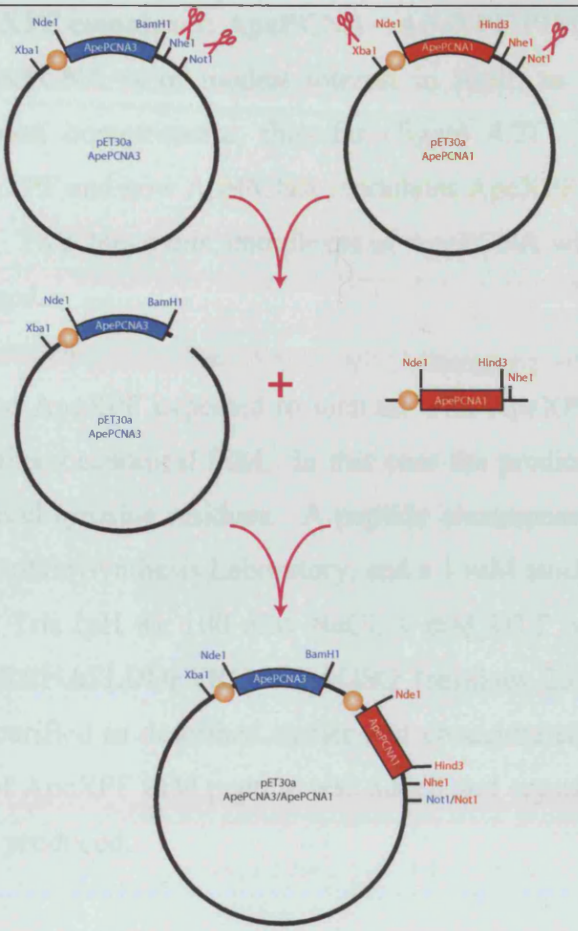

Figure 4.7 – Construction of ApePCNA3-ApePCNA1 dicistronic vector. This strategy relies on the presence of a unique Xba1 site before the ribosome binding site in pET vectors and the fact that Xba1 and Nhe1 sites have identical cohesive overhangs after cleavage. Orange circles indicate ribosome-binding sites.

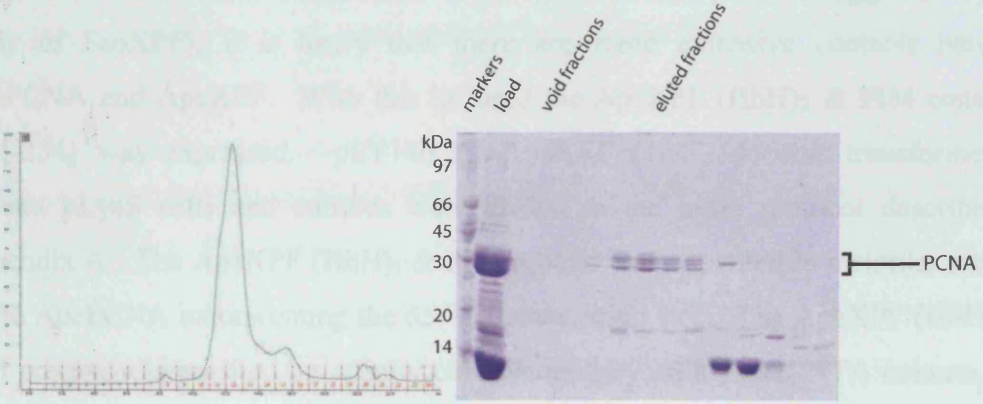

Figure 4.8 – Co-expression of ApePCNA subunits. Mass spectrometry confirmed that all three subunits were present in the complex and indicated a heterotrimeric complex had been formed. The yield of the complex was ~ 750 μg for 6 litres of bacterial cultures.

4.5 ApePCNA-ApeXPF complexes: ApePCNA – ApeXPF PIM.

The structure of ApePCNA is of modest interest in itself, as structures of other PCNAs, have all been homotrimeric, thus far (figure 4.2). The interaction of ApePCNA with ApeXPF and how ApePCNA modulates ApeXPF activity is however our primary interest. To achieve this, complexes of ApePCNA with various ApeXPF constructs were formed.

The minimum part of ApeXPF expected to interact with ApeXPF is the C-terminal tail region that contains a canonical PIM. In this case the predicted PIM is back-to-back phenylalanine and tyrosine residues. A peptide encompassing this region was synthesised by the Peptide Synthesis Laboratory, and a 1 mM stock of the peptide was made up in 20 mM Tris (pH 8), 100 mM NaCl, 1 mM DTT buffer. The peptide sequence was GGGRRHASLDDFYRGEGEAGSG (residues 233-254). ApePCNA was expressed and purified as described earlier and concentrated to 10 mg ml^{-1}. A 1:1.2 molar excess of ApeXPF PIM peptide was added and crystallisation trials were set up. No hits were produced.

4.6 ApePCNA-ApeXPF complexes: ApePCNA – ApeXPF $(HhH)_2$ & PIM.

The PIM has been shown to be crucial in the interaction of PCNAs with many binding partners (Warbrick; 2000). However, if the ApePCNA stimulates ApeXPF activity by mechanisms other than enhancement of substrate affinity, (as is suggested by the study of SsoXPF), it is likely that there are more extensive contacts between ApePCNA and ApeXPF. With this in mind the ApeXPF $(HhH)_2$ & PIM construct (164-254) was expressed. pET14b-3C ApeXPF (164-254) was transformed in Rosetta pLysS cells and cultures were grown to the basic protocol described in Appendix A. The ApeXPF $(HhH)_2$ & PIM protein was extracted in a similar method to the ApePCNA incorporating the 65 °C denaturation step. The ApeXPF $(HhH)_2$ & PIM protein was purified by affinity chromatography on a nickel-NTA column, and after washing, eluted by cleavage with 3C protease.

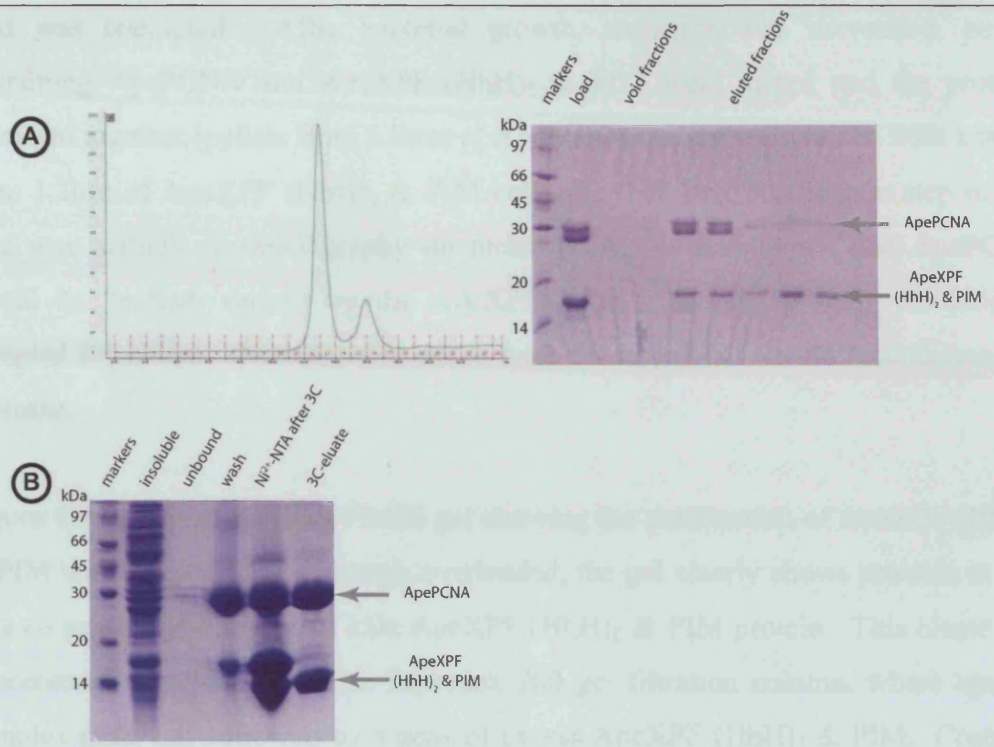

Figure 4.9 – Purification of an ApePCNA-ApeXPF (HhH)2 & PIM complex. [A] Gel filtration trace of ApePCNA- ApeXPF (HhH)2 & PIM complex. The major peak can be attributed to the 98 kDa complex. The minor peak is excess ApeXPF (HhH)2 & PIM. The SDS-PAGE gel of fractions from the gel filtration shows complex formation. [C] SDS-PAGE gel of ApePCNA- ApeXPF (HhH)2 & PIM complex purification using the His_6-tag on ApeXPF (HhH)2 & PIM.

The eluted ApeXPF $(HhH)_2$ & PIM protein was then gel filtered on a Superdex 75 column. The peak fractions were retained for complex formation with ApePCNA. Monodisperse gel filtered ApePCNA was added to a peak fraction of ApeXPF $(HhH)_2$ & PIM. The mixture was concentrated and loaded onto a Superdex 200 column. Although there was only a modest shift in the position of the ApePCNA peak, SDS-PAGE analysis indicated that complex formation had occurred, and that excess ApeXPF $(HhH)_2$ & PIM protein was eluted later (figure 4.9 (a)). The complex was concentrated to 10 mg ml^{-1} and a limited number of crystallisation trials were set up.

The limit as to the number of crystallisation trials set up was the amount of protein available. A yield of just 500 μg at this stage is only enough for three 96 well screens, using the Mosquito robot which can dispense 200 nl drops. In a bid to increase yields, co-extraction and purification of ApePCNA and ApeXPF $(HhH)_2$ &

PIM was attempted. After bacterial growth, induction and harvesting, pellets containing ApePCNA and ApeXPF $(HhH)_2$ & PIM were mixed and the proteins extracted together (pellets from 5 litres of ApePCNA culture were mixed with a pellet from 1 litre of ApeXPF $(HhH)_2$ & PIM culture). The first purification step in this case was affinity chromatography on nickel-NTA. It was hoped that ApePCNA would be "pulled- down" by the ApeXPF $(HhH)_2$ & PIM protein, resulting in complex formation. Proteins were eluted from the nickel-NTA with recombinant 3C protease.

Figure 4.9 (b) shows an SDS-PAGE gel showing the purification of ApeXPF $(HhH)_2$ & PIM with ApePCNA. Although overloaded, the gel clearly shows proteins at ~30 kDa co-purifying with the 10 kDa ApeXPF $(HhH)_2$ & PIM protein. This eluate was concentrated and loaded onto a Superdex 200 gel filtration column, where again a complex peak was followed by a peak of excess ApeXPF $(HhH)_2$ & PIM. Crucially the yield of complex improved significantly. Crystallisation trials were set-up using the vapour diffusion method and the Mosquito robot. A variety of sparse matrix and grid screens were set up. Initially no crystal hits were produced but after nine weeks crystals were found in a condition from the Hampton Research Index Screen (2.4 M sodium malonate, pH 7). Crystals were found in trays set up at 20 °C and 6 °C (figure 4.10 (a)). Sodium malonate is a cryoprotectant at concentrations greater than 3 M (Holyoak et al.; 2003) and thus the crystal was cryoprotected by raising the malonate concentration within the drop to 3 M. Only one crystal from the Mosquito drop was large enough for diffraction studies and was taken to the ESRF ID 23-2. The crystal diffracted well, to better than 2.5 Å (figure 4.10 (b)). However, the diffraction pattern showed evidence of multiple lattices and a dataset was not collectable. However, these test shots remain a promising lead and have been pursued by John Lally (Structural Biology Laboratory).

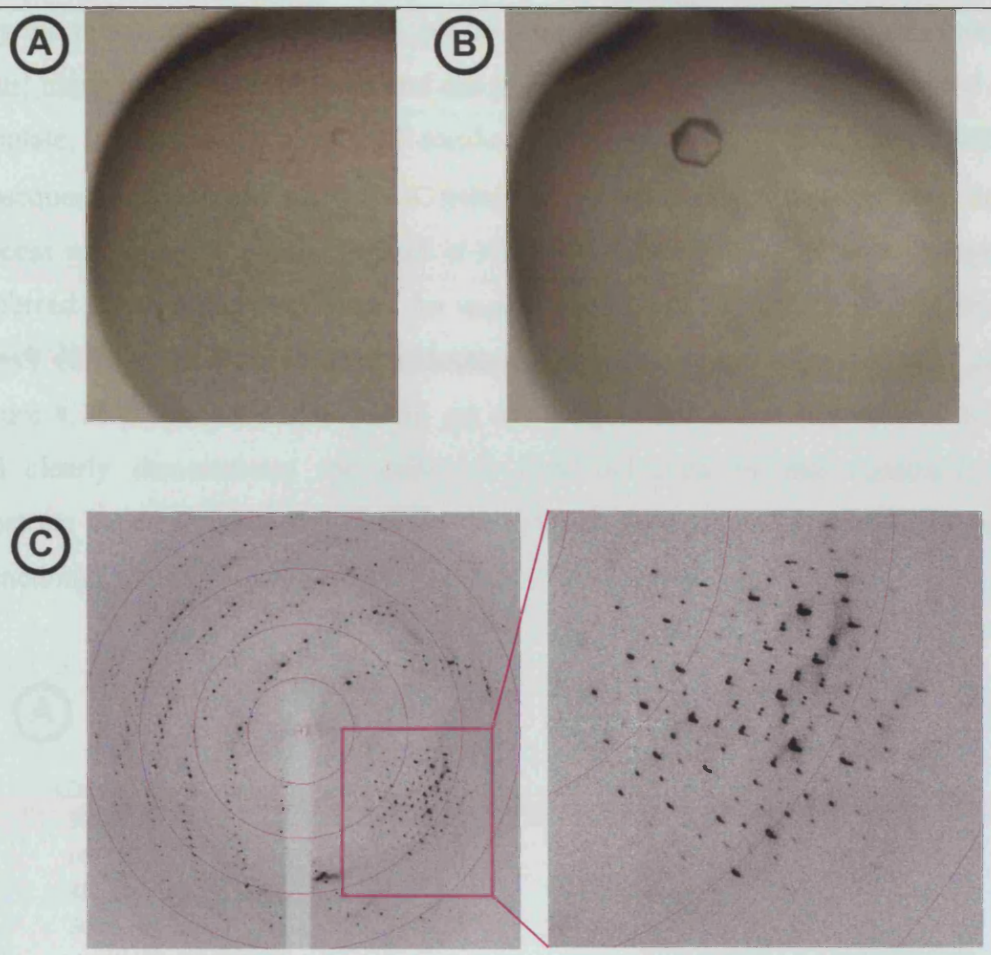

Figure 4.10 - ApePCNA – ApeXPF HhH2 & PIM crystals and diffraction. Complex crystals grown in sodium malonate at [A] 6 °C and [B] 20 °C. [C] Diffraction from the crystal shown in [B]. The data extends to 2.5 Å in the corners, although as the close-up shows there are multiple lattices. Circles are at 12, 6, 4 and 3 Å. Approximate cell dimensions are 74, 74, 54 Å.

4.7 ApePCNA-ApeXPF complexes: ApePCNA – ApeXPF .

Ultimately the structure of ApePCNA and full length ApeXPF would provide the most information as to how ApeXPF activity is regulated by ApePCNA. As discussed in section 4.3, a new construct encoding the entire full length of ApeXPF (1-254) was made. Initial expression trials of full length ApeXPF used Rosetta pLysS cells and used buffers containing at least 300 mM NaCl (in line with experience gained from the work described in Chapter 3). However, after purification of full length XPF on nickel-NTA, very little ApeXPF was produced (figure 4.11 (a)). At this point the construct design was re-evaluated. The ApeXPF (18-231) construct described in Chapter 3 expresses well, but as discussed the carboxy-terminal truncation made it unsuitable for this purpose. However, there was no obvious reason

why the full-length ApeXPF could not be truncated at the amino-terminus as before. Thus, using appropriate primers and the pET14b-3C ApeXPF (1-254) plasmid as a template, cDNA encoding ApeXPF residues 18-254 (ApeXPFΔN) was amplified and subsequently cloned into pET14b-3C using Nde1 and BamH1 restriction sites. In the process rare arginine codons present at residues 18, 20, 26 and 29 were changed to preferred *E.coli* arginine codons. An expression trial of this construct used Rosetta pLysS cells, a 25 °C overnight induction and buffers containing 300 mM NaCl. Figure 4.11 (b) shows a SDS-PAGE gel of the first purification step on nickel-NTA and clearly demonstrates the improved yield achieved by this construct. This construct represented a viable route to achieve a complex of ApePCNA and a "functional" ApeXPF.

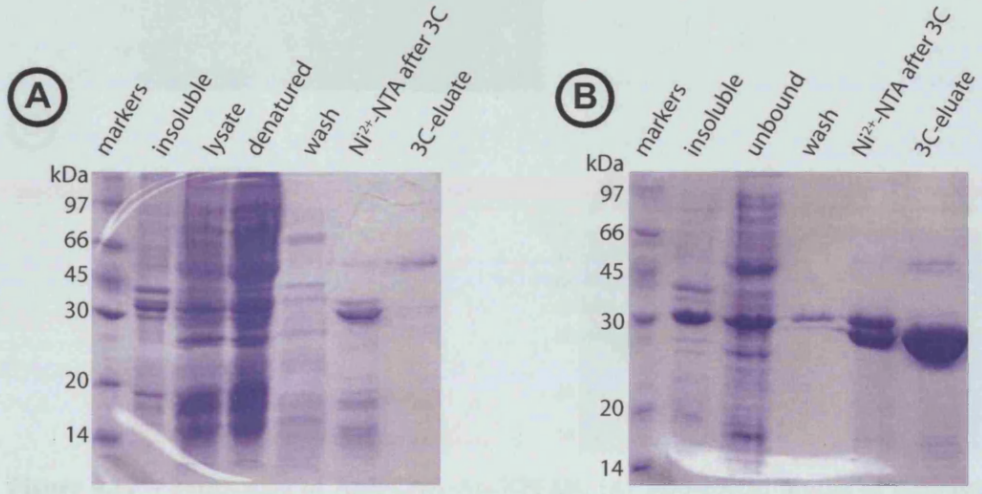

Figure 4.11 – Expression of ApeXPF constructs. [A] Full length ApeXPF (1-254). [B] ApeXPFΔN (18-254). As the gels show the expression level of ApeXPFΔN is a dramatic improvement on full-length XPF.

The ApeXPFΔN was then examined for its ability to form a complex with ApePCNA. Bacteria expressing ApePCNA and ApeXPFΔN were grown, induced and harvested as detailed in Appendix A. Cell pellets were then resuspended and mixed in a ratio of 6 litres of ApePCNA to 1 litre of ApeXPFΔN. This suspension was then sonicated and heat-treated as before. This lysate was loaded onto a Hi Trap chelating sepharose column, which had been pre-charged with nickel ions. The column was eluted using an imidazole gradient and the ÄKTA FPLC. The gradient elution was selected for two reasons. Firstly, earlier experiments attempting elution with 3C protease were unsuccessful. Secondly, it was critical that the correct concentration of imidazole was

used to elute the complex. Too little imidazole may not efficiently elute all the complex, but too little may have broken the complex. Figure 4.12 (a) shows the results of the imidazole elution. As the gel shows the ApePCNA and ApeXPFΔN complex appears to have formed. Fractions appearing to contain the binary complex were pooled and concentrated before gel filtration on a Superdex 200 column. As figure 4.12 (b) shows, the complex elutes at an appropriate elution volume.

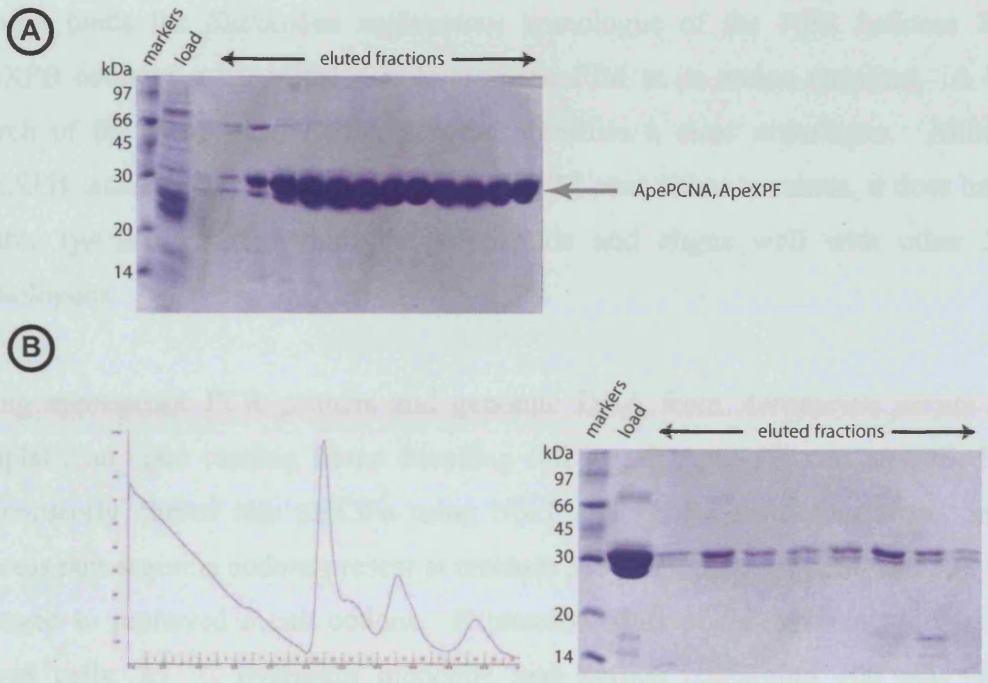

Figure 4.12 – Purification of ApePCNA-ApeXPFΔN. [A] SDS-PAGE profile of complex purification on Ni2+-sepharose. The column was eluted with an imidazole gradient. [B] Gel filtration profile of the complex. All three components co-purify together and the elution volume is consistent with a 140 kDa complex. (The X-axis indicates the volume in ml and the Y-axis the absorbance in mAU).

Although the structure of this complex would be of great interest, this was beyond the time limits of this study. The activity of this complex is discussed in section 4.10.

4.8 Expression and purification of ApeXPB

During the formative stages of this study, it became obvious that the binding of a homodimeric XPF to a heterotrimeric PCNA would create an asymmetric complex leaving one of the PCNA subunits without a binding partner. The study of the

interaction of SsoXPF with SsoPCNA indicates that the SsoXPF dimer interacts preferentially to two of the three SsoPCNA subunits. There are clear similarities between *Sulfolobus solfataricus* and *Aeropyrum pernix* and their enzymes; not least highlighted by the fact that SsoPCNA stimulates ApeXPF activity in vitro (Newman et al, 2005). Thus, it seems reasonable to assume that two of the ApePCNA subunits specifically bind to ApeXPF whilst the other does not. Unpublished data from Malcolm White's laboratory in St. Andrews has suggested that the third SsoPCNA subunit binds the *Sulfolobus solfataricus* homologue of the NER helicase XPB. SsoXPB contains a canonical double aromatic PIM at its amino-terminus. A blast search of the *Aeropyrum pernix* genome identifies a clear orthologue. Although ApeXPB lacks a canonical double aromatic PIM near either terminus, it does have a double tyrosine motif within the polypeptide and aligns well with other XPB homologues.

Using appropriate PCR primers and genomic DNA from *Aeropyrum pernix* as a template, an open reading frame encoding full length ApeXPB was amplified and subsequently cloned into pET30a using NdeI and EcoR1 restriction sites. In the process rare arginine codons present at residues 5, 9, 533, 534, 535, 536 and 538 were changed to preferred *E.coli* codons. Expression trials of this protein used Rosetta pLysS cells, 25 °C overnight induction and buffers containing 100 mM NaCl. Bacteria were grown, induced and harvested as usual and protein was extracted by sonication and heat denaturation as before. The first purification step was affinity chromatography on a Hi Trap heparin-sepharose column with elution using a NaCl gradient. After analysis by SDS-PAGE, fractions containing protein of the appropriate molecular mass (~66 kDa) were pooled, concentrated and loaded onto a Superdex 200 gel filtration column. A clean peak was eluted which contains the target ~66 kDa protein. The protein was identified by mass spectrometry (Protein Analysis Laboratory, Cancer Research UK) to be the expected ApeXPB protein.

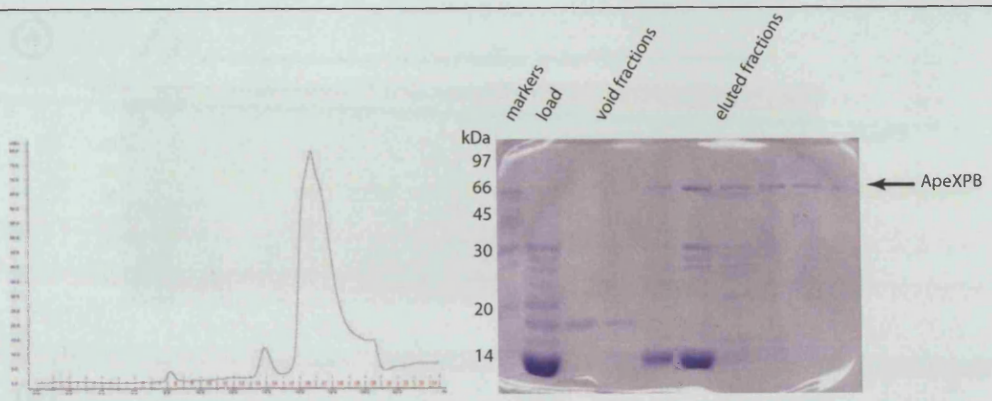

Figure 4.13– Expression and purification of ApeXPB. Gel filtration trace of ApeXPB together with the SDS-PAGE gel of peak fractions. (The X-axis indicates the volume in ml and the Y-axis the absorbance in mAU).

4.9 ApePCNA-ApeXPF complexes: ApePCNA – ApeXPF ΔN - ApeXPB

The ApeXPB was then examined for its ability to form a complex with ApePCNA and ApeXPF. Bacteria expressing ApePCNA, ApeXPFΔN and ApeXPB were grown, induced and harvested as detailed in Appendix A. Cell pellets were then resuspended and mixed in a ratio of 6 litres of ApePCNA to 1 litre of ApeXPFΔN to 5 litres of ApeXPB. The proteins were extracted and purified in an identical manner to the ApePCNA-ApeXPFΔN complex, with elution from a Ni^{2+}-sepharose column with imidazole. Figure 4.14 (a) shows the results of the imidazole elution. As the gel shows the ApePCNA, ApeXPFΔN, ApeXPB complex appears to have formed. Fractions appearing to contain the ternary complex were pooled and concentrated before gel filtration on a Superdex 200 column. As figure 4.14 (b) shows, the complex elutes at an appropriate elution volume, although the peak does have a slight shoulder to it, which may be indicative of heterogeneity.

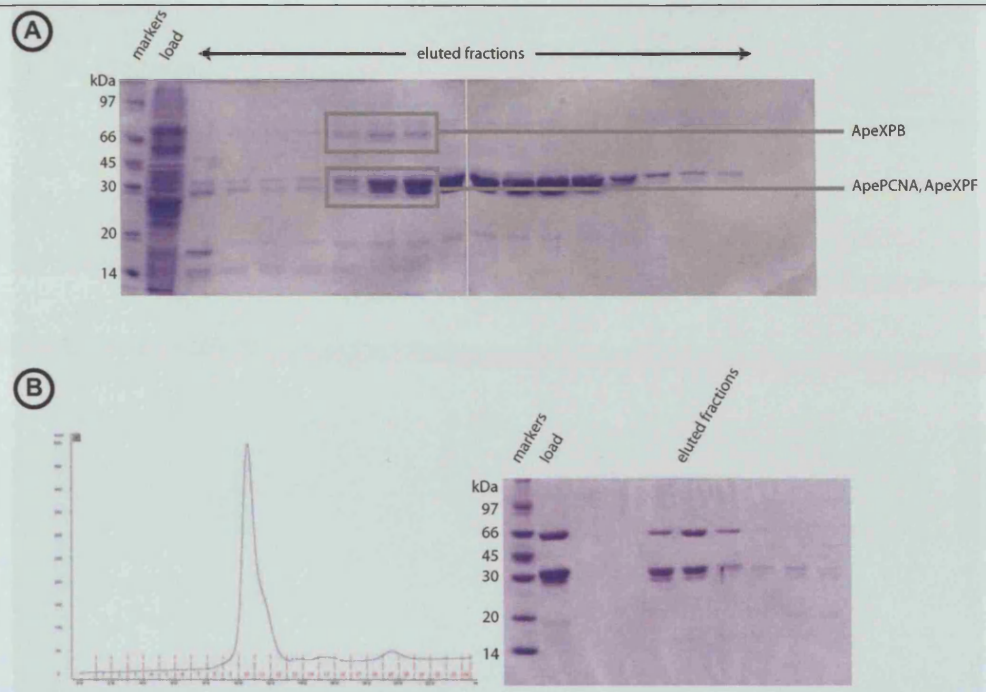

Figure 4.14 – Purification of ApePCNA-ApeXPFΔN-ApeXPB. [A] SDS-PAGE profile of complex purification on Ni2+-sepharose. The column was eluted with an imidazole gradient. Fractions containing all three components are boxed. [B] Gel filtration profile of the complex. All three components co-purify together and the elution volume is consistent with a 200 kDa complex. (The X-axis indicates the volume in ml and the Y-axis the absorbance in mAU).

The overall yield of ApePCNA/ApeXPFΔN/ApeXPB was modest (~ 500 μg from 6 litres + 1 litre + 5 litres of each culture respectively) and not suitable for a broad crystallisation screen. This coupled with the probable heterogeneity meant that the few crystallisation trials set up did not produce any hits. However, the theoretical mass of the ternary complex is 208 kDa means it could be large enough to study by electron microscopy. A collaboration with Dr. Hannes Ponstingl and Prof. Helen Saibil at Birkbeck College was initiated and is currently underway.

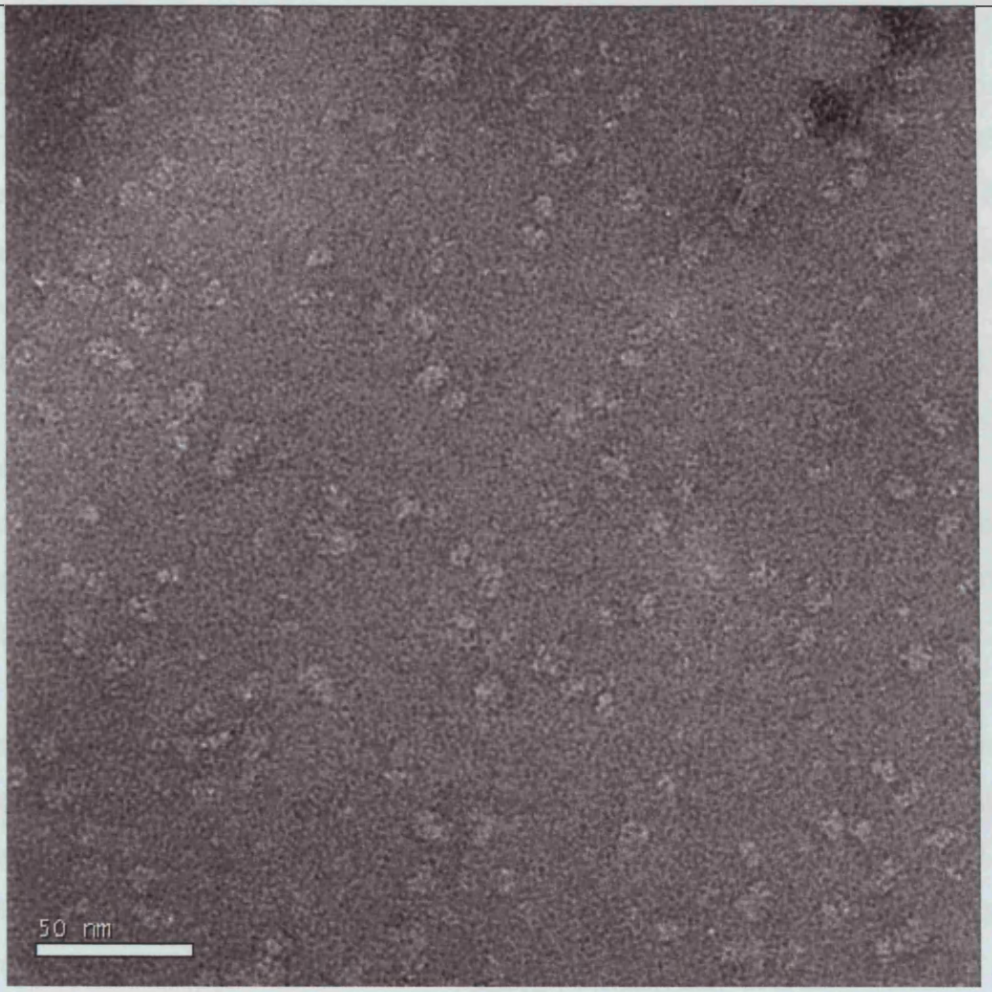

Figure 4.15 – Electron micrograph of ApePCNA- ApeXPFΔN- ApeXPB complex. Micrograph taken using a complex concentration of 0.025 mg ml-1 and at a magnification of 52,000 x g. The negative stain means the protein is white in the micrograph.

A sample of the ternary complex was prepared as before. An aliquot of the sample was applied to an EM grid and uranyl acetate was used in a negative staining experiment. Images were taken using the electron microscope at Birkbeck College. Although the electron microgram in figure 4.15 is difficult to decipher, ring structures with 6-fold symmetry with extra protein elements "hanging" off them were identified.

4.10 Activity of ApePCNA – ApeXPF complexes

The studies of the activity of the SsoXPF-SsoPCNA complex have provided the first information on how crenarchaeal XPFs work. However, the activity of ApeXPF-ApePCNA has not been characterised to date. Moreover there are no published studies on the effect of XPB in such a complex. To investigate the complex,

ApePCNA-ApeXPFΔN-ApeXPB and ApePCNA-ApeXPFΔN were prepared as discussed before. In order to verify that any perceived activity was due to the prepared complexes and not an *E.coli* contaminant, a catalytically dead form of ApeXPFΔN was engineered. Amino acid residue 68 of ApeXPF was mutated from aspartic acid to asparagine using the Stratagene QuikChange protocol. This mutation was functionally equivalent to the hXPF D720N mutation made in chapter 2 and originally in the mutagenesis study of Enzlin and Scharer (2002). It is expected that the role of D68 is as a general base, activating a water molecule as the nucleophile required for phosphodiester bond hydrolysis. Hence this D68N mutation was predicted to kill activity but not to affect the protein fold. Protein complexes of ApePCNA-ApeXPFΔN D68N with and without ApeXPB were expressed and purified as for the wild-type protein.

The protein complexes were assayed for endonuclease activity using ^{32}P-labelled substrates by Maureen Biggerstaff (Structural Biology Laboratory). Assay conditions were adapted from the SsoXPF studies (Roberts et al., 2003). Two synthetic DNA substrates were used. The first was a 3'-flap, which is generally regarded as a good substrate for SsoXPF. As figure 4.16 details, ApePCNA-ApeXPF cleaves this substrate efficiently in the presence and absence of ApeXPB. There is some residual activity in the D68N mutant experiments. However, the cleavage pattern is consistent with structure-specific activity. The second substrate was a splayed arm. The complexes also show activity against this substrate, although the cleavage appears to be less efficient and there is a difference in the presence and absence of ApeXPB. In this case the D68N mutation eliminates endonuclease activity completely.

These results show that the ApeXPFΔN protein expressed is active. However, no clear effect of ApeXPB on the activity could be detected.

4.11 Summary

In this chapter, a system for expression of the heterotrimeric PCNA from *Aeropyrum pernix* was established for the first time. After failed attempts at producing individual ApePCNA subunits, co-expression of all three ApePCNA subunits together yielded a clean, monodisperse complex.

Complexes consisting of ApePCNA and different ApeXPF constructs were pursued. Crystallisation trials of ApePCNA with an ApeXPF PIM peptide were unsuccessful. A more successful approach involved crystallisation trials with ApePCNA complexes with an ApeXPF $(HhH)_2$ & PIM construct. Initial crystals diffracted well at the ESRF and although the diffraction images clearly show multiple lattices, these crystals provide a promising start. Complexes of ApePCNA and ApeXPFΔN are also currently being crystallised.

A ternary complex of ApePCNA-ApeXPFΔN-ApeXPB was also formed. The yields of this complex were not sufficient for crystallisation. However, a collaboration has enabled structural studies by electron microscopy to be initiated. Preliminary results have identified intact ApePCNA rings with approximate 6-fold symmetry possibly with appropriate staining for ApeXPF and ApeXPB. Low-resolution structure determination by this method remains a possibility.

Finally, the endonuclease activity of ApeXPF-PCNA complexes has been examined. ApePCNA-ApeXPFΔN shows structure-specific activity towards 3'-flap and splayed arm substrates similar to that of SsoXPF-PCNA. The presence of ApeXPB does not appear to radically alter the activity of this complex.

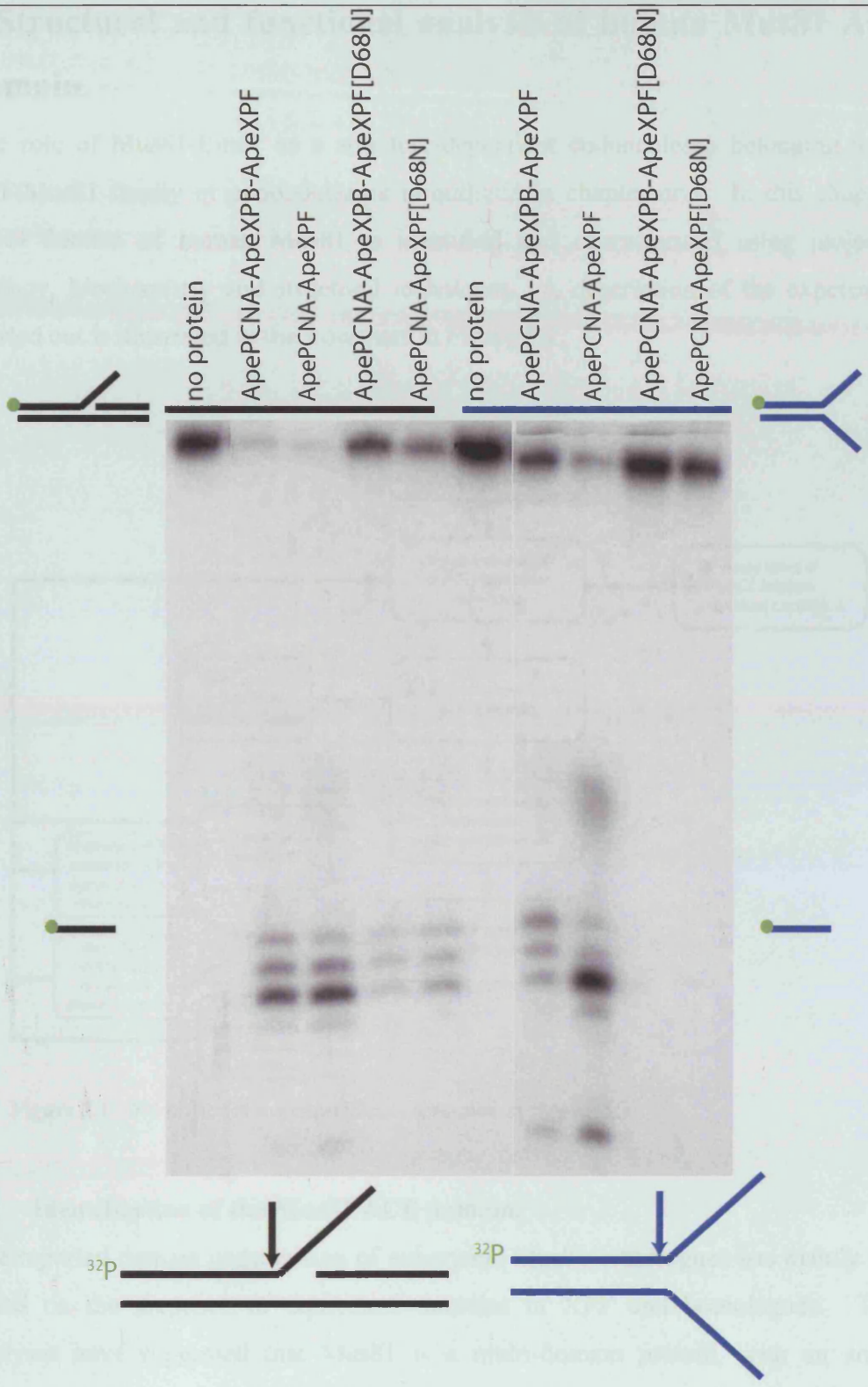

Figure 4.16 – ApeXPF activity. Autoradiograms showing the activity of ApeXPF complexes towards 3'-flap and splayed arm substrates. (Substrate sequences can be found in Appendix A).

5 Structural and functional analysis of human Mus81 ACE domain.

The role of Mus81-Eme1 as a structure-dependent endonuclease belonging to the XPF/Mus81 family of endonucleases is outlined in chapter one. In this chapter a novel domain of human Mus81 is identified and characterised using molecular biology, biochemistry and structural techniques. A description of the experiments carried out is illustrated in the flowchart in Figure 5.1.

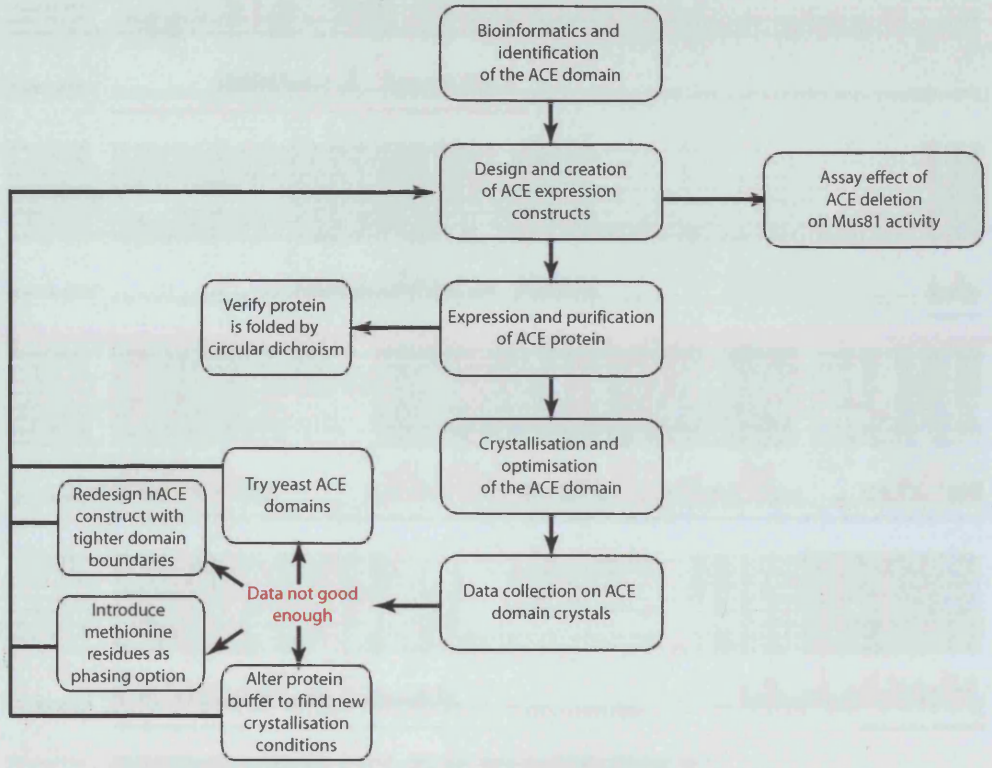

Figure 5.1 – Flowchart of the experiments described in chapter 5.

5.1 Identification of the Mus81 ACE domain.

The reported domain organisation of eukaryotic Mus81 orthologues has mainly been based on the presence of equivalent domains in XPF and homologues. These analyses have suggested that Mus81 is a multi-domain protein, with an amino-terminal HhH domain (Pfam PF00633), followed by a nuclease domain (Pfam PF02732) and finally a carboxy-terminal HhH domain.

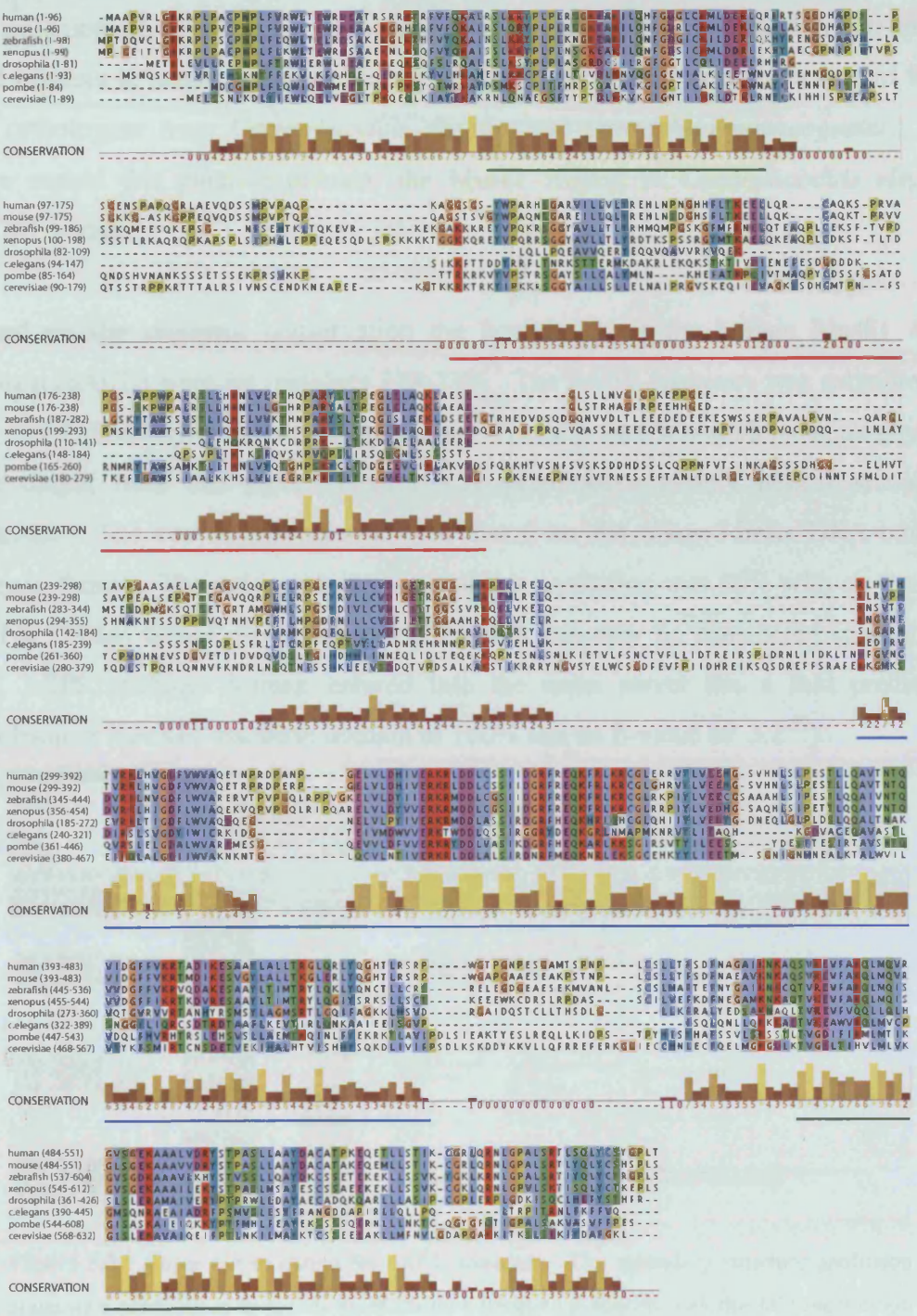

Figure 5.2 – Sequence conservation in eukaryotic Mus81 orthologues. Sequence numbers for each alignment panel are indicated. The blue line demarcates the nuclease domain; and the green lines, the HhH domains. The red line demarcates the ACE domain.

Human Mus81 is 551 amino acid residues in length (in fission yeast 572 residues) and thus a large proportion of Mus81 has no clearly identifiable domain assignment. Subsequent sequence alignment and secondary structure predictions identified an extra region of conservation between the amino-terminal HhH domain and the

nuclease domain (figure 5.2). This patch of conservation is approximately 100 amino acid residues in length and is present in all eukaryotic Mus81 orthologues, apart from the orthologues from *Caenorhabditis elegans* and *Drosophila melanogaster*. We have named this putative domain, the Mus81 Absent in *Caenorhabditis elegans* (ACE) domain.

Based on the sequence conservation the boundaries of the human Mus81 ACE domain (hACE) were set (residues 128-230). The hACE sequence was submitted to the 3D-PSSM/Phyre fold recognition server (http://www.sbg.bio.ic.ac.uk/phyre/). The output from this server, (figure 5.3), predicted a mainly helical secondary structure. The best fold prediction was that from the winged-helix DNA-binding domain family. The estimated precision of this prediction was 60% with an E-value of 1.9. These values mean that this prediction can only be considered speculative. (c.f. hXPF nuclease domain entered into the same server has a fold prediction precision of ApeXPF nuclease domain of 100% and an E-value of 3.2^{-18}).

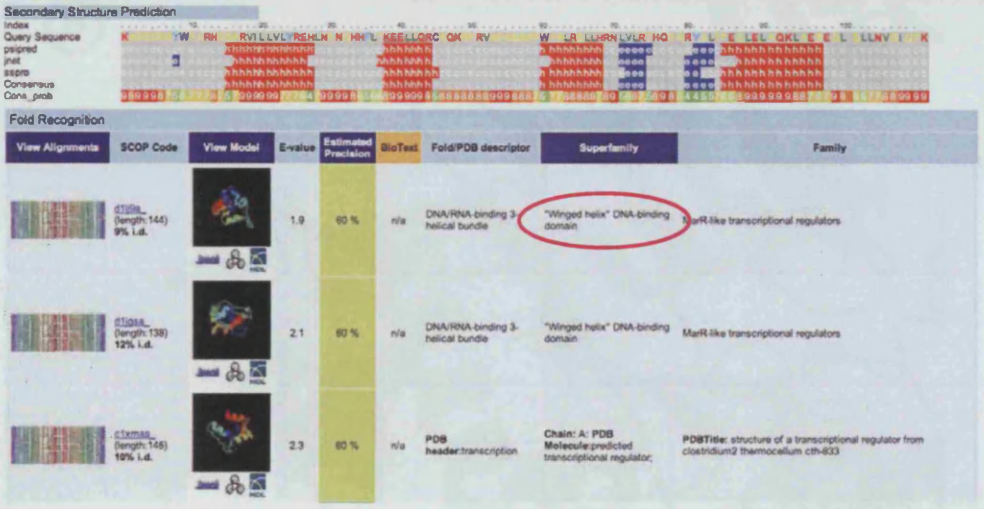

Figure 5.3 – Phyre server output for hACE sequence. The secondary structure prediction suggests a predominantly α-helical structure, with three β-strands with the fold recognition algorithm consistent with a winged helix domain (ringed).

Winged-helix domains (also known as forkhead domains) are well-characterised protein domains normally associated with DNA recognition. Studies of winged-helix proteins began with the discovery of HNF-3, a liver specific transcription factor

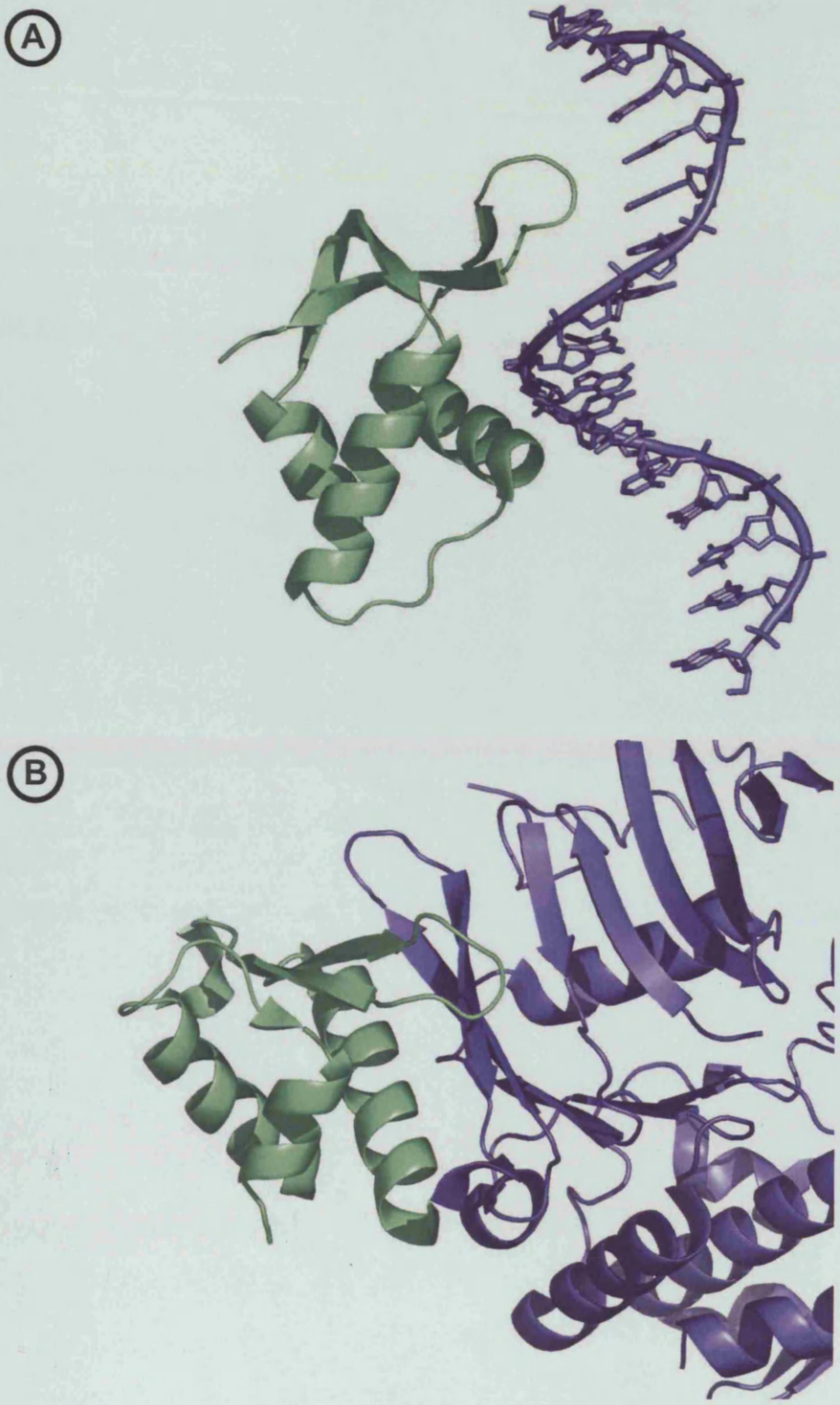

Figure 5.4 – Structure of winged-helix proteins. [A] Transcription factor RFX1 bound to its cognate DNA (PDB accession code 1DP7 (Gajiwala et al. 2000)). [B] SCC1 (green) bound to SMC1 (blue). Both proteins are involved in sister chromatid cohesion (PDB accession code 1W1W (Haering et al. 2004)).

(Gajiwala and Burley, 2000). Structures of winged-helix motifs, (figure 5.4), have shown them to be compact α/β folds, usually consisting of three α-helices and three β-strands. These secondary structure elements combine to form two wings that flank an α-helix, inspiring the motif's name. It is this α-helix that is often involved in the recognition of DNA by binding in a major groove.

The theoretical isoelectric point (pI) of the hACE domain is 10.0. Often proteins with such a preponderance of basic residues might be expected to bind to a polyanionic ligand such as DNA. This coupled with the winged-helix prediction could point to a role for the hACE in DNA binding and recognition. However, there are two pieces of information to temper this conclusion. Firstly, not all winged-helix domains are involved in binding DNA. Others are involved in protein-protein interactions. For example, the carboxy-terminal region of SCC1, a protein involved in sister chromatid cohesion and segregation is a winged-helix domain that has been shown to mediate an interaction with SMC1, another cohesion protein (figure 5.4, (Haering et al., 2004)). Secondly, recent literature has suggested residues 125-244 of hMus81 interact with the C-terminal domain of the human Bloom's Syndrome helicase BLM (hACE falls within these residue limits; (Zhang et al., 2005)). Thus the questions we sought to address were, what fold does hACE adopt and is it involved in Mus81 substrate recognition or does it mediate and regulate Mus81 protein-protein interactions.

5.2 Molecular cloning, expression and purification of hACE domain.

Based on the conservation observed from the sequence alignments discussed earlier, an open reading frame for hACE was amplified using appropriate PCR primers and a DNA template from a dicistronic pET28a hMus81-hEme1 plasmid (a gift from Alberto Ciccia, Genetic Recombination Laboratory, Cancer Research UK; details for cloning methods can be found in the general Materials and Methods in Appendix A). The hACE open reading frame was cloned into three different plasmid vectors; the first pET41a-3C, the second pGEX-6P2 and the third pET14b-3C. These constructs are depicted in Figure 5.5. On first inspection it may appear that that pET41a-3C-hACE and the pGEX-6P2-hACE plasmids are essentially the same but they are not. The difference is in the RNA polymerase promoter regions. Expression from pET41a-3C starts from a T7 RNA polymerase promoter, whilst pGEX-6P2 relies on

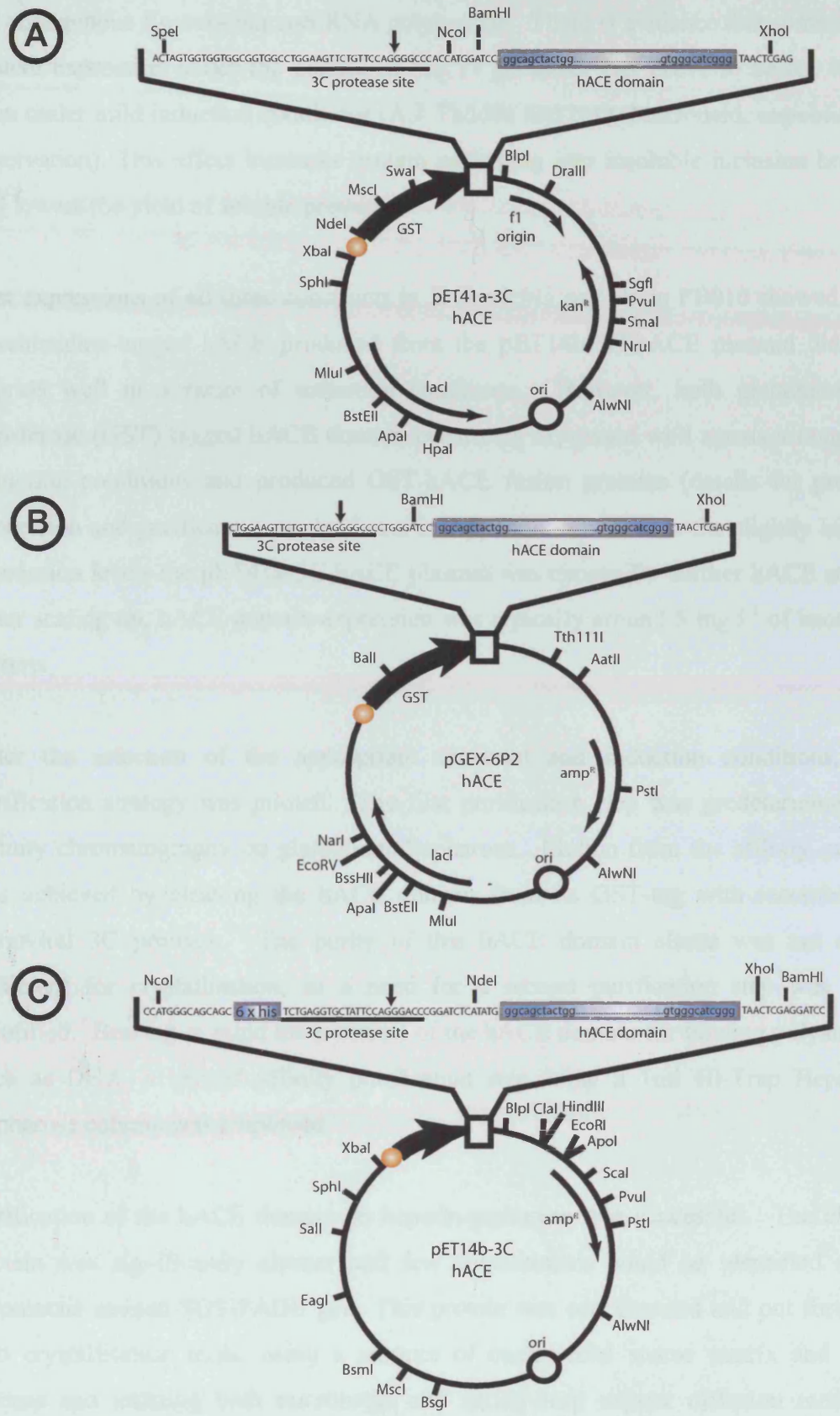

Figure 5.5 - Schematic diagram of hACE expression plasmids. [A] pET41a-3C hACE. [B] pGEX6P2-hACE. [C] pET14b-3C hACE.

the endogenous *Escherichia coli* RNA polymerase. There is evidence that sometimes protein expression under the control of the T7 promoter can prove to be too high, even under mild induction conditions (A.J. Fadden and N.Q. McDonald, unpublished observation). This effect increases protein packaging into insoluble inclusion bodies and lowers the yield of soluble protein.

Test expressions of all three constructs in *Escherichia coli* strain FB810 showed that hexahistidine-tagged hACE produced from the pET14b-3C-hACE plasmid did not express well in a range of induction conditions. However, both glutathione-S-transferase (GST) tagged hACE domain constructs expressed well across a range of induction conditions and produced GST-hACE fusion proteins (details for protein expression and purification can be found in Appendix A). Due to the slightly higher expression levels the pET41a-3C-hACE plasmid was chosen for further hACE study. After scaling up, hACE domain expression was typically around 5 mg l^{-1} of bacterial culture.

After the selection of the appropriate construct and induction conditions, the purification strategy was piloted. The first purification step was predetermined as affinity chromatography on glutathione-Sepharose. Elution from the affinity matrix was achieved by cleaving the hACE domain from its GST-tag with recombinant rhinoviral 3C protease. The purity of this hACE domain eluate was not quite sufficient for crystallisation, so a need for a second purification step was thus identified. Bearing in mind the potential of the hACE domain for binding polyanions such as DNA, a second affinity purification step using a 1ml Hi-Trap Heparin-Sepharose column was employed.

Purification of the hACE domain on heparin-sepharose was successful. The eluted protein was significantly cleaner and few contaminants could be identified on a Coomassie stained SDS-PAGE gel. This protein was concentrated and put forward into crystallisation trials, using a mixture of commercial sparse matrix and grid screens and utilizing both microbatch and sitting-drop vapour diffusion methods (details on crystallisation methods can be found in Appendix B). These crystallisation trials were unfortunately unsuccessful.

The lack of crystallisation hits led to concerns that the hACE protein was either not folded properly, or worse was not a discrete domain at all! With this in mind, circular dichroism experiments using hACE protein were performed.

Circular dichroism (CD) is a technique in which light is polarised into 2 circularly polarised components. Asymmetric molecules, such as proteins, absorb the two

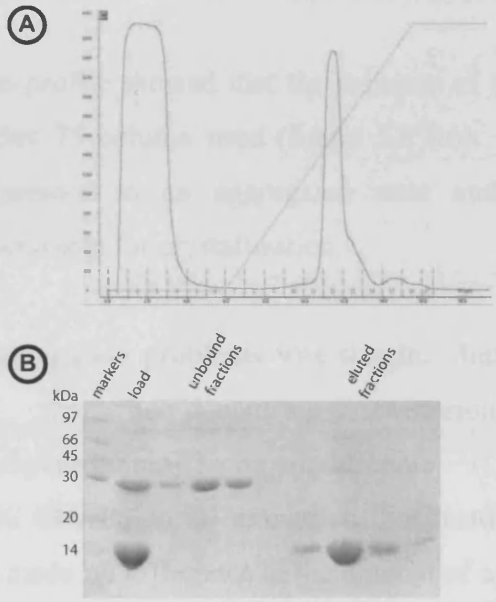

Figure 5.6 – Purification of hACE on heparin-sepharose. [A] ÄKTA-FPLC Trace. X-axis indicates the volume in ml and Y-axis is absorbance at 280 nm in mAU. The green line indicates the 0.1 – 1 M NaCl gradient employed. [B] SDS-PAGE gel of eluted fractions.

components differently. When one of these components is selectively absorbed, this yields light that is elliptical. This ellipticity can be measured and interpreted in terms of protein fold and secondary structure relative to a databse of known structure and CD spectra.

CD experiments were performed on hACE, using the apparatus of Dr. John Ladbury's group at UCL. The results of the experiment were merged and corrected for the contribution of the protein buffer to the spectrum. Figure 5.7 shows the hACE spectrum. Model protein secondary structures show characteristic CD spectra in far UV wavelengths. α-helices have strong ellipticity peaks near 192 nm with a broad trough between 200 and 240 nm (with minima at 208 and 223 nm). β-sheet spectra look similar but the maximum is at about 198 nm and the spectra in general are not as

intense. From figure 5.7 it is clear that hACE is folded and is largely α-helical in nature.

The confirmation that hACE was folded renewed confidence that crystallisation should be possible. To this end further characterisation of the protein was undertaken. Of particular interest was the oligomeric state of the protein and thus an analytical gel filtration experiment was performed to determine this.

The initial gel-filtration profile showed that the majority of the hACE protein ran in the void of the Superdex 75 column used (figure 5.8 (a)). This indicated that the hACE domain was present in an aggregated state and not a monodisperse, homogenous solution desirable for crystallisation.

A solution to these aggregation problems was sought. Initial attempts included a small amount of the non-denaturing zwitterionic detergent 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS) to a final concentration of 0.1 % CHAPS in all extraction, purification and storage buffers. However, its inclusion made no difference to the amount of aggregated hACE protein in the preparation. A second approach was to increase the concentration of sodium chloride, NaCl in the buffers. The rationale behind this approach is that an increased salt concentration provides increased counter-ions to polar side-chains and thus may encourage a monodisperse protein solution. The salt concentration was increased from 100 mM to 500 mM. hACE domain prepared in these new buffer conditions was cleaner after the first affinity chromatography step (higher salt concentrations often reduce non-specific interactions with affinity resins), and thus the eluate from this first step was concentrated and loaded straight onto a Superdex 75 gel filtration column.

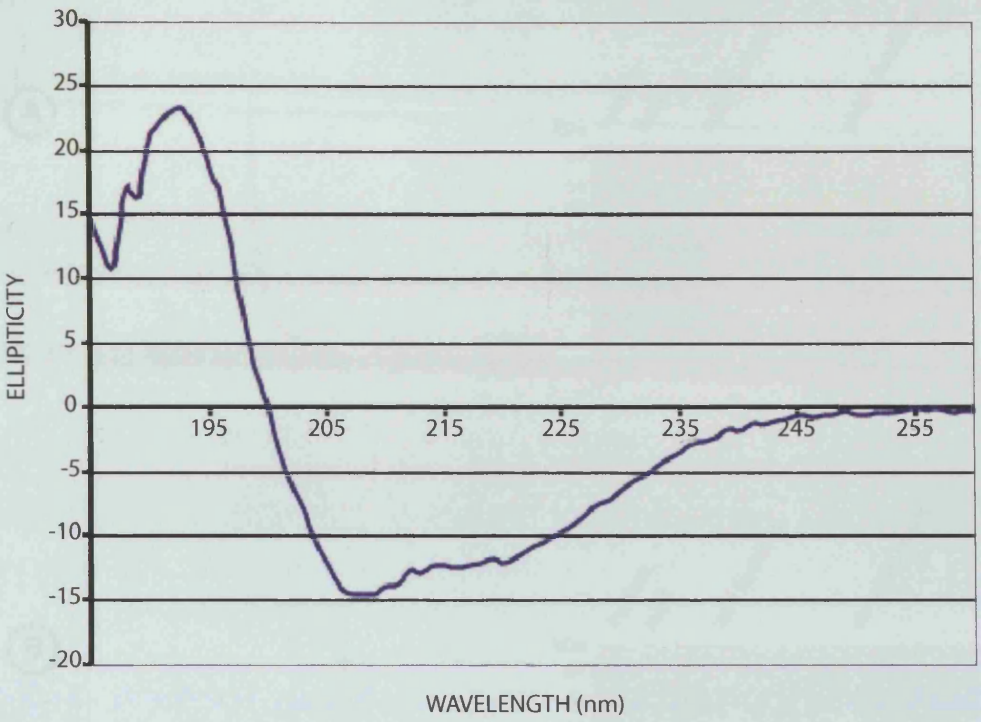

Figure 5.7 – Circular dichroism spectrum of hACE. The ellipiticity peak near 192 nm with a broad trough between about 200-240 nm is characteristic of a mostly α-helical protein.

As figure 5.8 (b) shows the majority of the hACE domain migrated with an apparent molecular weight of 9.3 kDa. This indicates that the hACE domain runs mainly as a monomer. Armed with pure, monodisperse protein further attempts were made to produce hACE crystals.

5.3 Crystallisation of hACE domain

Crystallisation trials were carried out using the hACE domain (20 mM Tris (8), 500 mM NaCl, 1 mM DTT) at a concentration of 10-12 mg ml^{-1} and a mixture of commercial sparse matrix and grid screens. Both sitting drop vapour diffusion and microbatch trials were performed. After 3 days small crystals were found in a sitting-drop ammonium sulphate ($(NH_4)_2SO_4$) grid screen. The condition was identified as 45 % saturated $(NH_4)_2SO_4$ buffered with 100 mM citric acid at pH 4. An optimisation tray was set up around this condition and larger hexagonal rods grew overnight in 41 % saturated $(NH_4)_2SO_4$ buffered with 100 mM citric acid at pH 3.5. Subsequently the substitution of citric acid for formic acid was found to improve the morphology of the crystals, resulting in fewer crystals growing together and single crystals suitable for diffraction studies (figure 5.9).

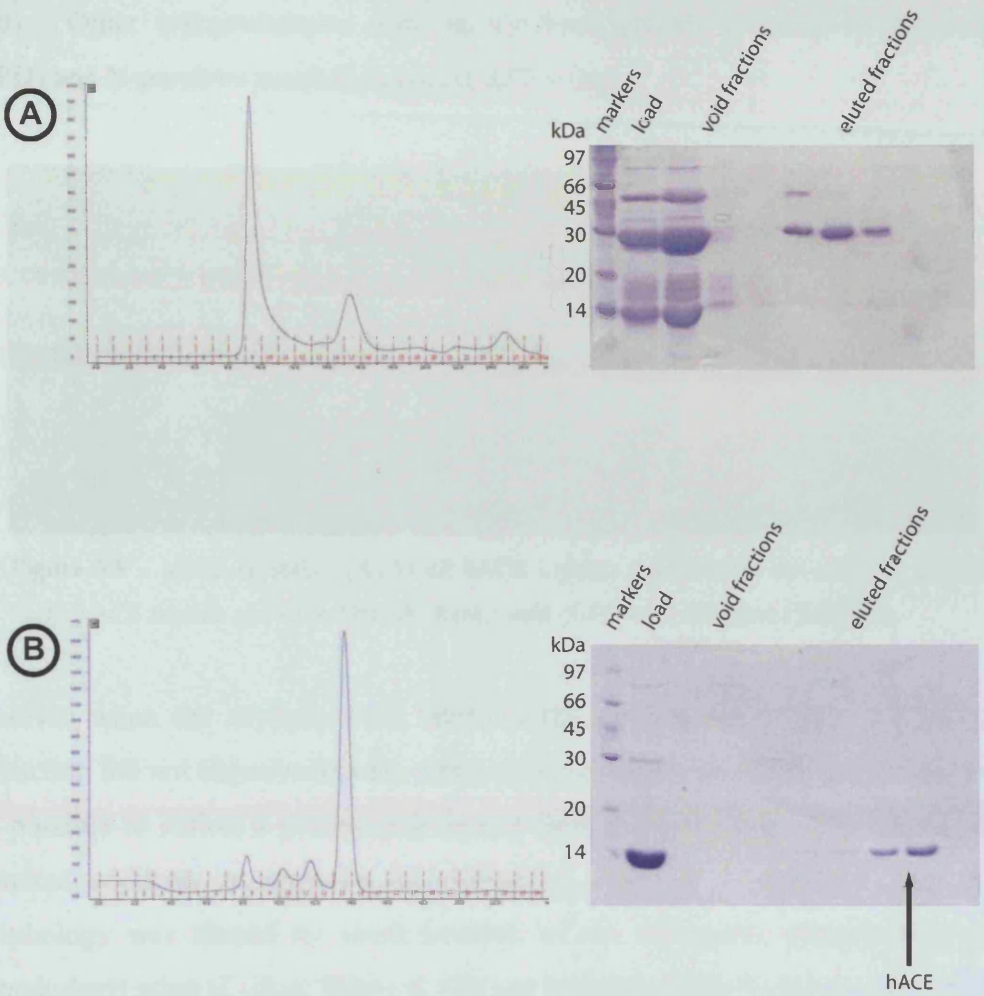

Figure 5.8 – Purification of hACE. [A] Gel filtration profile of hACE extracted in 100 mM NaCl. The majority of hACE protein elutes in the void fractions. [B] Gel filtration profile of hACE extracted in 500 mM NaCl. The majority of hACE protein elutes as a monodisperse species. The peak is consistent with the protein being monomeric. (X-axes indicate the volume in ml and Y-axes the absorbance at 280 nm in mAU.

To check that the crystals were not inorganic salt crystals, a minute amount of Hampton Research's Izit crystal dye was added to a drop with small crystals. Izit dye is a small molecule dye that fills solvent channels within protein crystals, colouring the crystals blue. Salt crystals do not possess large solvent channels, and thus the Izit dye cannot enter the crystal. In this case the crystals took up the dye and turned blue (figure 5.9 (a)). Initial diffraction studies using our in-house rotating anode X-ray generator provided further confirmation that the crystals were composed of protein. Initial images from the home source showed diffraction extending to 3.7 Å, using mother liquor and either 30 % glucose or 25 % glycerol as cryoprotectants (figure

5.10). Other cryoprotectants such as ethylene glycol, 2-methyl-2,4-pentanediol (MPD) and N-paratone resulted in poorer diffraction.

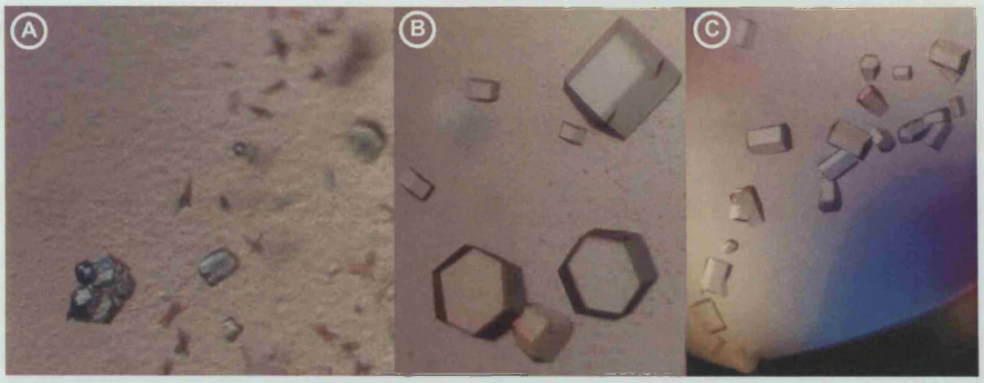

Figure 5.9 – hACE crystals. [A] Small hACE crystals stained with Izit dye. [B] & [C] Large hACE crystals grown in 100 mM formic acid (3.5), 41 % saturated $(NH_4)_2SO_4$.

However when the crystals were taken to the synchrotron (ESRF ID 14-1), the diffraction did not improve and the diffraction pattern showed high mosaicity. It was not possible to collect a processable dataset from these crystals. The use of simple chemical additives to improve the diffraction was not successful. The crystal morphology was altered by small amounts of the detergents nonaethylene glycol monododecyl ether ($C_{12}E_9$), Triton X-100 and hexadecyltrimethylammonium bromide (CTAB) (figure 5.11) but the diffraction was not improved and the crystals were very

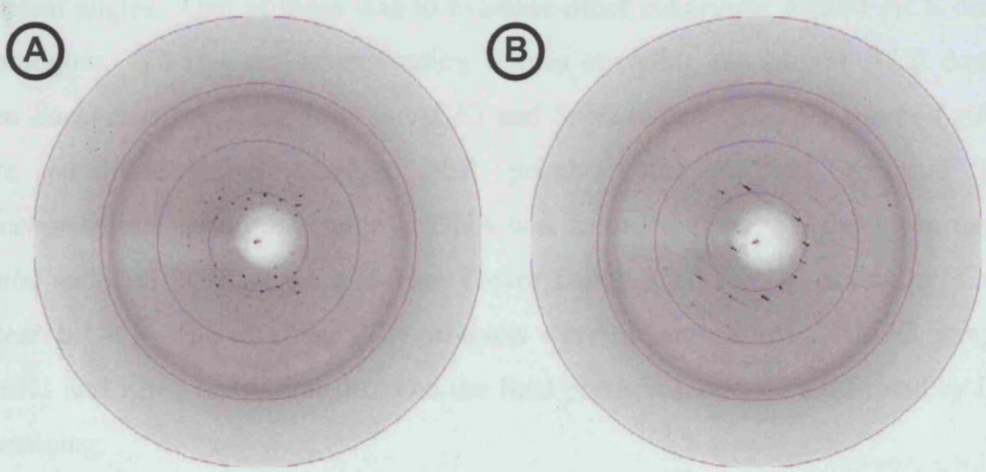

Figure 5.10 – hACE diffraction. Two test shots taken in-house using a rotating anode generator and a Mar detector. [A] 0 °. [B] 90 °. Data extends to 3.7 Å. Circles are at 10.4, 5.2, 3.5 and 2.6 Å. Cell parameters are a=60.6 Å, b=61.2 Å, c=150.2 Å ; α= 90 ° , β=90 °, γ=120 °.

difficult to handle. Various other techniques that have been shown to improve crystal diffraction such as crystal dehydration and incorporation of a small amount of cryoprotectant, such as glycerol or sucrose, into the precipitant mixture did not improve diffraction and often made it worse.

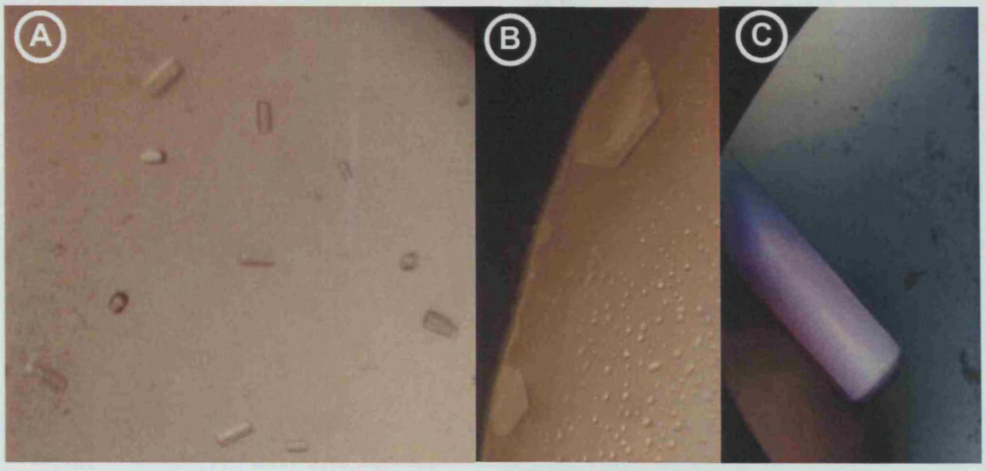

Figure 5.11 – hACE crystals. hACE was crystallised in the presence of the following detergent additives: [A] $C_{12}E_9$ (nonaethylene glycol monododecyl ether) [B] Triton X-100 [C] CTAB

5.4 Molecular cloning, expression and purification of yeast ACE domains.

Whilst attempts to improve the diffraction characteristics of the hACE domain crystals were underway, it seemed prudent to approach the problem from a number of different angles. One of these was to examine other eukaryotic Mus81 ACE domain orthologues. To this end open reading frames encoding the Mus81 ACE domains from *Saccharomyces cerevisiae* (scACE) and *Schizosaccharomyces pombe* (spACE) were amplified using suitable PCR primers and purified genomic DNA (*Saccharomyces cerevisiae* genomic DNA was from Promega, *Schizosaccharomyces pombe* genomic DNA was a gift from Trevor Duhig, Cell Cycle Laboratory, Cancer Research UK). The resultant PCR products were cloned into pET41a-3C using the BamH1 and Xho1 restriction sites and the final plasmid sequence confirmed by DNA sequencing.

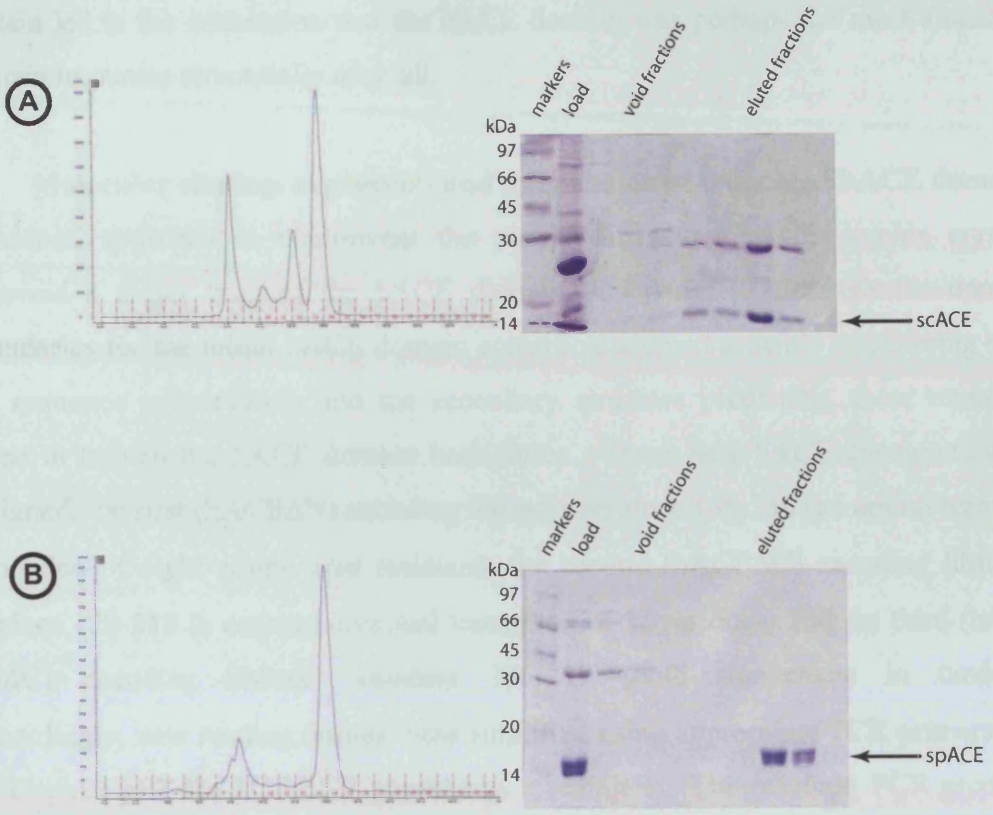

Figure 5.12 – Expression and purification of yeast ACE domains. [A] Gel filtration profile of scACE. The gel shows that scACE co-purifies with a bacterial contaminant. [B] Gel filtration profile of spACE. The gel shows spACE as a doublet, indicating degradation. (X-axes indicate the volume in ml and Y-axes the absorbance at 280 nm in mAU.

The domains were expressed and purified as for the hACE domain (figure 5.12). The scACE domain expressed well similar to its human counterpart and appeared to gel filter as a monomer. However, a bacterial contaminant consistently co-purified with the scACE protein, making protein purity an issue. The spACE domain is less well behaved, showing a doublet on a SDS-PAGE gel and was prone to precipitation. At first these problems were thought to be due to an inappropriate buffer pH (spACE pI ~ 8.5 and the pH of the Tris buffer used was initially 8.0) but use of a Tris buffer set to pH 7.0 did not circumvent these problems

Crystallisation trials using these proteins, a mixture of commercial sparse matrix and grid screens and the Mosquito robot did not produce any hits at all. Focussed trials around the crystallisation condition found for hACE were also set up. These were also unsuccessful. This coupled with the precipitation problems plaguing the spACE

protein led to the conclusion that the hACE domain was perhaps the most amenable domain to pursue structurally after all.

5.5 Molecular cloning, expression and purification of truncated hACE domain.

A second approach to circumvent the poorly diffracting hACE domain crystals followed a closer look at the hACE domain boundaries. Although the domain boundaries for the initial hACE domain constructs seemed sensible considering both the sequence conservation and the secondary structure prediction, there remained scope to tighten the hACE domain boundaries. Three new hACE constructs were designed; the first (hACEΔN) encoding hMus81 residues 136-230 (an amino-terminal truncation of eight amino acid residues), the second (hACEΔC) encoding hMus81 residues 128-219 (a carboxy-terminal truncation of 11 residues) and the third (hACE ΔNΔC) encoding hMus81 residues 136-219 (both truncations in tandem). Accordingly, new reading frames were amplified using appropriate PCR primers and the original pET41a-3C hACE plasmid as a template. The resultant PCR products were cloned into pGEX-6P2 using the restriction enzymes BamH1 and Xho1 and the final plasmids sequences confirmed by DNA sequencing.

The three new constructs were expressed and purified as for the original ACE domain (figure 5.13). All three constructs behave well, as for the original hACE construct and gel filter as monomers on a Superdex 75 column in similar high (500 mM) salt conditions. Preparations of all three gel filtered proteins were concentrated to 10-15 mg ml^{-1} by ultrafiltration.

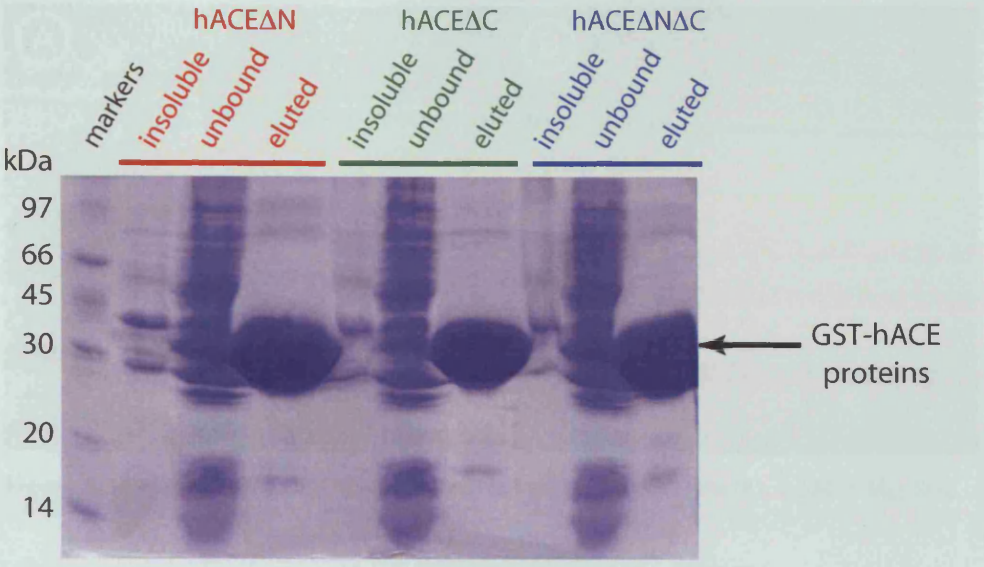

Figure 5.13 – Expression and purification of truncated hACE proteins. Initial purification of SeMet protein on glutathione-sepharose. The gel shows insoluble, unbound and glutathione eluted protein for each protein.

Crystallisation trials using these proteins, a mixture of commercial sparse matrix and grid screens and the Mosquito robot did not produce any hits in the short term. Focussed trials around the crystallisation condition found for the original hACE were also set up. These were unsuccessful. After between two and three months small crystals were found in a 200 nl sitting drop from the Nextal $AmSO_4$ suite containing hACEΔNΔC protein (figure 5.14). The conditions were slightly different from the original hACE domain hit. The conditions all contained 1.6 M $(NH_4)_2SO_4$ with a small selection of different buffers (100 mM HEPES pH 7.0, 100 mM Tris pH 8.0, 100 mM bicine pH 9.0). Small hACEΔNΔC crystals were also found in a 200 nl sitting drop from the Nextal pH clear suite. The condition in this case was 30 % PEG 6000 buffered with 100 mM HEPES pH 7.5. These crystals whilst promising were too small for diffraction studies. Attempts to reproduce these crystals have yielded similar small crystals unsuitable for diffraction studies.

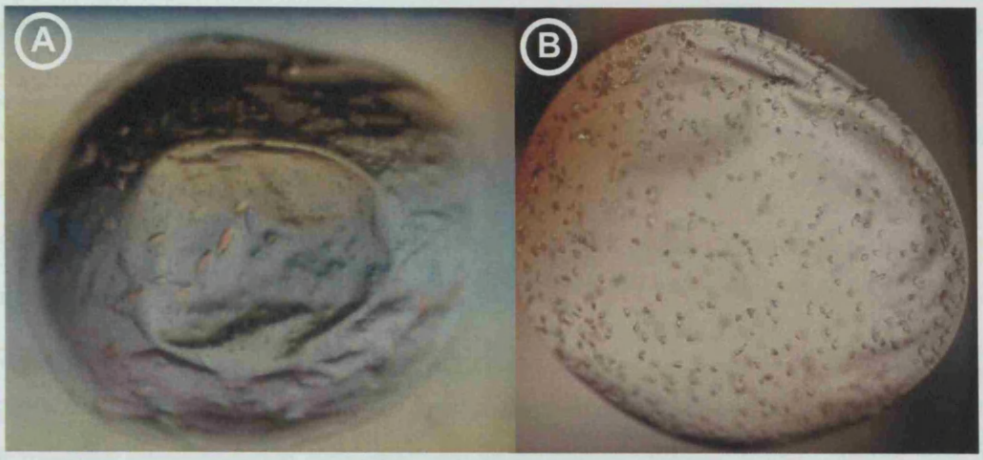

Figure 5.14 – hACEΔNΔC crystals. Crystals grown in 100 mM Tris (8), 1.6M $(NH_4)_2SO_4$.

5.6 Molecular cloning, expression and purification of selenomethionine hACE domain.

In parallel to attempts to improve the diffraction of the hACE domain crystals, the question of how to overcome the phase problem was raised. Since the possibility that the hACE domain is a winged-helix protein can be considered no more than speculative, molecular replacement was not considered a suitable way to obtain phases. Another common approach to consider was to derivatise hACE domain crystals with heavy atoms and then obtain experimental phases by multiple isomorphous replacement (MIR). Whilst this technique is conceptually simple, in this particular case it was potentially problematic as most of the repertoires of heavy atoms are non-reactive in the crystallisation conditions (pH 3.5), although there are a few platinum compounds that do react at low pH.

Another common method employed to tackle the phase problem is to analyse the anomalous scattering of selenium atoms to provide experimental phases. Most chemical elements introduced to within a crystal can provide anomalous scattering at an appropriate wavelength, but selenium is particularly attractive. This is because it is possible to produce recombinant protein in methionine-autotrophic *Escherichia coli* strains that contain selenomethionine (Se-Met) amino acid residues instead of methionine. This Se-Met containing protein can usually be purified and crystallised in similar conditions to the native protein, and the presence of Se-Met residues usually have little to no effects on the protein structure. The problem with this

approach in the case of hACE is that the domain does not contain any methionine residues apart from one vector-derived methionine at the extreme amino-terminus. As this methionine as at one of the polypeptide termini it is reasonable to suspect that it may well be disordered in the crystals and thus not useful for obtaining phases. A possible advantage to this approach is that in a few cases the introduction of Se-Met improves crystal quality and diffraction.

Sequence alignments of Mus81 ACE orthologues show that in some cases conserved hydrophobic positions are occupied by methionine (figure 5.15). From this, three conserved hydrophobic residues likely to be locked in the hydrophobic core of the protein fold were selected for mutation to methionine; leucine 185 (L185M), valine 194 (V194M) and leucine 212 (L212M). Suitable PCR primers were designed to effect these mutations using the Stratagene QuikChange protocol. Seven hACE methionine mutants were created in this way; three single mutants, three double mutants and one triple mutant. The presence of the mutations was confirmed by DNA sequencing. In order to maximise the opportunities for phasing, whilst minimizing disruption to the protein, the three double mutants were selected for expression.

The mutant hACE domains were expressed in minimal media supplemented with selenomethionine and purified as for the original hACE domain (figure 5.16). All three double mutants express well, and gel filtration experiments show that the mutant proteins are monodisperse and monomeric. The SeMet-labelled proteins were concentrated to 10-15 mg ml^{-1} by ultrafiltration before crystallisation.

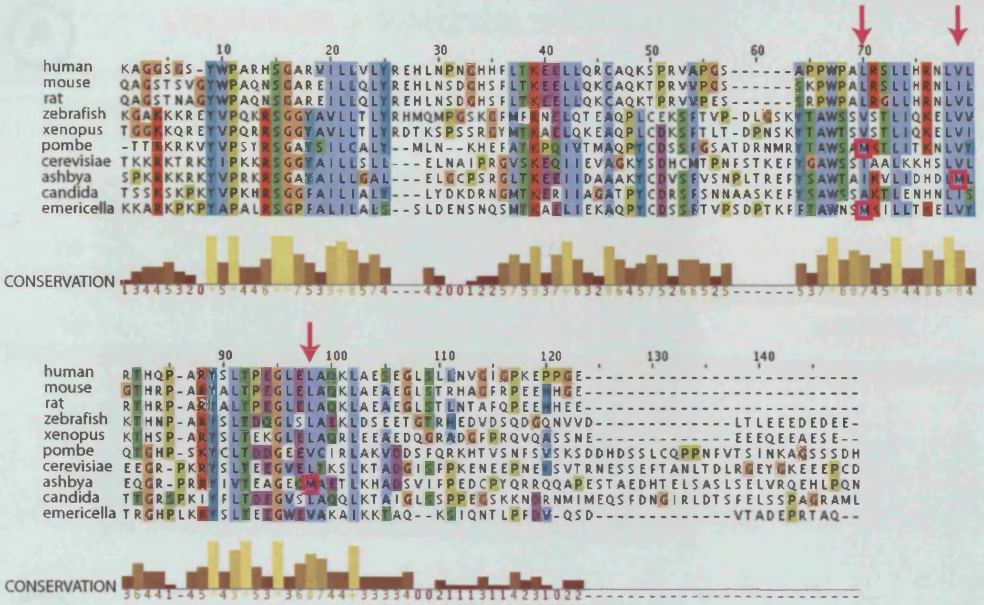

Figure 5.15 – Rationale for selection of residues for methionine mutagenesis. Arrows indicate sites selected for mutagenesis. Magenta boxes show methionine residues in ACE orthologues at these positions.

Crystallisation trials using these mutant proteins, a mixture of commercial sparse matrix and grid screens and the Mosquito robot did not produce any hits. Focussed trials around the crystallisation condition found for the original hACE were also set up. In the cases of hACE L185M L212M and hACE V194M L212M, this mini-screen produced only slight precipitate. However hACE L185M V194M produced tiny clusters of needles (figure 5.17 (a)). Whilst unsuitable for diffraction studies, these clusters were a useful start.

The formation of these clusters was very dependent on the concentration of the precipitant $(NH_4)_2SO_4$. For example for one batch of protein, drops with 40 and 40.2 % saturated $(NH_4)_2SO_4$ did not yield clusters but 40.4 and 40.6 % saturated $(NH_4)_2SO_4$ did. 40.8 and 41 % saturated $(NH_4)_2SO_4$ yielded only a slight precipitate.

It was possible in some cases to disturb these clusters so that individual crystals were broken away but the resultant crystals remained too small for diffraction studies. To improve the size of these crystals, streak seeding and microseeding experiments were set up. However, single crystals still did not grow and the clusters remained too small (figure 5.17 (b)).

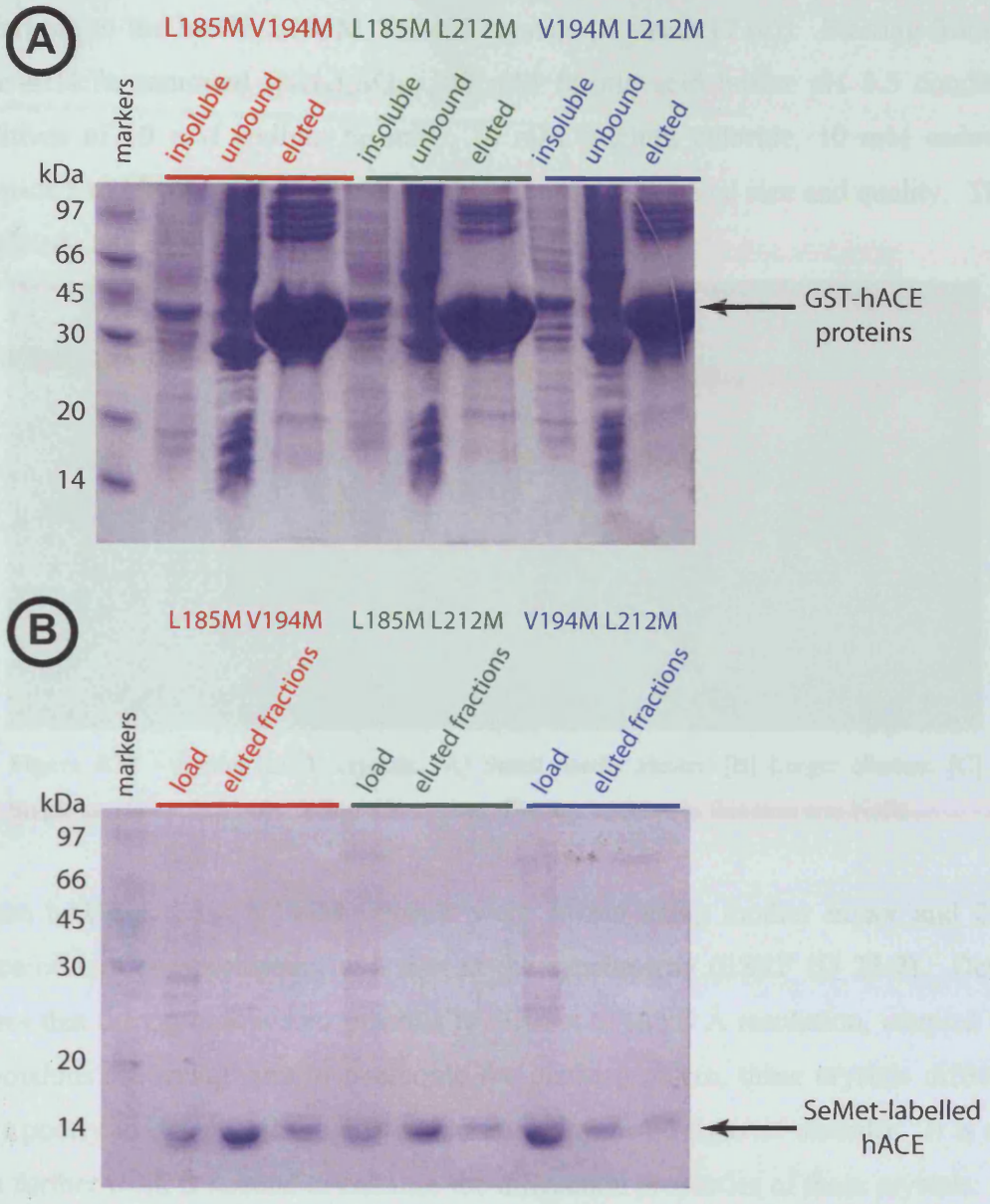

Figure 5.16 – Expression of selenomethionine labelled protein. [A] Initial purification of SeMet protein on glutathione-sepharose. The gel shows insoluble, unbound and glutathione eluted protein for each double mutant. [B] Purification of SeMet by gel filtration. Gel shows the load and the peak eluted fractions after gel filtration on a Superdex 75 column.

Another approach to improve the size and quality of protein crystals is the use of additives. The introduction of divalent cations, detergents, chaotropes, alcohols and other chemicals has been shown to affect crystallisation in many different ways. Screening additives generally involves adding a small amount of each additive to identical crystallisation drops based on the best condition identified thus far and analysing the effect of each additive. It is difficult to "second guess" which additives, if any, will improve crystallisation. In this case, additives made a significant

difference to the hACE L185M V194M crystals (figure 5.17 (c)). Starting from the base 41.4 % saturated $(NH_4)_2SO_4$, 100 mM formic acid buffer pH 3.5 condition, additives of 10 mM sodium bromide, 10 mM calcium chloride, 10 mM cadmium chloride and 10 mM manganese chloride, all improved crystal size and quality. These single crystals were suitable for diffraction studies.

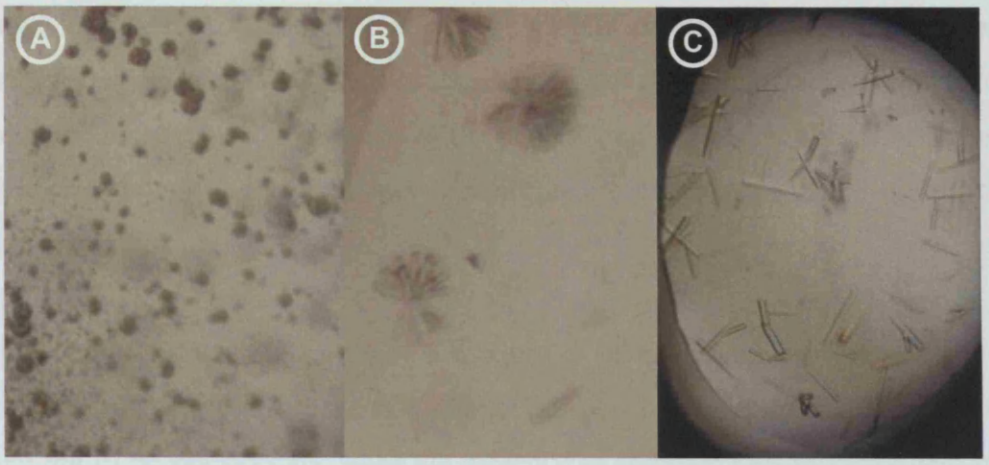

Figure 5.17 – SeMet hACE crystals. [A] Small needle cluster. [B] Larger clusters. [C] Single crystals suitable for diffraction studies. The key additive in this case was NaBr.

These hACE L185M V194M crystals were frozen using mother liquor and 25 % glycerol as a cryoprotectant, and shot at the synchrotron (ESRF ID 23-2). Despite hopes that the crystals would potentially diffract to sub 3 Å resolution, coupled with anomalous scattering data to overcome the phase problem, these crystals diffracted very poorly to ~ 8 Å and the diffraction images showed signs of disorder. It is clear that further work is needed to enhance the diffraction properties of these crystals.

5.7 Identification of a novel crystallisation conditions for the hACE domain

Given the difficulties in obtaining sub 3 Å resolution diffraction despite the size and exterior beauty of the original hACE crystals, a new crystallisation condition was sought. Given the amount of screening initially for a condition and the fact that the condition that was found was at an extreme of most crystallisation screens, it seemed reasonable that a change to the protein buffer may increase the likelihood of finding a novel crystallisation condition. Since the original hACE crystals were grown using $(NH_4)_2SO_4$ as a precipitant the protein buffer was modified to contain $(NH_4)_2SO_4$. Given the potential role for this domain in binding DNA, a polyanionic molecule, one could imagine that the sulphate ions present in $(NH_4)_2SO_4$ might be acting as counter-

ions to a DNA binding surface on the protein. Thus the hACE domain might be stabilised by the presence of sulphate ions in the buffer solution.

The hACE protein expression, extraction and purification was the same as normal, except all buffers contained 100 mM $(NH_4)_2SO_4$ and 50 mM NaCl (this gave an ionic strength equivalent to 250 mM NaCl). The protein expressed and purified as for the normal hACE preparations. Gel filtration analysis of these preparations showed more aggregated protein but the majority of the protein was retained in a monodisperse and monomeric state. The protein was concentrated to 10-15 mg ml^{-1} by ultrafiltration before crystallisation.

Crystallisation trials were set using a mixture of commercial sparse matrix and grid screens and the Mosquito robot and a variety of new hits were found. Many of these turned out to be salt crystals, since the presence of sulphate ions in the protein buffer meant that many conditions containing calcium ions produced crystals of calcium sulphate. However, novel protein crystals were obtained in a more neutral pH range 6.5 → 7.5 with low molecular weight PEGs as preciptants (30 % PEG 400, 25 % PEG MME 550).

These new conditions were attractive hits for a number of reasons. Firstly the initial crystals despite being in 200 nl drops were large enough for diffraction studies. Secondly low molecular weight PEGs make excellent cryoprotectants meaning the crystals could be scooped from the drops and flash-frozen for diffraction. Thirdly the neutral pH meant that the protein crystals should be more reactive with a wider range of heavy atoms for phasing.

A single crystal from a 200 nl drop was mounted and frozen directly in the cryo-stream of our home-source X-ray generator. Diffraction up to 3.3 Å was observed. This was the best diffraction seen for this project thus far, despite a non-optimal cryopreparation as shown by the presence of several ice rings (figure 5.18).

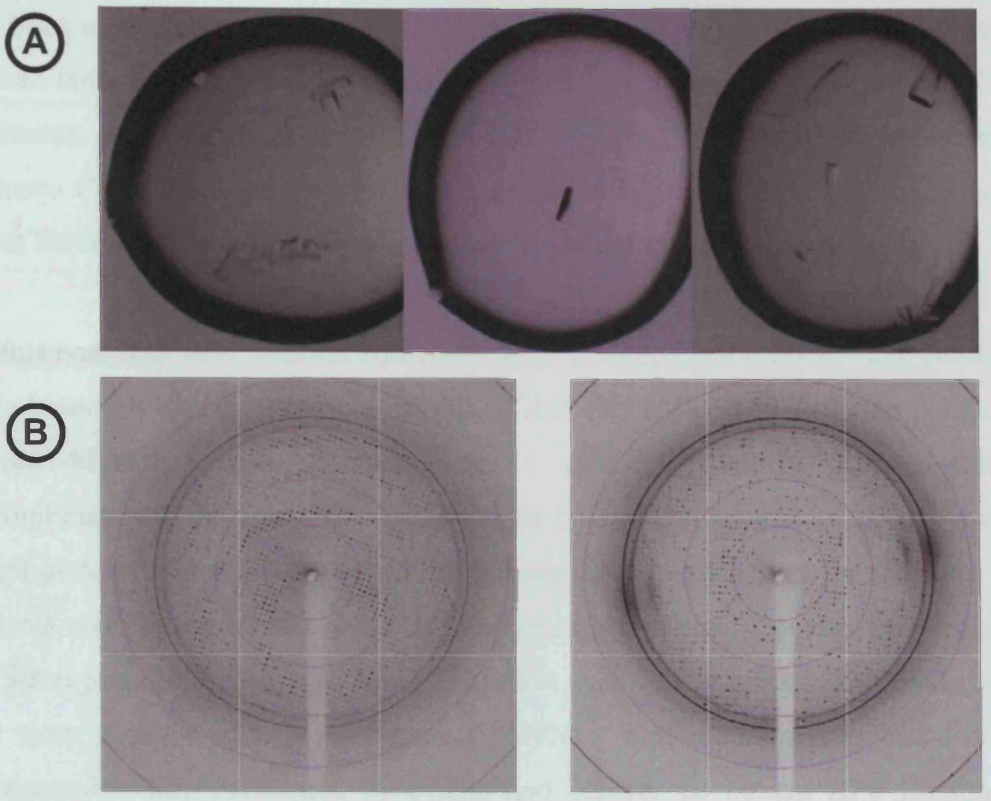

Figure 5.18 – hACE crystals and diffraction. hACE was extracted in a buffer containing $(NH_4)_2SO_4$. [A] Crystals grown using the Mosquito robot. Precipitants were either PEG400 or PEG MME 550. [B] Diffraction data collected at the ESRF ID29. Data extends to 2.9 Å. Circles are at 12, 6, 4, and 3 Å.

However, it proved difficult to reproduce these crystallisation hits in larger drops and with new batches of protein. Crystals from the original 200 nl drops were frozen and taken to the synchrotron (ESRF ID 29). Diffraction data from all the crystals were poor (~4-5 Å) except that for the crystal that was shot on the home-source initially. A dataset from this crystal was collected to 2.9 Å, although the data had a high mosaic spread. This may have been due in part to a poor freeze. Efforts to reproduce these crystals are ongoing.

5.8 Functional analysis of the human Mus81 ACE domain

Whilst efforts to obtain a crystal structure of the hACE domain form the bulk of this chapter, the role of hACE domain in the context of full-length hMus81 is also of interest. A complex of hMus81-hEme1 has been expressed recombinantly in *Escherichia coli* and has been shown to be active against 3'-flap and nicked duplex substrates in particular. The sequence of a second putative hMus81 partner, hEme2,

has been identified previously in the literature (Ciccia et al., 2003). The hMus81-hEme2 complex has recently been characterised as having activity towards 3'-flap substrates, and possibly a greater specific activity than the hMus81-hEme1 complex (Alberto Ciccia et al. manuscript in preparation). However, there is no information about the role of the hACE domain in these complexes.

To this end, four new plasmid constructs were designed and created. The constructs were based on two dicistronic plasmids, pGEX-KG hMus81-hEme1 and pGEX-KG hMus81-hEme2 (These plasmids were a gift from Alberto Ciccia, Genetic Recombination Laboratory, Cancer Research UK. Details about pGEX-KG can be found in Appendix A). As a control for subsequent activity assays, catalytically dead mutants were designed and made for both complexes using appropriate primers and the Stratagene QuikChange protocol. This was achieved by mutating hMus81 residue 339 from aspartic acid to asparagine (D339N). This mutation was based on the mutagenesis study performed by Enzlin and Schärer on human XPF (Enzlin and Scharer, 2002). By extension from that study, the D339N mutation is though to affect the binding of a water molecule involved in nucleophilic attack on the scissile phosphodiester bond, whilst leaving metal binding and the protein fold intact. This round of mutagenesis yielded two new dicistronic plasmids, pGEX-KG hMus81D339N-hEme1 and pGEX-KG hMus81D339N-hEme2. Subsequently, using appropriate primers, overlap PCR was employed to synthesize hMus81 open reading frames devoid of the hACE domain coding sequence. These hMus81ΔACE reading frames were re-cloned into the original dicistronic vectors using the restriction enzymes EcoR1 and Xho1. This yielded two new dicistronic plasmids, pGEX-KG hMus81ΔACE-hEme1 and pGEX-KG hMus81ΔACE-hEme2. In combination with the original wild-type complexes, this gave six plasmid constructs for expression.

Two groups have documented the expression of hMus81-hEme1 in *Escherichia coli* at low levels for functional work and thus this was not expected to be a problem (Ciccia et al., 2003; Ogrunc and Sancar, 2003). Expression of the hMus81 D339N mutant was not expected to be difficult either. However, a little trepidation preceded trials of the hMus81ΔACE constructs, as the effect of deleting the domain could not be foreseen. Expression, extraction and purification conditions were adapted from the

earlier study by Ciccia et al. (2003), except that Triton X-100 was replaced by 0.1 % CHAPS. Purification of these proteins was a three-step process. The first step was affinity chromatography on glutathione-sepharose. Proteins were eluted from this affinity matrix with excess reduced glutathione. These eluates were purified further on heparin-sepharose and Superose 12 columns (figure 5.19). Peak fractions from the superose 12 columns were aliquoted and flash frozen in liquid nitrogen.

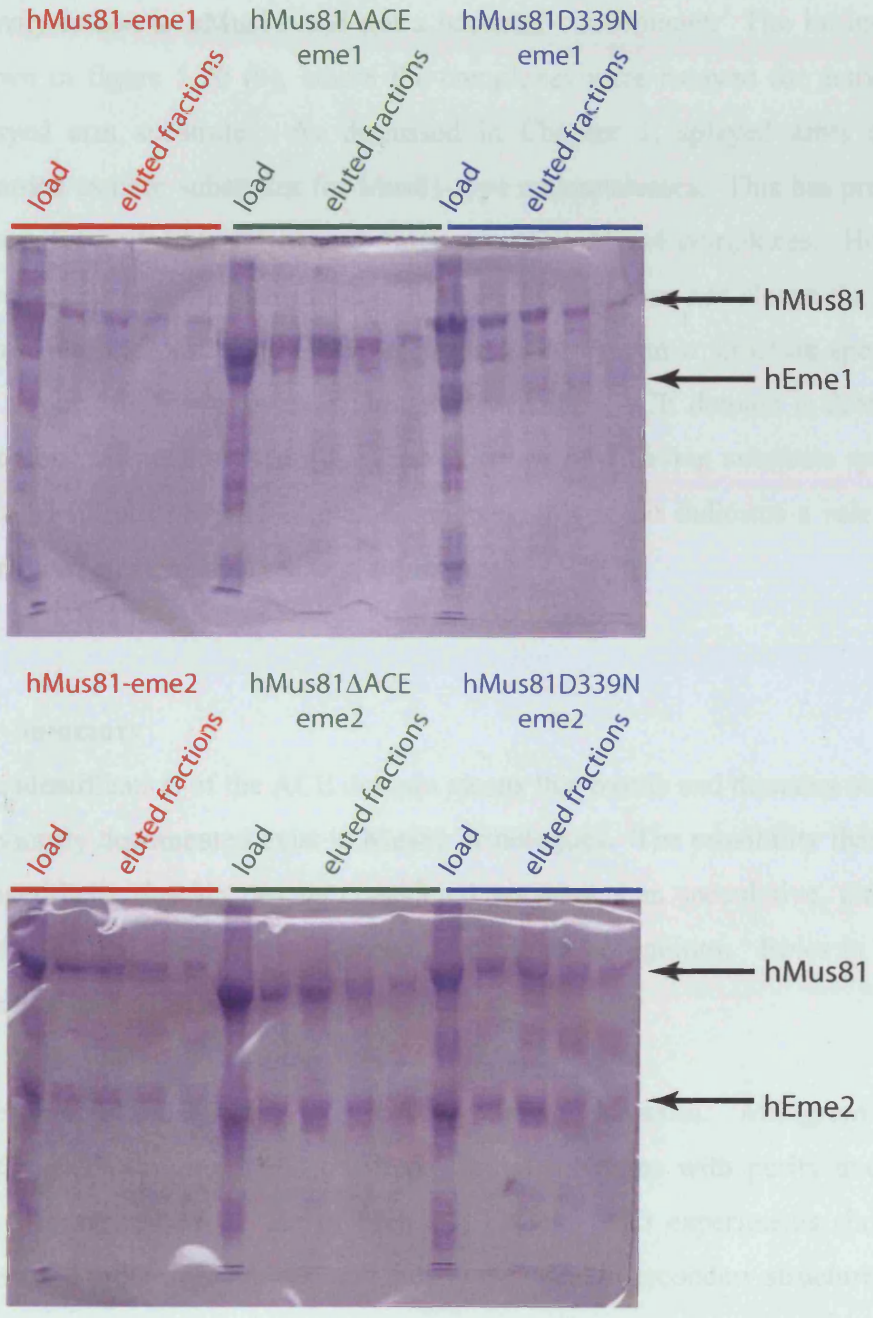

Figure 5.19 – Expression of Mus81complexes. SDS-PAGE gels of fractions taken after gel filtration on a Superose 12 column. The deletion of the ACE domain has a clear effect on hMus81's migration on the gels.

The protein complexes were assayed for endonuclease activity using ^{32}P-labelled substrates by Maureen Biggerstaff (Structural Biology Laboratory). Assay conditions were adapted from Ciccia et al. (2003). Against a 3'-flap substrate, (figure 5.20 (a)), the deletion of the ACE domain appeared to have little effect on hMus81 activity in the context of either hMus81-hEme1 or hMus81-hEme2. The D339N mutation efficiently killed the endonuclease activity and provided support that the observed activity is due to hMus81 and not a bacterial contaminant. The intriguing result is shown in figure 5.20 (b), where the complexes were assayed for activity against a splayed arm substrate. As discussed in Chapter 1, splayed arms are generally regarded as poor substrates for Mus81-type endonucleases. This has previously been shown for both hMus81-hEme1 and scMus81-scMms4 complexes. However, these experiments suggest that although hMus81-hEme1 does not cleave the splayed arm, hMus81-hEme2 efficiently cleaves the splayed arm in a structure-specific manner. On top of this, this activity is abrogated when the ACE domain is deleted from this complex. As well as being the first evidence of differing substrate specificities for hMus81-hEme1/hMus81-hEme2 complexes, this result indicates a role for the ACE domain in cleaving splayed arm substrates.

5.9 Summary

The identification of the ACE domain means that motifs and domains outside of those previously documented exist in Mus81 orthologues. The possibility that Mus81 has a winged-helix domain can be considered no more than speculative, although such a fold would be consistent with a role in substrate recognition. Roles in binding other proteins such as BLM remain possible.

The cloning and expression of hACE proved successful. Milligram quantities of hACE were expressed and purified. Initial problems with purity and aggregation were circumvented by use of high salt buffers. CD experiments showed that the expressed protein was folded and mostly α-helical in secondary structure.

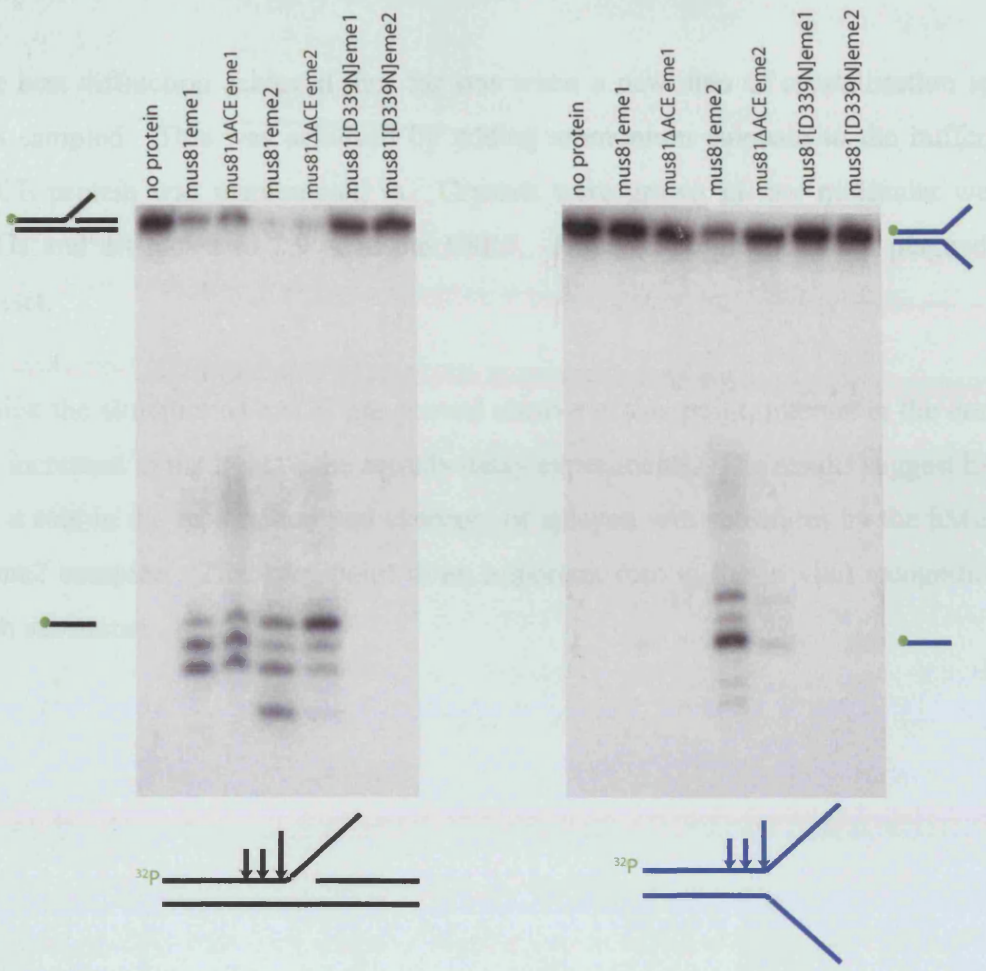

Figure 5.20 – Effect of hACE deletion on hMus81 activity. Autoradiograms showing the activity of Mus81 complexes towards 32-P labelled substrates. [A] 3'-flap structure. [B] Splayed arm. (Substrate sequences can be found in Appendix A).

Crystallisation of hACE protein was achieved. Crystals were grown using ammonium sulphate as a precipitant. Large single crystals were grown, although diffraction did not improve past 3.7 Å.

Whilst trying to improve these crystals by the use of detergents and additives, other approaches were taken to achieve an ACE structure. Attempts to express, purify and crystallise ACE domains from *S.cerevisiae* and *S.pombe* proved more problematic than for the human domain. Tightening domain boundaries at the amino and carboxy termini was only a qualified success with only small crystals grown for the double hACE ΔNΔC protein. The production of mutant selenomethionine-labelled protein was a success but the crystals produced diffracted poorly.

The best diffraction achieved thus far was when a new area of crystallisation space was sampled. This was achieved by adding ammonium sulphate to the buffer the hACE protein was manipulated in. Crystals were grown in low molecular weight PEGs and diffracted to 2.9 Å at the ESRF. However, high mosaicity plagued the dataset.

Whilst the structure of hACE has proved elusive to this point, interest in the domain has increased in the light of the activity assay experiments. The results suggest hACE has a role in the recognition and cleavage of splayed arm substrates by the hMus81-hEme2 complex. This may point to an important role in the in vivo recognition of such substrates.

6 Discussion and Conclusions

The aim of this thesis was to better define XPF/Mus81 architecture and to understand how XPF/Mus81 endonucleases recognise their substrate and the catalytic mechanism underlying the cleavage reaction. Progress towards these aims was made on a number of fronts as described in this thesis.

I - Enhanced yields of human XPF-ERCC1 were obtained using a modified expression vector and an *E.coli* host. This has had a profound impact on several aspects of the human XPF-ERCC1 project even though ultimately we have concluded that producing sufficient XPF-ERCC1 complex in *E.coli* to allow structural analysis is not feasible at present. Sufficient quantities of high quality, unaggregated XPF-ERCC1 allowed us to establish a functional assay for structure-specific endonuclease activity using fluorescently labelled substrates. This in turn allowed us to determine that the C-terminal fragment (629-905) of XPF bound to ERCC1 retains no detectable endonuclease activity. This implies that the helicase-like domain at the extreme N-terminus of XPF may play a role in catalysis or substrate engagement. From this we conclude that efforts to trap a full stem-loop substrate bound to a catalytically dead mutant of XPF-ERCC1 will require the helicase-like domain. Even with the availability of nano-drop dispensing crystallisation robots, a structural analysis of full-length hXPF-hERCC1 remains a very ambitious aim at this time. Despite optimisation, yields are still low and the protein remains susceptible to proteolysis. Alternative strategies at this point could involve screening further homologues of XPF-ERCC1 from other species. Preliminary efforts to express the *S.pombe* equivalent, the Swi10-rad16 complex, did not show any improved yields.

It was possible to produce a truncated form of XPF-ERCC1 with good yields, though as stated, the complex was not active. However, domains contributing to DNA-binding were present and we therefore undertook experiments to co-crystallise ERCC1-XPF (629-905) bound to various oligonucleotides and to XPA. ERCC1-XPF (629-905) had a tendency to aggregate suggesting further optimisation of buffer conditions is required and perhaps further truncations may be required. For example, ~ 40 residues before the start of the nuclease domain may be unstructured. Similarly,

from the crystal structure of the nuclease-like domain of ERCC1 (Tsodikov et al., 2005), residues 1-96 of ERCC1 are also unstructured and are not necessary for activity. A similar XPF-ERCC1 construct reported in the study of Tsodikov et al. suggests that XPF(655-905)-ERCC1(96-297) retained modest structure-specific activity, though this study required overnight incubations to observe any activity. Tsodikov and colleagues were unable to crystallise this truncated form of XPF-ERCC1.

Further modifications of full-length XPF-ERCC1 could improve yields further. Although with hindsight, changing three basic loops simultaneously in XPF may have been too ambitious, more focussed efforts to produce individual basic loop mutations could eliminate proteolysis and could conceivably enhance expression levels. The functional assay could be used to check whether loss of a given basic loop affected catalytic activity. Using the dicistronic vector described in this thesis, milligram quantities of recombinant XPF-ERCC1 were stockpiled and were sufficient to allow a screening project with Cancer Research Technology to be set up to identify an inhibitor of the human enzyme.

It had been hoped that this inhibitor could be co-crystallised with the *A.pernix* XPF nuclease domain. However, such a co-crystallisation experiment is contingent on the inhibitor of hXPF-ERCC1 not being so specific that it does not inhibit other XPF homologues. Indeed, preliminary data from Maureen Biggerstaff (Structural Biology Laboratory) has shown that the putative inhibitor does not appear to inhibit ApeXPF-ApePCNA nuclease activity or the activity of hMus81-hEme1 or hMus81-hEme2 complexes. If these preliminary results are borne out, it is unlikely that co-crystallisation trials of the inhibitor with ApeXPF nuclease domain would be successful.

Recent data by Tsodikov et al. (2005) described a similar reductionist approach to crystallise the ERCC1 nuclease-like domain and the heterodimer of XPF-ERCC1 $(HhH)_2$ domains. They were unable to crystallise the nuclease domain of XPF thus far and this perhaps remains an outstanding objective since, if inhibitors of human XPF-ERCC1 can be found that are competitive with substrate, then there will be a need for such a structure to contribute to a rational drug design program. Constructs coding for

the hXPF nuclease domain (residues 671-815) have been cloned to facilitate structure studies on a relevant nuclease:inhibitor complex.

II – During this thesis, our laboratory published results for the apo- and DNA-bound conformation of the *A.pernix* XPF (Newman et al., 2005). This structure contained the nuclease and $(HhH)_2$ domains as well as the connecting linker. The determination of the structure of ApeXPF complexed with DNA was not straightforward and required heavy atom derivatives and molecular replacement phases using a partially refined structure of the ApeXPF nuclease domain. Thus the production, crystallisation and data collection of selenomethionine-labelled ApeXPF nuclease domain described in this thesis made a significant contribution to the eventual structural analysis of an essentially intact ApeXPF (residues 18-231). Solving, building and refining the nuclease domain at 2.1 Å resolution provided an important learning experience and provided a structural platform for co-crystallisation with small molecule inhibitors that may be used in the future. Future experiments may also address the position of the scissile phosphate in the active site, whether the mechanism is truly a single-metal reaction centre and where the activated nucleophilic water is positioned in relation to other invariant XPF/Mus81 sidechains.

Disappointingly, efforts to co-crystallise ApeXPF with 3' flap substrates were unsuccessful. It is well documented that protein-DNA complexes can be very sensitive to the precise length and nature of the DNA fragments used. Whilst several crystals were grown in the presence of DNA, no diffraction was obtained suggesting further exploration of different DNA substrates may yet be warranted. A similar finding was made with ApeXPF in that diffraction quality crystals could be grown using oligonucleotides of 15 bases in length, whilst shorter or longer oligos gave crystals but no diffraction.

Our own data, combined with recent DNA footprinting data from Kosuke Morikawa's group in Osaka (Nishino et al., 2005b), suggests that model 2, in which DNA substrate is bent by roughly 90 degrees to allow it to engage both $(HhH)_2$ domains, may be an integral part of structure-specific recognition of double-strand/single-strand DNA junctions by XPF/Mus81. The preferred 3'-flap substrates for ApeXPF provides an upstream and downstream duplex that can be engaged by XPF. XPF-ERCC1 has a

different substrate preference, instead targeting stem-loop or splayed arm substrates. This provides a single region of duplex DNA that suggests a more elaborate and distinctive mode of substrate recognition. Model 2 also implies that ERCC1 may mimic the "inactive" subunit of ApeXPF that is proposed to engage the upstream duplex of a 3'-flap. If true, this could argue that ERCC1 plays an important role in substrate recognition and not just simply to stabilise the XPF subunit as previously proposed.

III – In order to produce more physiologically relevant complexes containing XPF, it was thought necessary to explore producing and reconstituting the *A.pernix* PCNA heterotrimer. By analogy to other crenarchaeal PCNA complexes, we wished to investigate whether the PCNA complex indeed stimulated XPF activity. An equivalent PCNA heterotrimer from *S. solfataricus* had been previously produced successfully at high yields by Bell and colleagues in Cambridge (Dionne et al., 2003). However, others had little success in crystallising the *S. solfataricus* XPF homologue. Moreover, the SsoPCNA only weakly stimulated ApeXPF catalytic activity, and it was felt that many groups had already initiated crystal screens using the SsoPCNA. Therefore some effort was made to establish an expression system for ApePCNA in *E.coli* and this proved ultimately to be successful. Since XPF is a homodimer and PCNA a heterotrimer, this suggested only two of the three subunits could be engaged by XPF. We found that under appropriate conditions, all three subunits co-purified with tagged XPF suggesting that the PCNA ring had indeed been reconstituted. More evidence that the PCNA complex forming a closed ring has come from preliminary negative-stained images measured by electron microscopy (Dr. Hannes Ponstingl and Prof. Helen Saibil at Birkbeck College). These clearly show a doughnut shaped object with associated additional density, potentially corresponding to XPF or XPB subunits.

Using the strategy established in Chapter 4, John Lally (Structural Biology Laboratory) has recently produced crystals of an ApePCNA-ApeXPFΔN complex (figure 6.1). Although the crystals are currently unsuitable for diffraction studies, they provide a promising lead and a structure of this complex would be hugely informative with regards to the activation of ApeXPF by ApePCNA. A functional assay for ApePCNA-ApeXPF has also been established in the laboratory for the first time using protein produced by this thesis work. Activity can be measured at 55 °C

indicating that it is now possible to conduct structure-function analysis of XPF-PCNA complexes for *A. pernix* for the first time. Future efforts could probe which PCNA subunits contribute to XPF stimulation together with structural efforts to trap the entire ApePCNA-ApeXPF complex. Forming complexes with DNA remains a major obstacle given that a PCNA-DNA complex has so far eluded many groups.

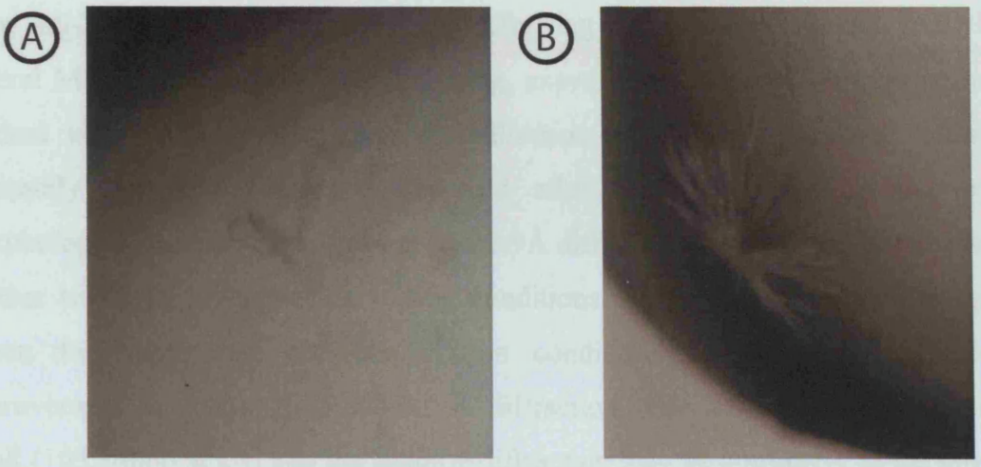

Figure 6.1 – Crystals of an ApePCNA-ApeXPF complex. Crystals grown in sitting-drop by vapour diffusion.

We included XPB in the XPF-PCNA preparations as we reasoned that the third PCNA subunit could bind a PIM-containing protein and hypothesised that XPB may play such a role. Furthermore, preliminary data from Professor Malcolm White (St. Andrew's University) suggested that *S. solfataricus* XPB could bind to PCNA (White, M.F. unpublished data). This would be in keeping with the notion of a PCNA "toolbelt", whereby multiple enzymes are loaded onto PCNA and each can swing into action as required. Future efforts will investigate the effect of PCNA on XPB helicase activity and whether the XPB helicase activity in turn affects XPF activity. One report has already eluded to XPB regulating XPF activity in human cells since XPB phosphorylation by casein kinase at its C-terminus appears to negatively regulate the incision carried out by XPF. If and how *A. pernix* XPB controls XPF activity remains to be shown and requires further validation.

IV - Perhaps the most frustrating aspect of this work has been the inability to determine the structure of the hMus81 ACE domain within the timeframe of this

thesis. The bioinformatic identification of the conserved ACE domain, unique to Mus81, suggested it was globular and would be a good crystallisation target. It was unclear from the sequence alignments whether the ACE domain would be involved in protein-protein interaction or whether it contributes towards DNA substrate specificity. Thus, there was a need for structural analysis to verify the overall fold of the domain and a functional analysis to elucidate the effect of deleting the domain from full-length Mus81, as well as characterising the individual domain itself from several Mus81 homologues. Initial cloning, expression and purification experiments worked very smoothly and early crystallisation hits looked promising. However ultimately the best diffraction observed after screening many related crystal morphologies and conditions gave at best 2.9Å diffraction with a large mosaic spread. Further tweaking of crystal harvesting conditions, dehydration and more extensive screen for appropriate cryo-preservation conditions may provide a sufficient improvement to obtain good sub-3.0 Å diffraction. The ACE domain is relatively small (103 amino acids) and the limiting diffraction may be a property of the intrinsic flexibility within the protein or alternatively poor lattice contacts within the crystal form. To identify regions of flexibility, we investigated carrying out NMR experiments in collaboration with Dr. Richard Harris at University College. At this point it was realised that the sample was of sufficient quality to have its structure determined by nuclear magnetic resonance (NMR). hACE protein has been expressed in *E.coli* grown in minimal media with $(^{15}NH_4)_2SO_4$ as the only nitrogen source. Figure 6.2 shows NMR spectra for the hACE domain. The spectra are good, with the 2D-15N NOESY spectrum containing an appropriate number of cross-peaks for the non-proline main chain amides. The preparation of a double ^{15}N-^{13}C labelled sample is also planned to aid structure determination in the near future. The expectation is that an NMR structure for the ACE domain would assist in determining the crystal structure. This could reveal the location of multiple sulphate ions required for crystal growth that may mimic interaction with a DNA phosphodiester backbone. We would then be in a position to mutagenise specific sidechains on the Mus81 ACE domain and probe the effect of such a mutation in the full-length enzyme.

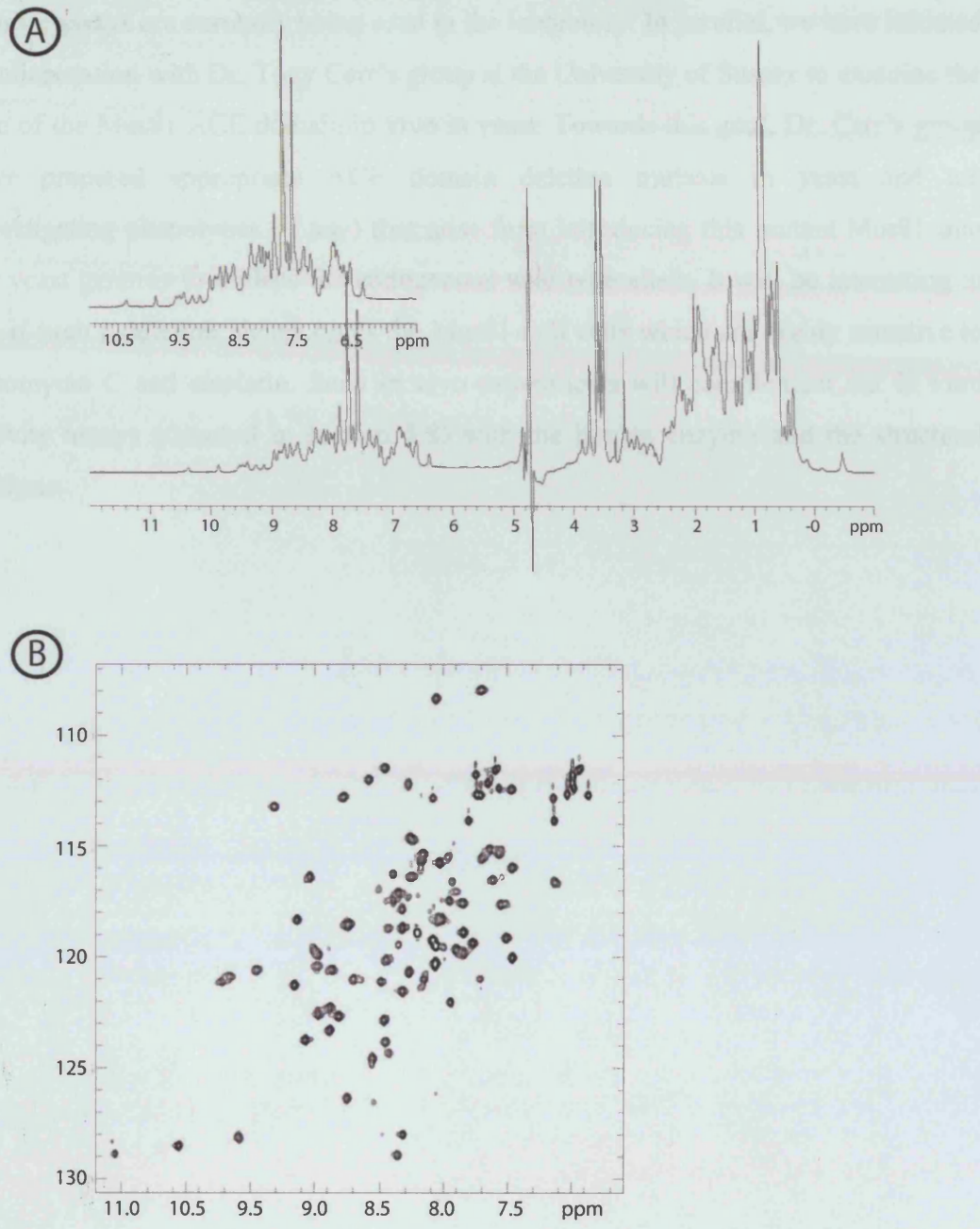

Figure 6.2 – hACE NMR spectra. [A] 1D- proton spectrum taken using unlabelled hACE protein. [B] 2D – ^{15}N NOESY spectrum. There are around 90 crosspeaks, an appropriate number for the hACE domain.

The availability of a gel-based assay for endonuclease activity allowed Mus81 complexes to be assayed using radiolabelled substrates. Preliminary work on Mus81-Eme1, as well as Mus81-Eme2 complexes, either with or without an ACE domain deletion, suggested a role in substrate interaction. Future efforts could therefore examine whether the ACE domain is able to bind DNA by employing a fluorescence polarisation assay to detect binding of single-stranded or double-stranded DNA.

Similar assays are currently being used in the laboratory. In parallel, we have initiated a collaboration with Dr. Tony Carr's group at the University of Sussex to examine the role of the Mus81 ACE domain in vivo in yeast. Towards this goal, Dr. Carr's group have prepared appropriate ACE domain deletion mutants in yeast and are investigating phenotypes (if any) that arise from introducing this mutant Mus81 into the yeast genome to replace the endogenous wild type allele. It will be interesting to see if such a deletion phenocopies the Mus81-null cells which are highly sensitive to mitomycin C and cisplatin. Such in vivo experiments will complement our in vitro activity assays (detailed in section 5.8) with the human enzyme and the structural analysis.

Appendix A Materials and Methods

A.1 Reagents and enzymes

Reagents were purchased from Sigma-Aldrich or Merck unless otherwise stated and were of the highest grade available. Enzymes used in molecular biology protocols were purchased from New England Biolabs (NEB).

A.2 Buffers, solutions and media

Common buffers and media used are described here.

Name	Composition
SDS-PAGE running buffer (10 X)	15.1 g Tris base, 94 g glycine, 50 ml 10 % SDS, up to 1000 ml with water
SDS-PAGE sample buffer	2 % SDS, 80 mM Tris-HCl (pH 6.8), 10 % glycerol, 0.02 % bromophenol blue, 2 mM 2-mercaptoethanol before use.
TAE buffer	40 mM Tris-acetate, 2 mM EDTA
TBE buffer	90 mM Tris-borate, 2 mM EDTA
DNA gel sample buffer (6 X)	0.25 % bromophenol blue, 0.25 % xylene cyanol FF, 30 % glycerol.
Luria-Bertani (LB) broth media	10 g NaCl, 10 g bacto-tryptone, 5 g bacto-yeast extract, made up to 1000 ml with water. Autoclaved.
LB-Agar	6 g agar added to 400 ml LB. Autoclaved.
SOC broth media	20 g bacto-tryptone, 5 g bacto-yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, 20 mM D-glucose; made up to 1000 ml with sterile water and autoclaved.

Table A.1 – Common buffers and media

A.3 Molecular cloning

A.3.1 Donated constructs

The following constructs were kindly donated for this work:

Construct	Donated by -
pET30b XPF-ERCC1 dicistronic	Prof Rick Wood/Dr. Matthew Newman
pET30b XPF	Prof Rick Wood/Dr. Matthew Newman
pET30b ERCC1	Prof Rick Wood/Dr. Matthew Newman
pET30b ERCC1-XPF (629-905)	Dr. Matthew Newman
pET15b XPA	Annabel Borg
pET14b-3C ApeXPF (18-231)	Dr. Matthew Newman
pET14b-3C ApeXPF (18-150)	Dr. Matthew Newman
pET28a hMus81-hEme1 dicistronic	Alberto Ciccia
pGEX-KG hMus81-hEme1 dicistronic	Alberto Ciccia
pGEX-KG hMus81-hEme2 dicistronic	Alberto Ciccia

Table A.2 – List of donated constructs

A.3.2 Plasmid vectors

The plasmid vectors used for protein expression were either from the pET (Novagen) or pGEX (Amersham) family of expression vectors. Figure A.1 shows the cloning and expression regions of three non-standard vectors used in this thesis; pET14b-3C, pET41a-3C and pGEX-KG.

A.3.3 Oligonucleotides

Oligonucleotides were synthesised by the Cancer Research UK Oligonucleotide Synthesis Department and purified by ethanol precipitations. Oligonucleotides for site-directed mutagenesis, crystallisation or those longer than 45 nucleotides were purified by HPLC. After closure of this facility in July 2004, all oligonucleotides were purchased from Sigma-Genosys.

A.3.4 Plasmid purification

Plasmid DNA was purified using Qiagen miniprep or maxiprep kits. Plasmid concentrations were calculated from absorption measurements at 260 and 280 nm using either a Pharmacia Biotech Ultrospec2000 spectrophotometer or a NanoDrop ND-1000 spectrophotometer.

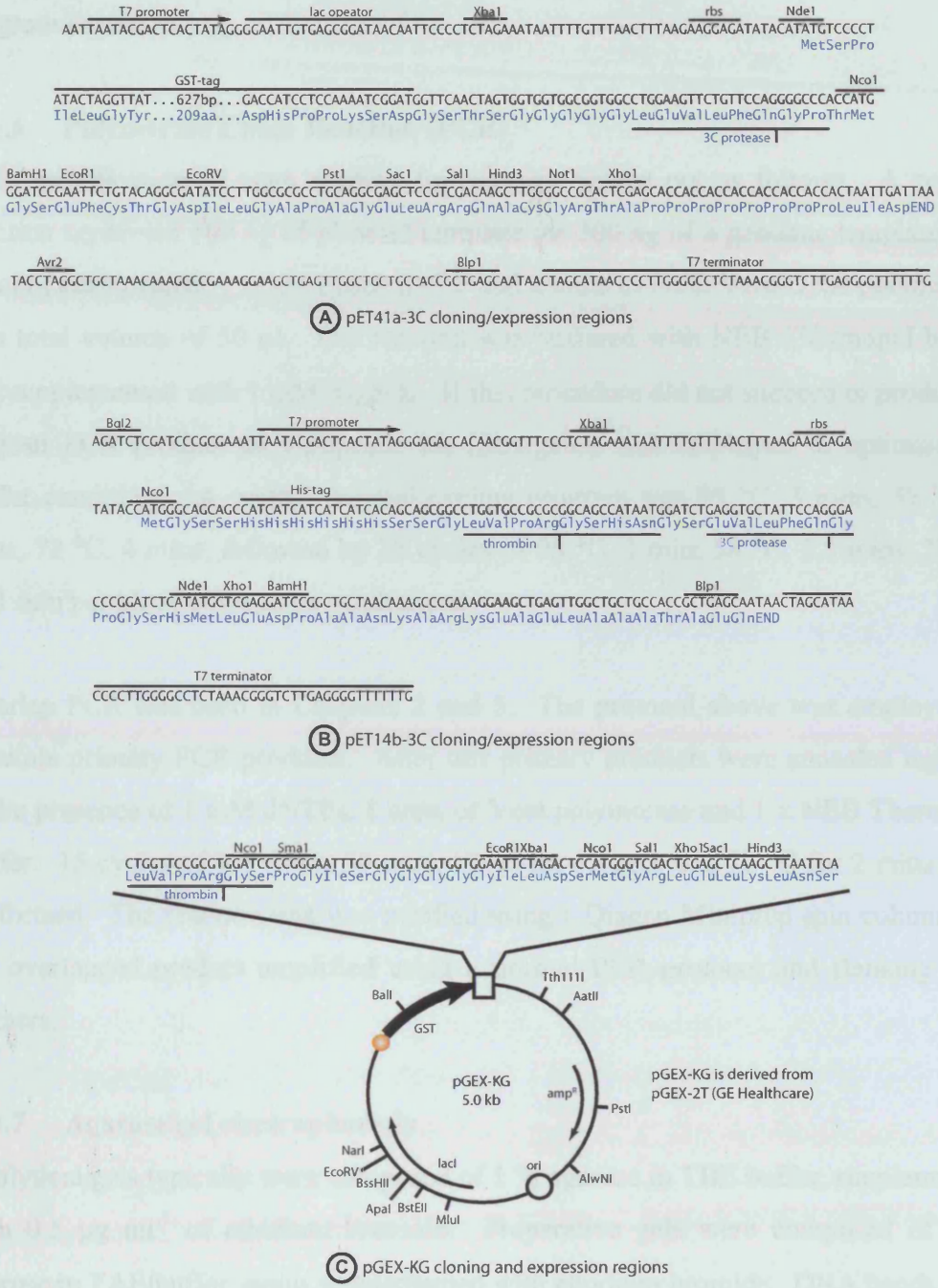

Figure A.1 – Cloning and expression regions of non-standard plasmid vectors used in this thesis.

A.3.5 DNA sequencing

Plasmid DNA was sequenced using the dideoxynucleotide chain terminating inhibitor method (Sanger et al., 1977) and the BigDye Terminator Cycle Sequencing kit (PE Applied Biosystems). Sequencing reactions were cleaned up using Qiagen DyeEx Sequencing kits and ran on an ABI Prism 3730 Capillary Sequencing Unit (Cancer

Research UK Equipment Park). Sequence data was analysed using the freeware program FinchTV.

A.3.6 Polymerase Chain Reaction (PCR)

PCR amplification of open reading frames was carried out as follows. A typical reaction contained 100 ng of plasmid template (or 200 ng of a genomic template), 10 pmol of each primer, 1 mM of each dNTP and 2 units of NEB Vent DNA polymerase in a total volume of 50 µl. The reaction was buffered with NEB Thermopol buffer and supplemented with 1 mM $MgSO_4$. If this procedure did not succeed in producing a clean PCR product an Optiprime kit (Stratgene) was employed to optimise the buffer conditions. A typical thermal cycling program was 95 °C, 5 mins; 58 °C, 4 mins, 72 °C, 4 mins; followed by 25 cycles of 95 °C, 1 min; 58 °C, 1.5 mins, 72 °C, @ 1 min per kb of DNA to be synthesised.

Overlap PCR was used in Chapters 2 and 5. The protocol above was employed to generate primary PCR products. After this primary products were annealed together in the presence of 1 mM dNTPs, 2 units of Vent polymerase and 1 x NEB Thermopol buffer. 15 cycles of 95 °C for 30 secs, 40 °C for 15 secs and 72 °C for 2 mins were performed. The reaction mix was purified using a Qiagen Miniprep spin column and the overlapped product amplified using a normal PCR protocol and flanking PCR primers.

A.3.7 Agarose gel electrophoresis

Analytical gels typically were composed of 1 % agarose in TBE buffer, supplemented with 0.5 µg ml^{-1} of ethidium bromide. Preparative gels were composed of 1 % agarose in TAE buffer, again supplemented with ethidium bromide. DNA bands were visualised on a UV-transilluminator, cut out with a fresh razor blade and the DNA extracted using Ultrafree-DA spin columns (Millipore). Further purification was achieved using the Qiagen miniprep kit.

A.3.8 Restriction digests

Restriction digests were performed using NEB enzymes with the recommended buffers for either single or double digests. Typically 1-2 units of enzyme were used per μg of DNA.

A.3.9 Ligation

Ligation reactions typically consisted of 50 ng of cut plasmid vector, a 3-5 fold molar excess of insert and 20 units of NEB T4 DNA ligase, buffered with T4 DNA ligase buffer, in a total volume of 10 μl, The reaction was left to incubate at room temperature for 1-3 hours and used to transform either DH5α (Gibco) or NovaBlue (Novagen) competent cells. Control transformations with no DNA, cut plasmid vector but no insert and uncut plasmid were used to assess the efficiency of ligation and transformation.

A.3.10 Transformation of chemically competent *Escherichia coli cells*

Frozen competent cells were thawed on ice and the transforming DNA was added to the cells and mixed gently. This mixture was incubated on ice for 30 minutes. The cells were transformed by heat shock for 1 minute at 42 °C and returned to ice for 2 minutes. If the plasmid of interest conferred ampicillin resistance, then the cells were plated out immediately. If the plasmid of interest conferred kanamycin resistance, then 900 μl of SOC media was added and the cells incubated at 37 °C for 1 hour before plating out. Selection of transformants was achieved by plating the cells on to LB-Agar plates containing the appropriate antibiotic and incubation at 37 °C overnight for colony formation.

A.4 Protein Expression

A.4.1 *Escherichia coli* host strains

Frozen stocks of electrocompetent cells of the following strains were used for protein expression described in this thesis.

A.4.2 Transformation of electrocompetent *Escherichia coli* cells

Frozen electrocompetent cells were thawed on ice and the transforming DNA was added to the cells and mixed gently. This mixture was then transferred to a cooled

EquiBio 2 mm electroporation cuvette. The cuvette was placed in a BioRad GenePulser electroporator and pulsed at 2.5 kV, 25 μF and 200 Ω. 1 ml of SOC media was added to the cuvette immediately. If the plasmid of interest was ampicillin resistant then the cells were plated out immediately. If the plasmid of interest conferred kanamycin resistance then the cuvette was incubated at 37 °C for 30 mins. Selection of transformants was achieved by plating the cells on to LB-Agar plates containing the appropriate antibiotic and incubation at 37 °C overnight for colony formation.

Name	**Antibiotic^R**	**Source**
FB810 (BL21(DE3) recA$^-$)	None	Lab stock
Rosetta	Cam	Novagen
Rosetta pLysS	Cam	Novagen
BL21-AI	None	Invitrogen
BL21-AI codon +	Cam	Invitrogen

Table A.3 – *E.coli* strains used for protein expression. Cam = chloramphenicol

A.4.3 Bacterial growth measurement

The growth of *Escherichia coli* cultures for protein expression was monitored by measuring the absorbance of an aliquot of the culture at 600 nm in a spectrophotometer, compared to a blank with just LB media.

A.4.4 Expression of recombinant protein in *Escherichia coli*

This section provides a framework around which expression experiments were built.

A single colony from a fresh transformation plate was used to inoculate 100 ml of LB media supplemented with the appropriate antibiotic(s) for the plasmid of interest. This starter culture was grown overnight with shaking at 37 °C. In the morning the starter culture was harvested by centrifugation and the supernatant discarded (this was to minimise transfer of β-lactamase to the large cultures). The pellet was then resuspended in 20 ml of fresh LB media and of this 2 ml was used to inoculate 1 l of

LB media (in a 2 l baffled flask) again supplemented with the appropriate antibiotic(s). These flasks were generally incubated with shaking at 37 °C until the optical density at 600 nm (OD_{600}) of the culture was greater than 0.6.

At this point the induction procedure was initiated. One of four induction protocols was used, depending on the protein of interest.

[A] The flasks remained at 37 °C and protein expression was induced by the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 250 μM. The cells were then incubated for a further 3 hours.

[B] The flasks remained at 37 °C and protein expression was induced by the addition of IPTG to a final concentration of 250 μM, plus L-arabinose to a final concentration of 0.2 %. The cells were then incubated for a further 5 hours.

[C] The flasks were cooled to 25 °C for 30 mins, and protein expression was induced by the addition of IPTG to a final concentration of 25 μM. The cells were then incubated for a further 3-4 hours at 25 °C.

[D] The flasks were cooled to 18 °C for 30 mins, and protein expression was induced by the addition of IPTG to a final concentration of 25 μM. The cells were then incubated overnight at 18 °C.

After induction the cells were harvested in a Beckman J-6B centrifuge at 4000 x g at 4 °C for 20 minutes. The cell pellets were re-suspended in 35 ml of fresh LB and harvested by centrifugation in a Beckman GS-6R centrifuge at 4000 x g at 4 °C for 20 minutes. The supernatant was discarded and the cell pellets stored at -80 °C.

Table A.4 shows the respective induction protocol for each protein or protein complex described in this thesis.

A.4.5 Expression of selenomethionine-labelled recombinant protein in *Escherichia coli*

(This section refers specifically to the production of selenomethionine-labelled (Se-Met) hACE domain protein discussed in section 5.6).

Protein expressed	Protocol	Notes
Full-length XPF-ERCC1	D	
ERCC1-XPF (629-905)	C	
GST-XPA	C	+ 10 mM $ZnSO_4$ in induction
GST-XPF – ERCC1	D	
ERCC1-XPFΔloops	D	
ApeXPFΔNΔC (18-231)	C*	
ApeXPF nuclease (18-150)	C*	
ApePCNA 1	B	
ApePCNA 2	B	
ApePCNA 3	B	
ApePCNA 1,2,3	B	
Full-length ApeXPF	C*	
ApeXPFΔN(18-254)	C*	
ApeXPF$(HhH)_2$ & PIM (164-254)	A	
ApeXPB	C	
hMus81 ACE	A	
scMus81 ACE	A	
spMus81 ACE	A	
hMus81 ACE ΔN	A	
hMus81 ACE ΔC	A	
hMus81 ACE ΔNΔC	A	
GST-hMus81-hEme1	D	
GST-hMus81ΔACE-hEme1	D	
GST-hMus81-hEme2	D	
GST-hMus81ΔACE-hEme2	D	

Table A.4 – Induction protocol for proteins and protein complexes described in this thesis. C* - Slow growth of these strains meant sometimes an overnight induction at 25 °C was used.

In order to express Se-Met labelled hACE protein, it was necessary to use a methionine auxotrophic *Escherichia coli* strain, in this case B834(DE3) cells (Novagen). The SelenoMet Media kit from Molecular Dimensions was used to provide a minimal methionine-free broth. This kit involved making two solutions; the first an autoclaved medium base containing the 19 other amino acids, the second a 0.22 μm filtered nutrient mix containing glucose, ferric chloride and other salts. Protein expression experiments were conducted as described above in section A.4.4 with the following changes. Overnight 100 ml starter cultures were grown in the minimal media described, supplemented with a small amount of regular methionine (50 $\mu g\ l^{-1}$). These starter cultures were harvested and the pellets washed extensively with water. The main cultures were grown using the minimal media described supplemented with a small amount of Se-Met (50 $\mu g\ l^{-1}$). To improve yields 500 ml cultures were grown in 2 l baffled flasks.

A.5 Protein purification and analysis

A.5.1 Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE was performed accordingly to the method of Laemmli (Laemmli, 1970). 12 % resolving gels were generally used to visualise all proteins and consisted of 12 % (w/v) acrylamide, 0.3 % (w/v) bisacrylamide, 300 mM Tris-HCl (pH 8.8), 0.1 % (w/v) SDS, 0.003 % tetramethylethylenediamine (TEMED) and 0.0003 % (w/v) ammonium persulphate (APS). The 3.8 % stacking gel used consisted of 3.8 % (w/v) acrylamide, 0.1 % (w/v) bisacrylamide, 125 mM Tris-HCl (pH 6.8), 0.1 % (w/v) SDS, 0.003 % TEMED and 0.0003 % APS.

The TEMED and APD were added last as they induce the polymerisation of the acrylamide. Gels were cast and run using apparatus purchased from Amersham Pharmacia Biotech. Samples to be analysed by SDS-PAGE were resuspended in SDS-sample buffer and incubated at 95 °C for 5 minutes. Gels were run in 1 x SDS-PAGE running buffer at a constant voltage of 100 V whilst the samples permeated the stacking gel, and subsequently 200 V as the samples ran through the resolving gel. Electrophoresis was continued until the bromophenol blue dye front had just run off the bottom of the gel.

Gels were stained at room temperature with Coomassie Blue stain (0.1 % (w/v) Coomassie PhastBlue (Pharmacia), 30 % (v/v) methanol, 10 % (v/v) glacial acetic acid). Gels were destained at room temperature in destain solution (45 % (v/v) methanol and 10 % (v/v) glacial acetic acid).

A.5.2 Protein concentration

Protein concentrations were determined in one of two ways. The first was by Bradford assay (Bradford, 1976) using an appropriate reagent from BioRad and bovine serum albumin to construct a standard curve. This assay is inexpensive and sensitive, although the results with different proteins can vary widely (especially for arginine rich proteins). The second was using absorption measurements at 280 nm from a NanoDrop ND-1000 spectrophotometer and theoretical molar extinction coefficients calculated from the Expasy ProtParam server (http://www.expasy.org/tools/protparam.html). This method is rapid and uses very little protein, although it is susceptible to contamination.

A.5.3 Purification of His_6-tagged proteins

The following sections provide a framework around which purification experiments were built. Tables A.5, A.6, A.7 and A.8 detail the buffers that were used in the purification of proteins described in this thesis (unless alternative buffers are given explicitly in the text).

A.5.3.1 Batch purification

Frozen pellets were thawed at room temperature and resuspended in lysis buffer, supplemented with 10 mM benzamidine and 1 mM PMSF (both protease inhibitors). The cells were sonicated on ice using a MSE Soniprep 150 sonicator at 12 μm. 6 x 30 secs bursts were employed with rest periods on ice of at least 60 secs and end-over-end mixing. Cell debris was removed by centrifugation in a Beckman Allegra 64R centrifuge at 29,200 x g, for 30 mins at 4 °C. The supernatant was mixed on a roller shaker at 4 °C for 90 mins with Ni^{2+}-NTA affinity resin (Qiagen) pre-washed in lysis buffer (typically 1 ml of Ni^{2+}-NTA was used for 10 mg of protein expected). The resin was collected from the unbound fraction in a column and washed thoroughly prior to elution. Elution was achieved either by the addition of imidazole-

containing elution buffer or by addition of 3C protease to a slurry of the Ni^{2+}-NTA resin in lysis buffer.

Depending on the experiment in hand, proteins were dialysed overnight to remove the imidazole and then 3C protease was added to cleave the protein at 4 °C overnight. SDS-PAGE was used to monitor the efficiency of purification and cleavage.

Protein expressed	Lysis buffer	Wash buffer	Elution buffer
Full length XPF-ERCC1	20 mM HEPES (7.4) 500 mM NaCl, 2 mM 2-ME, 2mM $MgCl_2$, 10% glycerol, 0.01% CHAPS, 5 mM imidazole.	Lysis + 20 mM imidazole	Lysis + 250 mM imidazole
ERCC1-XPF (629-905)	50 mM Tris (8), 300 mM NaCl, 2 mM 2-ME	Lysis + 20 mM imidazole	Lysis + 20 mM imidazole
ERCC1-XPF Δloops	20 mM HEPES (7.4) 500 mM NaCl, 2 mM 2-ME, 2mM $MgCl_2$, 10% glycerol, 0.01% CHAPS, 5 mM imidazole.	Lysis + 20 mM imidazole	Lysis + 20 mM imidazole
ApeXPFΔNΔC (18-231)	20 mM Tris (8), 500 mM NaCl, 2 mM 2-ME	Lysis + 20 mM imidazole	Lysis + 250 mM imidazole
ApeXPF nuclease (18-150)	20 mM Tris (8), 500 mM NaCl, 2 mM 2-ME, 0.5 % CHAPS	Lysis + 20 mM imidazole	Lysis + 250 mM imidazole
Full length ApeXPF (1-254)	50 mM Tris (8), 300 mM NaCl, 2 mM 2-ME	Lysis + 20 mM imidazole	Lysis + 3C protease
ApeXPFΔN (18-254)	50 mM Tris (8), 300 mM NaCl, 2 mM 2-ME	Lysis + 20 mM imidazole	Lysis + 3C protease
ApeXPF$(HhH)_2$ & PIM (164-254)	50 mM Tris (8), 100 mM NaCl, 2 mM 2-ME	Lysis + 20 mM imidazole	Lysis + 3C protease
ApePCNA- ApeXPF$(HhH)_2$ & PIM (164-254)	50 mM Tris (8), 100 mM NaCl, 2 mM 2-ME	Lysis + 20 mM imidazole	Lysis + 3C protease
ApePCNA- ApeXPFΔN (18-254)	50 mM Tris (8), 300 mM NaCl, 2 mM 2-ME	Lysis + 20 mM imidazole	Lysis + 20 mM → 500 mM imidazole
ApePCNA- ApeXPFΔN (18-254)-ApeXPB	50 mM Tris (8), 300 mM NaCl, 2 mM 2-ME	Lysis + 20 mM imidazole	Lysis + 20 mM → 500 mM imidazole

Table A.5 – Buffers for the purification of His_6-tagged proteins

A.5.3.2 Gradient purification

Imidazole gradients were used to purify His_6-tagged *Aeropyrum pernix* complexes in Chapter 4. Cell pellets were resuspended and sonicated as for the batch purification, although being thermophilic proteins all manipulations were performed at room temperature. After clearing the cell debris by centrifugation, the lysates were heated in a water batch at 65 °C to denature the *E.coli* host proteins. The precipitated *E.coli* proteins were removed by centrifugation at 29,200 x g, for 30 mins at 22 °C. The supernatant was loaded onto a 1 ml Hi-Trap chelating sepharose column (Amersham) pre-charged with nickel sulphate using a syringe. After loading the column was

transferred to the room temperature ÄKTA FPLC. Unbound protein was washed from the column with 5 column volumes of loading buffer. An imidazole gradient (from 20-500 mM imidazole) over 10 column volumes was generally used to elute the protein(s) of interest. Fraction collection was automated and aliquots of these fractions were analysed by SDS-PAGE.

A.5.4 Purification of GST-tagged proteins

Protein expressed	Lysis buffer	Wash buffer	Elution buffer
GST-XPA	50 mM Tris (8), 150 mM NaCl, 1 mM DTT	Lysis w/ 20 mM Tris (8)	Wash + 20 mM GSH
GST-XPF – ERCC1	20 mM HEPES (7.4) 500 mM NaCl, 2 mM 2-ME, 2mM $MgCl_2$, 10% glycerol, 0.01% CHAPS.	Lysis	Lysis + 20 mM GSH
GST- hMus81 ACE domain constructs	50 mM Tris (8), 500 mM NaCl, 1 mM DTT	Lysis w/ 20 mM Tris (8)	Wash + 3C protease
GST- scMus81 ACE domain	50 mM Tris (8), 500 mM NaCl, 1 mM DTT	Lysis w/ 20 mM Tris (8)	Wash + 3C protease
GST – spMus81 ACE domain	50 mM Tris (8), 500 mM NaCl, 1 mM DTT	Lysis w/ 20 mM Tris (8)	Wash + 3C protease
GST – hMus81-hEme1/hEme2 complexes	50 mM phosphate (8), 300 mM NaCl, 2 mM 2-ME, 10 % glycerol, 0.2 % CHAPS	Lysis	Lysis + 20 mM GSH

Table A.6 – Buffers for the purification of GST-tagged proteins

Cell pellets were resuspended and sonicated as for His_6-tagged proteins. The cleared cell lysate was added to 1 ml of glutathione-sepharose affinity resin (Amersham) prewashed in lysis buffer (typically 1 ml of resin was used for 10 mg of protein expected). The affinity resin was collected from the unbound fractions and was washed thoroughly in wash buffer. Proteins were either eluted with wash buffer supplemented with 20 mM reduced glutathione (pH 8) or by addition of 3C protease (Structural Biology Laboratory stock) to a slurry of the affinity resin in wash buffer.

A.5.5 Heparin-sepharose affinity chromatography

Heparin-sepharose affinity chromatography was performed using the ÄKTA FPLC apparatus. In general 1 ml HiTrap Heparin columns (Pharmacia) were used with the suggested flow-rates and pressure limits. Depending on the volume to be bound to the column one of two approaches was taken. For volumes of 10 ml and under, the column was connected to the ÄKTA FPLC and washed with 5 column volumes of loading buffer. The protein was then injected onto the column. Alternatively for large volumes the column was washed with 5 column volumes of loading buffer with a syringe, before the sample was loaded with a syringe. After loading the column was

transferred to the ÄKTA FPLC. In both cases, unbound protein was washed from the column with 2 column volumes of loading buffer. A NaCl gradient elution up to 1 M NaCl over 10 column volumes was generally used to elute the protein(s) of interest. Fraction collection was automated and aliquots of these fractions were analysed by SDS-PAGE.

Protein expressed	Lysis buffer	Load buffer	Elution buffer
Full length XPF-ERCC1 (2^{nd} purification step)	As from Ni^{2+}NTA, 20 mM HEPES (7.4), 50 mM NaCl, 2 mM 2-ME, 10% glycerol, 0.01% CHAPS, 0.5 mM EDTA	As from Ni^{2+} NTA-resin	Gradient: lysis → 1 M NaCl
ApePCNA subunits	50 mM Tris (8), 100 mM NaCl, 1 mM DTT	As lysis	Gradient: lysis → 1 M NaCl
ApePCNA123 complex	50 mM Tris (8), 100 mM NaCl, 1 mM DTT	As lysis	Gradient: lysis → 1 M NaCl
ApeXPB	50 mM Tris (8), 100 mM NaCl, 1 mM DTT	As lysis	Gradient: lysis → 1 M NaCl
hMus81 ACE (2^{nd} purification step)	From GSH-resin, 20 mM Tris (8), 100 mM NaCl, 1 mM DTT	As from GSH-resin	Gradient: → 1 M NaCl
GST – hMus81-hEme1/hEme2 complexes (2^{nd} purification step)	From GSH-resin, 50 mM phosphate (8), 300 mM NaCl, 2 mM 2-ME, 10 % glycerol, 0.2 % CHAPS, 20 mM GSH	As from GSH-resin	Gradient: → 1 M NaCl

Table A.7 – Buffers for the purification of proteins on Heparin-Sepharose

A.5.6 Size exclusion chromatography

Protein expressed	Isocratic elution buffer
Full length XPF-ERCC1	20 mM HEPES (7.4), 50 mM NaCl, 0.5 mM EDTA, 0.5 mM DTT, 2 mM $MgCl_2$ 10% glycerol, 0.01% CHAPS.
ApePCNA subunits	20 mM Tris (8), 100 mM NaCl, 1 mM DTT
ApePCNA123 complex	20 mM Tris (8), 100 mM NaCl, 1 mM DTT
ApeXPB	20 mM Tris (8), 100 mM NaCl, 1 mM DTT
ApeXPF$(HhH)_2$ & PIM (164-254)	20 mM Tris (8), 100 mM NaCl, 1 mM DTT
ApePCNA- ApeXPF$(HhH)_2$ & PIM (164-254)	20 mM Tris (8), 100 mM NaCl, 1 mM DTT
ApePCNA- ApeXPFΔN (18-254)	20 mM Tris (8), 300 mM NaCl, 1 mM DTT
ApePCNA- ApeXPFΔN (18-254)-ApeXPB	20 mM Tris (8), 300 mM NaCl, 1 mM DTT
hMus81 ACE (2^{nd} purification step)	20 mM Tris (8), 500 mM NaCl, 1 mM DTT
scMus81 ACE domain (2^{nd} purification step)	20 mM Tris (8), 500 mM NaCl, 1 mM DTT
spMus81 ACE domain (2^{nd} purification step)	20 mM Tris (8), 500 mM NaCl, 1 mM DTT
GST – hMus81-hEme1/hEme2 complexes (3^{rd} purification step)	50 mM phosphate (8), 300 mM NaCl, 2 mM 2-ME, 10 % glycerol, 0.2 % CHAPS

Table A.8 – Buffers for purification of proteins by size exclusion (gel filtration) chromatography

Size exclusion chromatography (also referred to as gel filtration) was performed using a ÄKTA FPLC apparatus (Pharmacia). Columns used in this thesis (Superdex 75, superdex 200 and superose 12) were all purchased from Pharmacia and were used with the suggested flow-rates and pressure limits, and stored after use in 20 % (v/v)

ethanol. Columns were generally equilibrated with two column volumes of buffer. After this the sample was injected onto the column and another column volume of buffer was used for the isocratic elution of the proteins of interest. Fraction collection was automated and aliquots of these fractions were analysed by SDS-PAGE. The columns were calibrated using proteins of a known mass (n.b. calibration curves of these columns are not definitive as gel filtration relies on both the mass and the shape of the protein, given by the Stokes' radius).

A.6 Activity assays

Substrates used for Mus81 and ApePCNA-XPF activity assays were derived from the study of Ciccia et al. (2003). Three synthetic DNA strands were used in combination to generate the 3'-flap and splayed arm substrates.

XO1-GACGCTGCCGAATTCTACCAGTGCCTTGCTAGGACATCTTTGCCCACCTGCAGGTTCACCC

XO3-CATGGAGCTGTCTAGAGGATCCGACTATCGA

XO4 –ATCGATAGTCGGATCCTCTAGACAGCTCCATGTAGCAAGGCACTGGTAGAATTCGGCAGCGT

Strand XO1 was ^{32}P-5'-labelled. For the 3'-flap all three strands were annealed together. For the splayed arm just strands XO1 and XO4 were anneled together. The assay conditions used were identical to those of Ciccia et al. for Mus81 and those of Roberts et al. (Roberts et al., 2003) for ApePCNA-XPF. The reaction products were separated on a 10 % sequencing gel for 2 hours followed by exposure to a Phosphoimager screen or autoradiography using Biomax film for 2 hours.

A.7 Circular dichroism

Circular dichroism spectra were measured on a Aviv 202SF spectropolarimeter using 60 µl samples in a 0.1 mm quartz cuvette at room temperature. Scans were carried out from 190 to 260 nm with a bandwidth of 1 nm and a scan rate of 20 nm/min.

Appendix B Crystallisation and crystallographic theory

B.1 Protein Crystals

B.1.1 Crystallisation

The basic principle in protein crystallisation is bringing a protein solution to a limited degree of supersaturation (Ducruix and Griege, 1992). A phase diagram (figure B.1) can be used to illustrate how supersaturated protein solutions yield crystals. There are three regions of the phase diagram where the solution is supersaturated.

(a) The precipitation zone where excess protein molecules come out of solution and form amorphous aggregates.

(b) The nucleation zone where excess protein molecules come out of solution and form crystalline aggregates.

(c) The metastable zone where conditions for large well ordered crystals growth can be found, provided nucleation has occurred.

Although the aim is bringing a protein solution to a limited degree of supersaturation, the method defining conditions appropriate for the crystallisation of any particular macromolecule is still very much one of trial and error. In this work, two techniques of setting up crystallisation trials were used to find a path via the nucleation zone and into the metastable zone (figure B.1). These were vapour diffusion (using either hanging or sitting drops) and microbatch (using either paraffin oil or Al's oil (Hampton Research)). A variety of sparse matrix and grid screens were used, most of which were purchased from Hampton Research or Nextal.

The setup of vapour diffusion experiments typically involved 1-2 μl of concentrated protein sample being mixed with 1-2 μl of a precipitant solution taken from a 500 μl precipitant reservoir to form a "drop". The initial precipitant concentration in the drop was therefore lower than that in the reservoir. In order to reach equilibrium the reservoir draws water from the drop, thus increasing the concentration of precipitant and protein in the drop. It is hoped that this effect drives the drop into the nucleation zone. The removal of some protein molecules into crystal nuclei reduces the protein concentration, thus moving the drop into the metastable zone in which crystal growth may occur.

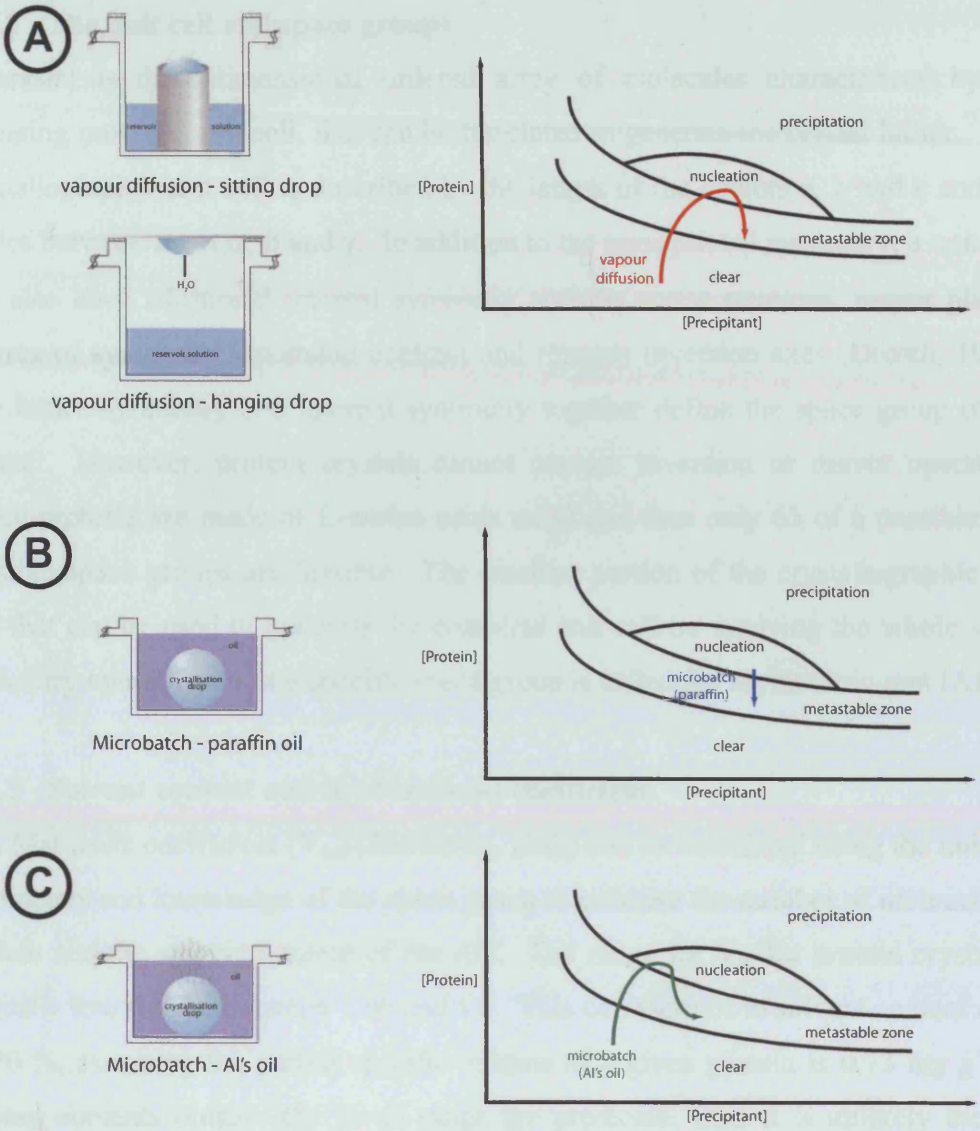

Figure B.1 – Phase diagrams for different crystallisation techniques. [A] Vapour diffusion. [B] Microbatch with paraffin oil. [C] Microbatch with Al's oil. Adapted from Hampton Research.

Microbatch experiments work in a slightly different way. Typically 1 µl of concentrated protein sample was mixed with 1 µl of a precipitant solution to form a drop under a layer of oil. In microbatch experiments with paraffin oil, little evaporation from the drop is possible and thus in this scenario, the precipitant must alter the level of protein supersaturation so that it falls directly into the nucleation zone (figure B.1). Microbatch experiments with Al's oil, allows some evaporation from the drop and thus the phase diagram and theory is more akin to that of vapour diffusion.

B.1.2 The unit cell and space groups

A crystal is three-dimensional ordered array of molecules characterised by its repeating unit, the unit cell, that can be translated to generate the crystal lattice. The crystallographic unit cell is described by the length of the vectors a, b and c and the angles between them α, β and γ. In addition to the translational symmetry, a unit cell can also have additional internal symmetry such as screw rotations, mirror planes, centres of symmetry (inversion centres) and rotation inversion axes (Drenth, 1994). The lattice symmetry and internal symmetry together define the space group of the crystal. However, protein crystals cannot contain inversion or mirror operations (since proteins are made of L-amino acids only) and thus only 65 of a possible 320 possible space groups are feasible. The smallest portion of the crystallographic unit cell that can be used to generate the complete unit cell by applying the whole set of symmetry operations of the specific space group is called the asymmetric unit (AU).

B.1.3 Solvent content and the Matthews coefficient.

The Matthews coefficient (V_m) (Matthews, 1968) can be calculated using the unit cell parameters and knowledge of the space group to estimate the number of molecules of protein and the solvent content of the AU. The range for V_m for protein crystals is typically found to be between 1.66 and 4.0. This corresponds to solvent content of 25 to 70 %, assuming the partial specific volume of a given protein is 0.73 mg g^{-1}. If solvent contents outside the given range are predicted, then it is unlikely the AU contains that number of molecules. The Matthews coefficient is given by:

Equation B.1
$$V_m = \frac{V}{MZ}$$

Where V_m = Matthews coefficient; V = volume of the unit cell, M = the molecular mass of the protein and Z = the number of molecules in the asymmetric unit.

B.2 X-ray diffraction by crystals

B.2.1 Bragg's law

X-rays are scattered by the electrons within the atoms that make up proteins. A charged particle that interacts with an electromagnetic wave experiences a force,

which caused it to oscillate. This causes it to act as a new source of radiation and thus "scatter" the incident radiation. In order to observe constructive interference, the X-rays must scatter in phase in a relationship given by Bragg's Law (Equation B.2)

Equation B.2 $$n\lambda = 2d\sin\theta$$

Where θ = angle of diffraction, λ = wavelength, n = integer value and d = inter-planar spacing between atoms.

The unit cells of protein crystals are much larger than the wavelength of the radiation used and so in order to measure data for all reflections the crystal is rotated to bring successive reflections into a diffracting position. The rotation method was employed for all data collection in this thesis and is where the crystal is rotated about a single axis through a narrow, defined angle (normally 0.5-1 °) during each exposure to the X-rays. All reflections that satisfy Bragg's law over each angle are recorded as diffraction spots on a single image.

Rearrangement of Bragg's law shows that $\theta \propto (1/d)$. This means that it is easier to represent the diffraction data using a reciprocal lattice based on 1/d that varies as sinθ. This reciprocal lattice can be used to describe the unit cell edges as a/h, b/k and c/l where h, k and l are integers and are called the Miller indices. The notation (hkl) used to describe the plane defined by the h, k and l Miller indices. The space described by the Miller indices is referred to as reciprocal space.

B.2.2 X-ray sources and detectors

Although some of the crystals described in this thesis diffracted weakly on the rotating anode X-ray source at Cancer Research UK, the datasets of the ApeXPF nuclease domain described were collected at the ESRF, Grenoble. Synchrotron radiation sources produce X-rays by forcing electrons travelling a high speed to adopt a curved path by the use of dipole magnets. This change of direction emits X-ray radiation. Synchrotrons produce polychromatic radiation from which specific wavelengths can be chosen using a monochromator.

The intensities of all the diffracted X-rays need to be measured in order to determine a crystal structure. A detecting device is required to measure the position and intensity of the diffracted beams. Accurate measurement of the diffracted pattern was performed using either a MAR storage phosphor-based image plate or a charge-coupled device (CCD). Image plates offer a wider dynamic range and generally larger active detector area than CCD detectors, but also require a longer read-out time. This can significantly limit the speed of data collection, especially at synchrotrons.

B.2.3 Cryocrystallography

The useful life of a protein crystal exposed to synchrotron radiation is limited by radiation damage. In order to minimise damage created by reactive species generated by absorption of high-energy photons, data were collected from crystals cooled to 100 K in a stream of cold nitrogen gas. Under these conditions, water in the solvent channels of the crystals and the surrounding liquor would usually (depending on the precipitant condition) form ice crystals. Ice crystals disrupt crystal order and give rise to diffraction, which may partially obscure protein diffraction. Crystals were therefore pre-soaked in a cryoprotectant solution that freezes to form a "glass". Cryoprotectant solutions are typically composed of a slightly increased concentration of the components of the crystal mother liquor together with glycerol, ethylene glycol or PEG400. It is usual for several cryoprotectants to be tested, for example in Chapter 5, many cryoprotectants were tried for hACE crystals. Crystals were supported during data collection using nylon loops (Hampton Research) held in a steel pin in turn attached to a magnetic base.

B.2.4 Data collection

During data collection, two factors govern the strategy to secure a complete dataset. The first is the symmetry of the reciprocal lattice. The second is the accuracy of the measure reflections. This depends on factors such as the exposure time, how strongly the crystal diffracts and the detector. It is useful to record each reflection multiple times as this allows its value to be determined with greater accuracy.

All of the considerations for a complete dataset were taken into account by using the STRATEGY option of the Mosflm program (Leslie, 1992) to calculate the initial oscillation angle, the oscillation range and the number of images needed to obtain a

complete dataset. These values were determined by the mosaic spread, the unit cell parameters and the crystal-to-detector distance. Mosaic spread is a measure of the total angle over which a particular diffraction spot satisfies Bragg's law. Spreading of reflections results from micro-heterogeneity in the orientation, size and cell parameters of the mosaic blocks that make up the crystal.

B.2.5 Data processing and reduction

From this point on, the experiment becomes a computational affair. The data processing program Mosflm (Leslie, 1992) was used for all diffraction data in this thesis. The first step in the data analysis is to obtain an accurate estimate of the space group and the unit cell parameters by autoindexing based on a one-dimensional Fast Fourier Transform (FFT) method. Then Mosflm was used to predict diffraction spot positions on the detector, estimate the X-ray background, and integrate the reflection intensities by profile fitting. A number of parameters describing the characteristics if the detector (geometry, sensitivity, dynamic range, distance for the crystal), the crystal (mosaic spread, cell dimensions, absorption) and incident beam (wavelength, divergence) are refined during this process.

Raw integrated intensities from each image must be corrected for differences in incident beam intensity between frames, crystal decay due to radiation damage, orientation dependent absorption by the crystal, and other systematic errors. The programs Scala (Evans, 1993) was used to calculate appropriate scale factors for each processed image, apply these scale factors, and merge with the corrected intensities. The statistical indicator R_{sym} is used to assess the quality of the data, along with the ratio of observed intensity to noise ($I/\sigma I$).

Equation B.3 $$R_{sym} = \frac{\sum |I - \langle I \rangle|}{\sum |\langle I \rangle|}$$

Where I is the intensity of an individual reflection, and <I> is the mean intensity of the set of equivalent reflections including I.

The merged intensities were converted to structure factor amplitudes using the program Truncate (Collaborative Computational Project, 1994). At this point the full set of unique reflections theoretically measurable for the resolution and space group

of the dataset are generated and a subset of 5 % of the reflections were "flagged" for omission from subsequent refinement.

B.3 Molecular replacement

Every reflection has both an amplitude (F) and a phase (α) determined by the positions and scattering power of the atoms contributing to that reflection.

Equation B.4 $\mathbf{F}(hkl) = F(hkl).\exp[i\alpha(hkl)]$

Where **F**(hkl) is the structure factor for the reflection of Miller indices h, k, l; F(hkl) is the structure factor amplitude and α(hkl) is the phase.

The diffraction pattern is the Fourier transform of the diffracting object (in this case the electron density for the molecules of the crystal), and thus the electron density, ρ (xyz) can be calculated by the Fourier summation of the structure factors, **F**(hkl), that define the diffraction pattern.

Equation B.5 $$\rho(xyz) = \frac{1}{V}\sum_{hkl} |F_{hkl}| \exp[2\pi i(hx + ky + lz) + i\alpha_{hkl}]$$

Where $|F_{hkl}|$ is the structure factor amplitude of reflection (hkl) including the temperature factor, V is the volume of the unit cell, ρ is electron density, α is phase angle and xyz are the co-ordinates of the atom in the unit cell.

In the absence of a lens capable of focussing the diffracted X-rays to form an image, detectors record only the intensity of the diffracted beam and the phase information is lost. This is generally referred to as the phase problem. In order to construct the electron density of the molecule, it is necessary to provide approximate phases. When the structure of a homologous protein is available (as it was in the case of the ApeXPF nuclease domain in Chapter 3), the method of molecular replacement can be used to solve the phase problem.

Molecular replacement consists of finding the rotation (a matrix [C]) and translation (d) that relates a molecule of known structure (X1) to the molecule of interest (X2).

Equation B.6 $X2 = [C].X1 + \underline{d}$

Molecular replacement was carried out using the program Molrep (Vagin and Teplyakov, 1997) which splits this process into three parts, a search for the rotation function [C], followed by a search for the translation, $\underline{d}$, and finally rigid-body refinement of the positioned molecule(s). First the Patterson function, P_c, calculated from the model is compared with that calculated using the observed data, P_o, for all orientations of the model to find a rotation function, R, that best matches the model to the observed data. Solutions are ranked according to R-factor and correlation co-efficient, which describe the agreement between the observed and calculated structure factor amplitudes (Fo, Fc) and Patterson functions (Po, Pc) respectively. The Patterson function, $P(\underline{u})$ is the Fourier transform of the square of the structure factor amplitudes, and can therefore be calculated directly from the measured intensities without need for prior knowledge of the phases. Peaks in the Patterson function indicate interatomic vectors.

Equations B.8 and B.9

$$P(\underline{u}) = \sum_v \rho(\underline{x}).\rho(\underline{x} + \underline{u})dV = \frac{2}{V}\sum_{hkl} F(hkl)^2.\cos[2\pi(hu + kv + lw)]$$

$$F_h^{\,2} = I(hkl)$$

Where $P(\underline{u})$ is the value of the Patterson function at a point specified by the vector $\underline{u}(u,v,w)$; $\rho(\underline{x})$ and $\rho(\underline{x} + \underline{u})$ are the electron density at positions $\underline{x}$ and $\underline{x} + \underline{u}$; $F(hkl)^2$ is the square of the structure factor amplitude for reflection hkl (i.e. the intensity I(hkl)) abd V is the unit cell volume.

The intermolecular vectors for the best rotation function solutions are then searched for a corresponding translation function, which positions the rotated molecule with respect to the symmetry axes of the unit cell. Finally, the transformations showing the best agreement are applied to the model, and rigid-body refinement is carried out to optimise agreement between the calculated and observed structure factor amplitudes.

B.4 Refinement

Models obtained in by molecular replacement were improved by an iterative process of interactive model building using the graphics package COOT (Emsley and Cowtan, 2004) and refinement using REFMAC 5 (Murshudov et al., 1997). The goal of refinement is to maximise the correlation between the correct stereochemical model and the observed structure factors. This is measured by two statistical factors known as R_{fact} and R_{free} (defined in equation B.10). The R_{free} is calculated from the random subset of 5 % of reflections taken after the merged intensities were converted to structure factor amplitudes. These reflections are excluded from the refinement and this acts as a method of cross-validation.

EquationB.10

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

Where $|F_{obs}|$ and $|F_{calc}|$ are the experimental and calculated structure factor amplitudes respectively for a reflection hkl either belonging to the working set of reflections (R_{fact}) or to the test set of reflections (R_{free})

Refinement programs use a target function in order to obtain the best possible model. REFMAC 5 (Murshudov et al., 1997) employs a maximum likelihood function using Bayes' theorem which suggests that the best model is that most consistent with the observations. Restrained refinement was performed in this thesis, which means that the textbook information known about protein structures was used as additional information within the refinement. Such information included known bond lengths, bond angles and permitted φ and ϕ dihedral angles. Refinement is considered complete when R_{fact} and R_{free} have converged and the geometry of the model has been positively assessed by programs such as PROCHECK (Laskowski et al., 1993).

References

Abraham, J., Lemmers, B., Hande, M. P., Moynahan, M. E., Chahwan, C., Ciccia, A., Essers, J., Hanada, K., Chahwan, R., Khaw, A. K., et al. (2003). Eme1 is involved in DNA damage processing and maintenance of genomic stability in mammalian cells. EMBO J **22**, 6137-6147.

Araujo, S. J., and Wood, R. D. (1999). Protein complexes in nucleotide excision repair. Mutat Res **435**, 23-33.

Avery, O.T., MacLeod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal type. J. Exp. Med. **79**, 137-158

Bastin-Shanower, S. A., Fricke, W. M., Mullen, J. R., and Brill, S. J. (2003). The mechanism of Mus81-Mms4 cleavage site selection distinguishes it from the homologous endonuclease Rad1-Rad10. Mol Cell Biol **23**, 3487-3496.

Begley, T. J., and Samson, L. D. (2003). AlkB mystery solved: oxidative demethylation of N1-methyladenine and N3-methylcytosine adducts by a direct reversal mechanism. Trends Biochem Sci **28**, 2-5.

Boddy, M. N., Gaillard, P. H., McDonald, W. H., Shanahan, P., Yates, J. R., 3rd, and Russell, P. (2001). Mus81-Eme1 are essential components of a Holliday junction resolvase. Cell **107**, 537-548.

Boddy, M. N., Lopez-Girona, A., Shanahan, P., Interthal, H., Heyer, W. D., and Russell, P. (2000). Damage tolerance protein Mus81 associates with the FHA1 domain of checkpoint kinase Cds1. Mol Cell Biol **20**, 8758-8766.

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

Bruck, I., and O'Donnell, M. (2001). The ring-type polymerase sliding clamp family. Genome Biol **2**, REVIEWS3001.

Camenisch, U., Dip, R., Schumacher, S. B., Schuler, B., and Naegeli, H. (2006). Recognition of helical kinks by xeroderma pigmentosum group A protein triggers DNA excision repair. Nat Struct Mol Biol **13**, 278-284.

Carr, A. M., Schmidt, H., Kirchhoff, S., Muriel, W. J., Sheldrick, K. S., Griffiths, D. J., Basmacioglu, C. N., Subramani, S., Clegg, M., Nasim, A., and et al. (1994). The rad16 gene of Schizosaccharomyces pombe: a homolog of the RAD1 gene of Saccharomyces cerevisiae. Mol Cell Biol **14**, 2029-2040.

Caruthers, J. M., and McKay, D. B. (2002). Helicase structure and mechanism. Curr Opin Struct Biol **12**, 123-133.

Ciccia, A., Constantinou, A., and West, S. C. (2003). Identification and characterization of the human mus81-eme1 endonuclease. J Biol Chem **278**, 25172-25178.

Collaborative Computational Project, N. (1994). The CCP4 Suite: Programs for Protein Crystallography. Acta Crystallographica D50, 760 760 763.

Daimon, K., Kawarabayasi, Y., Kikuchi, H., Sako, Y., and Ishino, Y. (2002). Three proliferating cell nuclear antigen-like proteins found in the hyperthermophilic archaeon Aeropyrum pernix: interactions with the two DNA polymerases. J Bacteriol **184**, 687-694.

de Laat, W. L., Appeldoorn, E., Jaspers, N. G., and Hoeijmakers, J. H. (1998). DNA structural elements required for ERCC1-XPF endonuclease activity. J Biol Chem **273**, 7835-7842.

De Silva, I. U., McHugh, P. J., Clingen, P. H., and Hartley, J. A. (2000). Defining the roles of nucleotide excision repair and recombination in the repair of DNA interstrand

cross-links in mammalian cells. Mol Cell Biol **20**, 7980-7990.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System. World Wide Web.

Dendouga, N., Gao, H., Moechars, D., Janicot, M., Vialard, J., and McGowan, C. H. (2005). Disruption of murine Mus81 increases genomic instability and DNA damage sensitivity but does not promote tumorigenesis. Mol Cell Biol **25**, 7569-7579.

Dionne, I., Nookala, R. K., Jackson, S. P., Doherty, A. J., and Bell, S. D. (2003). A heterotrimeric PCNA in the hyperthermophilic archaeon Sulfolobus solfataricus. Mol Cell **11**, 275-282.

Doe, C. L., Ahn, J. S., Dixon, J., and Whitby, M. C. (2002). Mus81-Eme1 and Rqh1 involvement in processing stalled and collapsed replication forks. J Biol Chem **277**, 32753-32759.

Doherty, A. J., Serpell, L. C., and Ponting, C. P. (1996). The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. Nucleic Acids Res **24**, 2488-2497.

Drenth, J. (1994) Principles of Protein X-ray crystallography. Springer-Velag

Ducruix, A. and Griege, R. (1992). Crystallisation of Nucleic Acids and Proteins, IRL Press.

Eichman, B. F., Mooers, B. H., Alberti, M., Hearst, J. E., and Ho, P. S. (2001). The crystal structures of psoralen cross-linked DNAs: drug-dependent formation of Holliday junctions. J Mol Biol **308**, 15-26.

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

Enzlin, J. H., and Scharer, O. D. (2002). The active site of the DNA repair endonuclease XPF-ERCC1 forms a highly conserved nuclease motif. EMBO J **21**,

2045-2053.

Evans, P. R. (1993). Data reduction. Paper presented at: CCP4 Study Weekend on Data Collection & Processing.

Fricke, W. M., Bastin-Shanower, S. A., and Brill, S. J. (2005). Substrate specificity of the Saccharomyces cerevisiae Mus81-Mms4 endonuclease. DNA Repair (Amst) **4**, 243-251.

Friedberg, E.C., Walker, G.C., Siede, W., Wood, R.D., Schultz, R.A., Ellenberger, T. (2006). DNA repair and mutagenesis (Second Edition). American Society for Microbiology.

Gaillard, P. H., and Wood, R. D. (2001). Activity of individual ERCC1 and XPF subunits in DNA nucleotide excision repair. Nucleic Acids Res **29**, 872-879.

Gajiwala, K. S., and Burley, S. K. (2000). Winged helix proteins. Curr Opin Struct Biol **10**, 110-116.

Garinis, G. A., Jans, J., and van der Horst, G. T. (2006). Photolyases: capturing the light to battle skin cancer. Future Oncol **2**, 191-199.

Gulbis, J. M., Kelman, Z., Hurwitz, J., O'Donnell, M., and Kuriyan, J. (1996). Structure of the C-terminal region of p21(WAF1/CIP1) complexed with human PCNA. Cell **87**, 297-306.

Guzder, S. N., Torres-Ramos, C., Johnson, R. E., Haracska, L., Prakash, L., and Prakash, S. (2004). Requirement of yeast Rad1-Rad10 nuclease for the removal of 3'-blocked termini from DNA strand breaks induced by reactive oxygen species. Genes Dev **18**, 2283-2291.

Haering, C. H., Schoffnegger, D., Nishino, T., Helmhart, W., Nasmyth, K., and Lowe, J. (2004). Structure and stability of cohesin's Smc1-kleisin interaction. Mol Cell **15**, 951-964.

Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K., and Pease, L. R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene **77**, 51-59.

Hollingsworth, N. M., and Brill, S. J. (2004). The Mus81 solution to resolution: generating meiotic crossovers without Holliday junctions. Genes Dev **18**, 117-125.

Horton, N. C., and Perona, J. J. (1998). Role of protein-induced bending in the specificity of DNA recognition: crystal structure of EcoRV endonuclease complexed with d(AAAGAT) + d(ATCTT). J Mol Biol **277**, 779-787.

Ikegami, T., Kuraoka, I., Saijo, M., Kodo, N., Kyogoku, Y., Morikawa, K., Tanaka, K., and Shirakawa, M. (1998). Solution structure of the DNA- and RPA-binding domain of the human repair factor XPA. Nat Struct Biol **5**, 701-706.

Interthal, H., and Heyer, W. D. (2000). MUS81 encodes a novel helix-hairpin-helix protein involved in the response to UV- and methylation-induced DNA damage in Saccharomyces cerevisiae. Mol Gen Genet **263**, 812-827.

Jair, K. W., Martin, R. G., Rosner, J. L., Fujita, N., Ishihama, A., and Wolf, R. E., Jr. (1995). Purification and regulatory properties of MarA protein, a transcriptional activator of Escherichia coli multiple antibiotic and superoxide resistance promoters. J Bacteriol **177**, 7100-7104.

Kelman, Z., and White, M. F. (2005). Archaeal DNA replication and repair. Curr Opin Microbiol **8**, 669-676.

Komori, K., Fujikane, R., Shinagawa, H., and Ishino, Y. (2002). Novel endonuclease in Archaea cleaving DNA with various branched structure. Genes Genet Syst **77**, 227-241.

Komori, K., Hidaka, M., Horiuchi, T., Fujikane, R., Shinagawa, H., and Ishino, Y. (2004). Cooperation of the N-terminal Helicase and C-terminal endonuclease

activities of Archaeal Hef protein in processing stalled replication forks. J Biol Chem **279**, 53175-53185.

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

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature **227**, 680-685.

Lally, J., Newman, M., Murray-Rust, J., Fadden, A., Kawarabayasi, Y., and McDonald, N. (2004). Crystallization of the xeroderma pigmentosum group F endonuclease from Aeropyrum pernix. Acta Crystallogr D Biol Crystallogr **60**, 1658-1661.

Laskowski, R. A., Macarthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK - a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography **26**, 283-291.

Lee, J. H., Hwang, G. S., and Choi, B. S. (1999). Solution structure of a DNA decamer duplex containing the stable 3' T.G base pair of the pyrimidine(6-4)pyrimidone photoproduct [(6-4) adduct]: implications for the highly specific 3' T --> C transition of the (6-4) adduct. Proc Natl Acad Sci U S A **96**, 6632-6636.

Lee, J. H., Park, C. J., Shin, J. S., Ikegami, T., Akutsu, H., and Choi, B. S. (2004). NMR structure of the DNA decamer duplex containing double T*G mismatches of cis-syn cyclobutane pyrimidine dimer: implications for DNA damage recognition by the XPC-hHR23B complex. Nucleic Acids Res **32**, 2474-2481.

Lehmann, A. R. (2003). DNA repair-deficient diseases, xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy. Biochimie **85**, 1101-1111.

Leslie, A. G. W. (1992). Recent changes to the MOSFLM package for processing film and image plate data. Joint CCP4 + ESF-EAMCB Newsletter on Protein

Crystallography 26.

Li, L., Elledge, S. J., Peterson, C. A., Bales, E. S., and Legerski, R. J. (1994). Specific association between the human DNA repair proteins XPA and ERCC1. Proc Natl Acad Sci U S A **91**, 5012-5016.

Li, L., Peterson, C. A., Lu, X., and Legerski, R. J. (1995). Mutations in XPA that prevent association with ERCC1 are defective in nucleotide excision repair. Mol Cell Biol **15**, 1993-1998.

Liu, Y., Masson, J. Y., Shah, R., O'Regan, P., and West, S. C. (2004). RAD51C is required for Holliday junction processing in mammalian cells. Science **303**, 243-246.

Matsumiya, S., Ishino, Y., and Morikawa, K. (2001). Crystal structure of an archaeal DNA sliding clamp: proliferating cell nuclear antigen from Pyrococcus furiosus. Protein Sci **10**, 17-23.

Matthews, B. W. (1968). Solvent content of protein crystals. J Mol Biol **33**, 491-497.

McHugh, P. J., Spanswick, V. J., and Hartley, J. A. (2001). Repair of DNA interstrand crosslinks: molecular mechanisms and clinical relevance. Lancet Oncol **2**, 483-490.

McPherson, J. P., Lemmers, B., Chahwan, R., Pamidi, A., Migon, E., Matysiak-Zablocki, E., Moynahan, M. E., Essers, J., Hanada, K., Poonepalli, A., et al. (2004). Involvement of mammalian Mus81 in genome integrity and tumor suppression. Science **304**, 1822-1826.

McWhir, J., Selfridge, J., Harrison, D. J., Squires, S., and Melton, D. W. (1993). Mice with DNA repair gene (ERCC-1) deficiency have elevated levels of p53, liver nuclear abnormalities and die before weaning. Nat Genet **5**, 217-224.

Meetei, A. R., Medhurst, A. L., Ling, C., Xue, Y., Singh, T. R., Bier, P., Steltenpool, J., Stone, S., Dokal, I., Mathew, C. G., et al. (2005). A human ortholog of archaeal

DNA repair protein Hef is defective in Fanconi anemia complementation group M. Nat Genet **37**, 958-963.

Motycka, T. A., Bessho, T., Post, S. M., Sung, P., and Tomkinson, A. E. (2004). Physical and functional interaction between the XPF/ERCC1 endonuclease and hRad52. J Biol Chem **279**, 13634-13639

Mullen, J. R., Kaliraman, V., Ibrahim, S. S., and Brill, S. J. (2001). Requirement for three novel protein complexes in the absence of the Sgs1 DNA helicase in Saccharomyces cerevisiae. Genetics **157**, 103-118.

Murshudov, G., A.Vagin, A., and E.J., D. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr **53**, 240-255.

Newman, M., Murray-Rust, J., Lally, J., Rudolf, J., Fadden, A., Knowles, P. P., White, M. F., and McDonald, N. Q. (2005). Structure of an XPF endonuclease with and without DNA suggests a model for substrate recognition. EMBO J **24**, 895-905.

Niedernhofer, L. J., Essers, J., Weeda, G., Beverloo, B., de Wit, J., Muijtjens, M., Odijk, H., Hoeijmakers, J. H., and Kanaar, R. (2001). The structure-specific endonuclease Ercc1-Xpf is required for targeted gene replacement in embryonic stem cells. EMBO J **20**, 6540-6549.

Nishino, T., Komori, K., Ishino, Y., and Morikawa, K. (2003). X-ray and biochemical anatomy of an archaeal XPF/Rad1/Mus81 family nuclease: similarity between its endonuclease domain and restriction enzymes. Structure (Camb) **11**, 445-457.

Nishino, T., Komori, K., Ishino, Y., and Morikawa, K. (2005b). Structural and functional analyses of an archaeal XPF/Rad1/Mus81 nuclease: asymmetric DNA binding and cleavage mechanisms. Structure (Camb) **13**, 1183-1192.

Nishino, T., Komori, K., Tsuchiya, D., Ishino, Y., and Morikawa, K. (2005a). Crystal structure and functional implications of Pyrococcus furiosus hef helicase domain

involved in branched DNA processing. Structure **13**, 143-153.

Ogrunc, M., and Sancar, A. (2003). Identification and characterization of human MUS81-MMS4 structure-specific endonuclease. J Biol Chem **278**, 21715-21720.

Park, C. J., and Choi, B. S. (2006). The protein shuffle. Sequential interactions among components of the human nucleotide excision repair pathway. Febs J **273**, 1600-1608.

Pegg, A. E., and Byers, T. L. (1992). Repair of DNA containing O6-alkylguanine. Faseb J **6**, 2302-2310.

Pellegrini, L., Yu, D. S., Lo, T., Anand, S., Lee, M., Blundell, T. L., and Venkitaraman, A. R. (2002). Insights into DNA recombination from the structure of a RAD51-BRCA2 complex. Nature **420**, 287-293.

Pingoud, A., Fuxreiter, M., Pingoud, V., and Wende, W. (2005). Type II restriction endonucleases: structure and mechanism. Cell Mol Life Sci **62**, 685-707.

Rafferty, J. B., Ingleston, S. M., Hargreaves, D., Artymiuk, P. J., Sharples, G. J., Lloyd, R. G., and Rice, D. W. (1998). Structural similarities between Escherichia coli RuvA protein and other DNA-binding proteins and a mutational analysis of its binding to the holliday junction. J Mol Biol **278**, 105-116.

Roberts, J. A., and White, M. F. (2005). An archaeal endonuclease displays key properties of both eukaryal XPF-ERCC1 and Mus81. J Biol Chem **280**, 5924-5928.

Roberts, J. A., Bell, S. D., and White, M. F. (2003). An archaeal XPF repair endonuclease dependent on a heterotrimeric PCNA. Mol Microbiol **48**, 361-371.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A **74**, 5463-5467.

Schmidt, H., Kapitza-Fecke, P., Stephen, E. R., and Gutz, H. (1989). Some of the swi genes of Schizosaccharomyces pombe also have a function in the repair of radiation

damage. Curr Genet **16,** 89-94.

Shao, X., and Grishin, N. V. (2000). Common fold in helix-hairpin-helix proteins. Nucleic Acids Res **28**, 2643-2650.

Sijbers, A. M., de Laat, W. L., Ariza, R. R., Biggerstaff, M., Wei, Y. F., Moggs, J. G., Carter, K. C., Shell, B. K., Evans, E., de Jong, M. C., et al. (1996). Xeroderma pigmentosum group F caused by a defect in a structure-specific DNA repair endonuclease. Cell **86**, 811-822.

Sijbers, A. M., van Voorst Vader, P. C., Snoek, J. W., Raams, A., Jaspers, N. G., and Kleijer, W. J. (1998). Homozygous R788W point mutation in the XPF gene of a patient with xeroderma pigmentosum and late-onset neurologic disease. J Invest Dermatol **110**, 832-836.

Singh, S., Folkers, G. E., Bonvin, A. M., Boelens, R., Wechselberger, R., Niztayev, A., and Kaptein, R. (2002). Solution structure and DNA-binding properties of the C-terminal domain of UvrC from E.coli. EMBO J **21**, 6257-6266.

Soultanas, P., and Wigley, D. B. (2001). Unwinding the 'Gordian knot' of helicase action. Trends Biochem Sci **26**, 47-54.

Thayer, M. M., Ahern, H., Xing, D., Cunningham, R. P., and Tainer, J. A. (1995). Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. EMBO J **14**, 4108-4120.

Tian, M., Shinkura, R., Shinkura, N., and Alt, F. W. (2004). Growth retardation, early death, and DNA repair defects in mice deficient for the nucleotide excision repair enzyme XPF. Mol Cell Biol **24**, 1200-1205.

Tomkinson, A. E., Bardwell, A. J., Bardwell, L., Tappe, N. J., and Friedberg, E. C. (1993). Yeast DNA repair and recombination proteins Rad1 and Rad10 constitute a single-stranded-DNA endonuclease. Nature **362**, 860-862.

Tripsianes, K., Folkers, G., Ab, E., Das, D., Odijk, H., Jaspers, N. G., Hoeijmakers, J. H., Kaptein, R., and Boelens, R. (2005). The structure of the human ERCC1/XPF interaction domains reveals a complementary role for the two proteins in nucleotide excision repair. Structure **13**, 1849-1858.

Tsodikov, O. V., Enzlin, J. H., Scharer, O. D., and Ellenberger, T. (2005). Crystal structure and DNA binding functions of ERCC1, a subunit of the DNA structure-specific endonuclease XPF-ERCC1. Proc Natl Acad Sci U S A **102**, 11236-11241.

Vagin, A., and Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. J Applied Crystallography **30**, 1022-1025.

van Hoffen, A., Balajee, A. S., van Zeeland, A. A., and Mullenders, L. H. (2003). Nucleotide excision repair and its interplay with transcription. Toxicology **193**, 79-90.

Wah, D. A., Hirsch, J. A., Dorner, L. F., Schildkraut, I., and Aggarwal, A. K. (1997). Structure of the multimodular endonuclease FokI bound to DNA. Nature **388**, 97-100.

Warbrick, E. (2000). The puzzle of PCNA's many partners. Bioessays **22**, 997-1006.

Westerveld, A., Hoeijmakers, J. H., van Duin, M., de Wit, J., Odijk, H., Pastink, A., Wood, R. D., and Bootsma, D. (1984). Molecular cloning of a human DNA repair gene. Nature **310**, 425-429.

Whitby, M. C. (2005). Making crossovers during meiosis. Biochem Soc Trans **33**, 1451-1455.

Whitby, M. C., Osman, F., and Dixon, J. (2003). Cleavage of model replication forks by fission yeast Mus81-Eme1 and budding yeast Mus81-Mms4. J Biol Chem **278**, 6928-6935.

White, M. F. (2003). Archaeal DNA repair: paradigms and puzzles. Biochem Soc Trans **31**, 690-693.

White, M. F., and Bell, S. D. (2002). Holding it together: chromatin in the Archaea. Trends Genet **18**, 621-626.

Xiao, W., Chow, B. L., and Milo, C. N. (1998). Mms4, a putative transcriptional (co)activator, protects Saccharomyces cerevisiae cells from endogenous and environmental DNA damage. Mol Gen Genet **257**, 614-623.

Yagi, T., Matsumura, Y., Sato, M., Nishigori, C., Mori, T., Sijbers, A. M., and Takebe, H. (1998). Complete restoration of normal DNA repair characteristics in group F xeroderma pigmentosum cells by over-expression of transfected XPF cDNA. Carcinogenesis **19**, 55-60.

Yanez, R. J., Rodriguez, J. M., Nogal, M. L., Yuste, L., Enriquez, C., Rodriguez, J. F., and Vinuela, E. (1995). Analysis of the complete nucleotide sequence of African swine fever virus. Virology **208**, 249-278.

Yang, W (2004) Crystallization of protein-DNA complexes. http://mcl1.ncifcrf.gov/nihxray/Tips-and-Tricks_Crystallization_Protein-DNA.html

Zhang, R., Sengupta, S., Yang, Q., Linke, S. P., Yanaihara, N., Bradsher, J., Blais, V., McGowan, C. H., and Harris, C. C. (2005). BLM helicase facilitates Mus81 endonuclease activity in human cells. Cancer Res **65**, 2526-2531.

Zhu, X. D., Niedernhofer, L., Kuster, B., Mann, M., Hoeijmakers, J. H., and de Lange, T. (2003). ERCC1/XPF removes the 3' overhang from uncapped telomeres and represses formation of telomeric DNA-containing double minute chromosomes. Mol Cell **12**, 1489-1498.