Structural study of Phage lambda proteins

Naer Abdul Bari Alkaabawi

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

Molecular Biology and Biotechnology

November-2017

The University Of Sheffield.

Abstract

Lambda phage proteins

Four tail proteins of the Lambda bacteriophage, Rap endonuclease, NinH, Ea31 and Ea10, have previously been expressed in *E. coli*. Here, the methods are reported that were used to optimize the expression, purification and crystallization of these proteins. The structure determination of Ea10 is described.

No crystals have yet been obtained for the purified full-length Rap endonuclease. Therefore, two truncated versions of the Rap endonuclease, which removed the first 30 and 70 residues of the disordered N-terminus, were produced to potentially increase the chance of protein crystallization.

Overexpression experiments for NinH and Ea31 proteins were carried out with a His- tag or an MBP-tag, respectively, for solubility reasons. Both proteins were purified and set up in many crystallization trials. Crystals were not observed. Although tag cleavage for these proteins was performed to increase possibilities of crystallization, It was difficult to obtain a suitable level of purified protein for setting up crystallization trials.

Heavy-atom derivatized crystals were obtained for Ea10, from which its structure was solved by X-ray diffraction to 2.7 Å. It was found to form a dimeric structure with beta strands swapped between a beta sheet observed in each monomer. As yet, no strongly homologous structures have been found for Ea10 to help predict its main function. Structural comparisons for Ea10 via the Dali Server suggest a notable similarity with part of the Q-beta replicase core complex, indicating an association with RNA. Unpublished data from the laboratory of Dr. Gary Sharples (University of Durham, UK) suggests that it might have a DNA binding capability. However, assays described here have not been able to confirm this. Experiments are still ongoing to unravel the mystery of the Ea10 function.

Structure determination for the full length DnaD protein of *Bacillus subtilis,* which is a 232 amino acid primosomal protein that binds to supercoiled forms of DNA and converts them to open forms without nicking, was attempted, involving the N- and C-terminal domains which have never been solved together. Both domain structures (N-terminal and C-terminal) were published independently using crystallography and NMR assay, respectively. High resolution crystals (1.6 Å) with different space groups were collected. These crystals were confirmed as the N-terminal domain hits. Despite the use of a protease inhibitor was used for the prevention of a protein cleavage, N-terminal crystals were formed. Protein-protein interaction assay with DnaA protein for stability increase has not given any progress.

Acknowledgment

First of all, I would like to acknowledge and thank deeply my supervisor Dr. John B Rafferty for giving me the opportunity to make the first step in the scientific research world. For years of encouragement, guidance, explanation of crystallography work and optimism throughout my PhD study I am sincerely grateful.

The grateful and thankful continue to my collaborator Prof. Gary Sharples, who supported and provided me from the first time of my working with Lambda proteins, Dr. Panos Sultanas for providing DnaD and DnaA proteins. Furthermore, I would like thank Professor Jon Sayer and his group for letting me use equipments in his laboratory and for all the advice and support. In addition, for the great assistance and advice on protein purification by Dr. Sveta Sedelnikova, many thanks.

Secondly, I will never forget the big help, advice and support of Prof. Dave Rice and Dr. Pat Baker especially on structural analysis. I would also like to thank the soul and the lab organizer represented by Fiona HF Rodgers and all of the Postdocs and PhD students in Crystallography group starting: Dr. Claudine Bisson, Dr. Jason Wilson, Dr. Hayley Owen and Sam Dix…etc. who always offer me good suggestions during the Study period.

Many thanks for all of my colleagues in the lab: Dr. Sami Melebari, Dr. Israa Al-hawani, Dr. Adli Aziz for their friendship and assistance. I would like to say thank you so much to all of my friends and their family namely: Dr. Adnan Al-mosawi, Dr. Ehsan Al-Bermany and Dr. Murtakab Al-hejaji, for their Kindness and moral support during my study. Special thanks the gorgeous people Dr. Inam Lafta and Dr. Muslim AlEdami, Dr. Azhir Al-kaabi, Officer Wisam Al-kaabi, Dr. Haider Alshabbani, Dr. Salah Al-Obaidi, and Dr. Basheer Al-waely, for shown all the kindness, support and confidence. I also express my deep gratitude toward an amazing and wonderful girl, Roxanne Lau, for sharing her time, knowledge, kindness and assistance throughout my study journey, thanks a lot.

Further, I want to thank my family: my father, mother, sisters and my brothers for their unconditional and infinity love and support. Likewise, I am grateful to my wounded and oppressed country, Iraq. I hope from all of my heart to see my country safer, greater and stronger and recuperate of all terrorism and extremists forever. Finally, I would like to thank the Higher Committee for Education Development in Iraq (HCED), office of prime minister, to give me this opportunity to complete my PhD in the United Kingdom.

Table of contents

List of figures

List of tables

Abbreviations and Symbols

Species Abbreviations

B. subtilis ……..Bacillus subtilis
E. coli ……Escherichia coli
T4 ……Bacteriophage T4
T5 ……Bacteriophage T5

Chemical Abbreviations

APS ……Ammonium persulphate
Ca2+ ……..Calcium ion
DTT …… Dithiothreitol
EDTA…… Ethylenediaminetetraacitic acid
HEPES …… 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
IPTG…… Isopropyl-1-thio-!-d-galactopyranoside
K+ …… Potassium ion
MES ……2-(N-morpholino)ethanesulfonic acid
Mg2+ …… Magnesium ion
Mn2+…….. Manganese ion
Na+……Sodium ion
PEG ……..Poly ethylene glycol
SDS ……. Sodium dodecylsulphate
Sm3 …… Samarium ion
Zn2+ …….. Zinc ion

Miscellaneous Abbreviations

3D …….. Three dimensional
BLAST …….. Basic Local Alignment Search Tool
Cat1……..Catalytic metal ion binding site 1
Cat2 ……..Catalytic metal ion binding site 2
Cα ……Alpha carbon of amino acid
CCP4……Collaborative Computational Project No. 4
CFE……..Cell Free Extract
DLS……Diamond Light Source, Oxfordshire, UK
DNA……..Deoxyribonucleic acid
dA…… Deoxyadenosine
ds…….. Double stranded
ExoIX…….. Exonuclease IX from Escherichia coli
HTH……Helix-Turn-Helix
HhH……..Helix-hairpin-Helix
Ile……..Isoleucine
JCSG…….. ..Joint Centre for Structural Genomics
LB……..Luria-Bertani media
PAGE……Polyacrylamide gel Electrophoresis
PDB……..Protein DataBank
RMSD……Root mean square deviation
WT……Wild type

Units

° ……Degree
°C……..Degree Celcius
Cm……..Centimetre
µ…… Micro Meter
Å…….…… Ångstrom
Da…… Dalton
kDa…… Kilo dalton
g……..Gram
K…… Kelvin
M…….. Molar
m…….. Milli
MW…… Molecular weight
OD600 ……..Optical density at 600 nm
Rpm…… Rotations per minute
s…… Second
v/v ……Volume to volume
v/w ……Volume to weight
xg……Acceleration of gravity

Crystallographic Terms

a, b, c ……Real space unit cell dimensions
α, β, γ……Real space unit cell angle
AU……Asymmetric unit
B-factor ……..Crystallographic temperature factor
ƒ. …….. Atomic scattering factor
F(hkl) ……Structure factor for reflection hkl
| F(hkl)|…….. Structure factor amplitude for reflection hkl
h, k, l ……..Reciprocal lattice points
I …….. Diffraction intensity
λ ……..Wavelength
LLG ……..Log-likelihood gain
MR …… Molecular replacement
P(uvw) ……..Patterson function
R ……Crystallographic R-factor
Rfree ……..R-factor for free set of reflection
RFZ…….. Rotation function Z-score
Rmerge…… R-factor relating agreement between symmetry related reflections
Rp.i.m ……Multiplicity weighted merging R-factor
Rwork …… R-factor for working set of reflection
TFZ…….. Translation function Z-score
u, v, w ……Patterson space coordinates
V …….. Volume of the unit cell
x, y, z ……..Real space coordinates

1 Chapter 1: Introduction

1.1 General background about bacteriophage

Bacteriophages are captivating microorganisms (viruses) that infect and invade bacteria. The name bacteriophage came from a Greek term (Phagen) meaning bacteria-eater by d'Herelles in 1917 (Fruciano and Bourne, 2007). The bacteriophages have played and still play very important roles (genetically and molecular biologically) in bacteria such as changing the phenotypes, converting non-pathogenic strains to pathogenic, or regulating bacteria populations in varying environmental conditions (Kovall and Matthews, 1997). The size of genomes, the diversity of the nucleic acid content in genotypes and the phenotype characteristics were used in order to classify bacteriophages. Although, this type of classification is currently followed, it is meaningless with regard to the evolutionary development. All bacteriophages were organized under the order of *Caudoviridae* in terms of possessing a tail, and later on, they were classified into three main families based on the size and contractibility of the tail. For example, the long contractible tail belongs to *Myoviridae* and the long non-contractible tail belongs to *Podoviridae*, whereas *Siphoviridae* represents those that are possessing a short non-contractible tail (Maniloff and Ackermann, 1998). In 1976, the first genome was completely sequenced (RNA-genome of bacteriophage MS2) by Walter Fries, and a year later Fred Sanger sequenced the complete DNA- genome (5386 bp) of Phage Φ-X174 (Koonin and Galperin, 2003).

Recently, it has become easy for scientists to study and understand the evolution of bacteriophages using either isolated bacteriophages to obtain direct sequences or sequence analyses that reveal the presence of pro-phage sequences within bacterial genomes (Canchaya et al., 2003). The presence of powerful facilities for sequencing with developing technologies has created bioinformatics databases, which have at least 750 sequenced bacteriophage genomes for study for example at the EBI (European Bioinformatics institute) (Hatfull and Hendrix, 2011). It has become easy to compare many phage sequences in the whole genome background, as well as to analyze the phage gene content in order to classify genomes functionally (Rohwer and Edwards, 2002). Bacteriophage genomes contain many commonly related gene patterns, which

are classified as mosaic genes, and which have the capability to be recombined during the genetic exchanges between phages inside the infected cell based on the molecular theory of phage evolution. Generally, many genes are carried on even the smallest phage genome, all of which are essential for different functions, such as replication, transcription, lysis, regulation, and morphology. However, the precise role of many genes has not been discovered. The genome of *Myovirusstop* T4 consists of 230 genes, of which only 130 genes are characterized functionally. Increasing the size of a genome will increase the complexity of a bacteriophage and the interference with the host cell activities (Rohwer and Edwards, 2002).

Comparative genome analysis, which reveals the similarities and the differences between individual genes, gene clusters, and whole genomes, assists phage classification and taxonomy by using phage genome sequences for the investigation of phage evolution. This helps a full understanding of the relationship between an individual phage and phage groups (Proux et al., 2002).

1.2 Lambda phage

Lambda phage (*λ*) is considered as one of the most popular bacteriophage for genetic study. It was first discovered in 1951 by Esther Lederberg (University of Wisconsin, Madison), and was the second phage genome to be sequenced (Sanger et al., 1982). It was discovered in *Escherichia coli* (strain K12) that had been mutated by ultraviolet irradiation, after finding unusual plaques with turbid centers in certain mutant stocks. These strains were identified as lysogenic as a result of carrying *λ phage.* It was the starting point for the study of lambda lifestyles (lytic and lysogenic). The importance of *λ phage* increased experimentally in 1961, after the discovery of a group of genes from suppressor mutants of *λ* (*amber* mutations) by Allan Campbell, that are responsible for the *λ phage* lytic cycle. In 1968, Clarence first identified pro-phage mutations, which were fatal for lytic growth of *λ phage*, after the isolation of more genes from amber mutations by Sandy Parkinson (Mount et al., 1968).

A genetic study successfully revealed many mutations of *λ phage* that were identified as essential genes for plaque formation. However, some genes were identified that were

not part of plaque formation but were considered essential due to their roles in the lytic cycle. These included the Rz and Rz1 genes that encode two small lysis proteins, which are responsible for disrupting the outer membrane of host cells and necessary for cell lysis (Berry et al., 2012).

Additionally, the genetic analysis of lysogeny also discovered the small *cro* gene and revealed that the absence of this gene affects lytic growth and leads to the lysogenic state (Eisen et al., 1970). The actual length of lambda DNA was determined in heteroduplex form by electron microscope analysis that replaced the recombination frequency units for construction of a Lambda gene map in base pairs (Eisen et al., 1970). Structurally, the lambda phage particle consists of a head and a tail, which can have tail fibers (Figure 1.1 A). The head contains one DNA molecule, which is double-stranded, linear, and has 48490 base pairs (Little, 1967).

The *λ* genome was found to have complementary 12 base single strand parts at each 5' end (non-blunt or *cos* ends), and these segments play a very important role as sticky ends to rebind or circle the *λ* DNA after invading host cells (Figure 1.1 B) (Hershey and Burgi, 1965). The first attempts to sequence the lambda DNA for genetic and biological interests were in 1968 by using these cohesive ends of the lambda genome (Wu and Kaiser, 1968), but it was not easy to determine the exact sequence of 12 bp. However, the complete genome of lambda virus was determined in 1982 as dsDNA using the termination method of dideoxynucleotide chain sequencing. The phage often has 12-14 different large protein complexes with more than 1000 protein molecules creating the entire particle (Sanger et al., 1982).

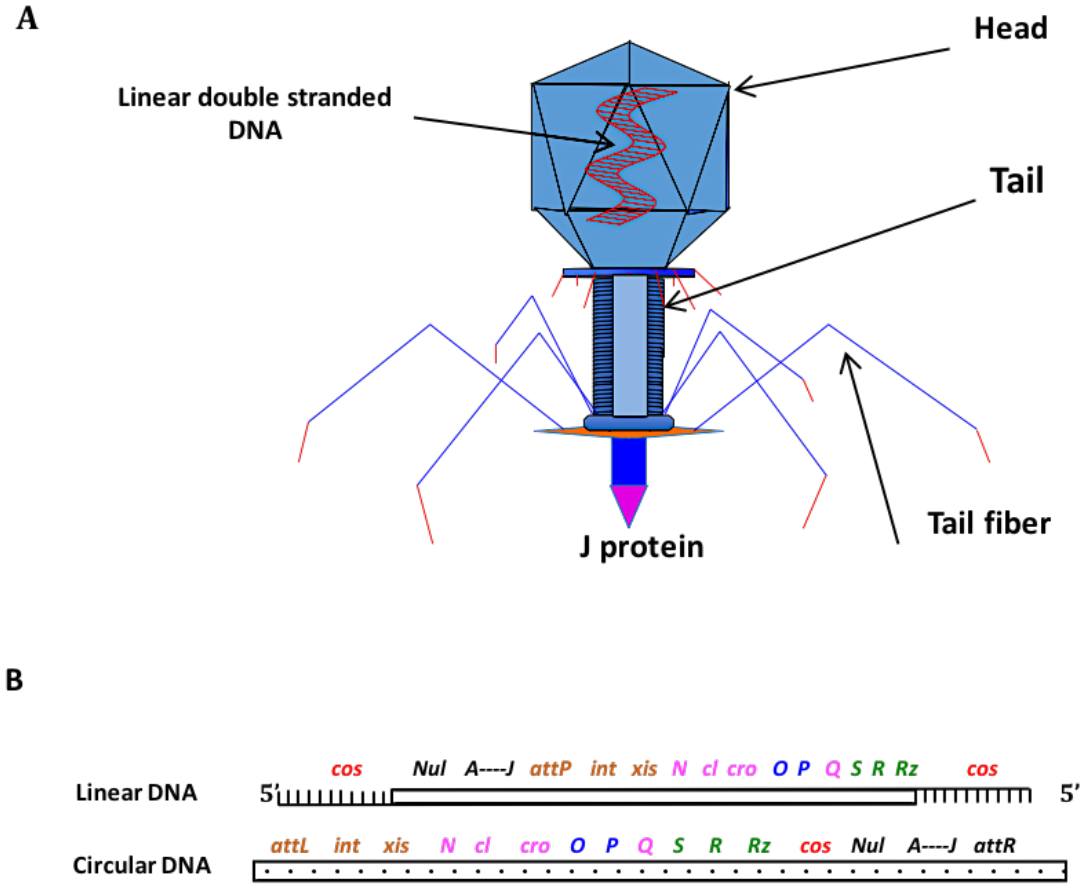


Figure 1.1: λ Bacteriophage particle.

A) Bacteriophage lambda virion structure with its main components (Head, Tail, Tail fiber, and Genomic DNA). **B**) The λ DNA after invasion (linear and circular forms) shows the main functional regions represented by the essential proteins. For example, Lysis proteins (green), replication proteins (blue), recombination proteins (Browen), regulation proteins (pink), and cohesive (cos) ends(red) comprising 12 complementary nucleotides at each 5' end that are responsible for circularizing the λ DNA after injection into a host bacterial cell. This figure is drawn by PowerPoint program.

1.3 Lambda phage replication in *E. coli*

Generally, the first step of a replication cycle is the attachment of the λ phage to the surface of the host cell by the interaction between the λ phage J-protein, which is encoded by the 3.4 kbp J gene, at the phage tail tip (Katsura, 1983), and the *E. coli* host cell outer membrane Lambda receptor, which is encoded by the *LamB* gene and provides non-specific channels to pass different types of hydrophilic molecules such as maltose and maltodextrin through the cell membrane (Ferenci et al., 1980).

When the lambda phage infects the *E. coli* cell, two different lifecycles may occur: lytic or lysogenic (Schwartz, 1975) (Figure 1.2).

1.3.1 Lytic cycle

The lytic cycle starts after injecting the DNA of lambda phage across the host cell outer membrane and then the inner membrane via the mannose permease complex (manX, Y, Z proteins) into the host cell (Erni et al., 1987). The λ phage linear nucleic acid genome immediately will be circularized via its single-stranded sticky ends by using the host DNA ligase protein, and is then replicated. The initial RNA transcription occurs without phage involvement from P_L and P_{RE} promoters based on the action of the host *rho* protein, and the initial transcription pathway ends after the phage *N* and *cro* gene sequences (Roberts, 1969). The N protein is produced due to the translation of P_L short RNA and allows continuing RNA transcription and inhibits the activity of *rho* protein, as well as inducing the λ recombination functions. On the other hand, the *cro* protein is yielded by the translation of the initial RNA transcript from $P_{RE,}$ which induces the replication function (Franklin, 1974). The phage genes are transcribed and translated to allow production of phage proteins. The phage proteins are assembled to form phage particles, which are released from the host cell after breaking down the cell wall via the activity of the lysis proteins (R, S, Rz, and Rz1) that work together to disrupt the host cell wall (Brüssow 2005; Sao-Jose et al. 2000) (Figure 1.2).

1.3.2 Lysogenic cycle

The lysogenic cycle is second method of λ phage reproduction, which does not produce a progeny phage or lyse the infected cell. The λ DNA can form a circular replicon in the cytoplasm but alternatively can integrate in to the host cell genome to produce a prophage. This naturally is the ideal pathway for the λ phage when environmental conditions prevent its replication in the bacterial cells as with the lytic cycle. In this case, the host cells stay alive and reproduce normally carrying a prophage. The prophage DNA replicates along each time with the host cell replication itself, yielding more cells with copies of the prophage. However, exposure to ultraviolet or certain chemicals induces the prophage DNA to separate from the host genome and proceed to the lytic cycle as shown in Figure 1.2 (Campbell and Reece 2005).

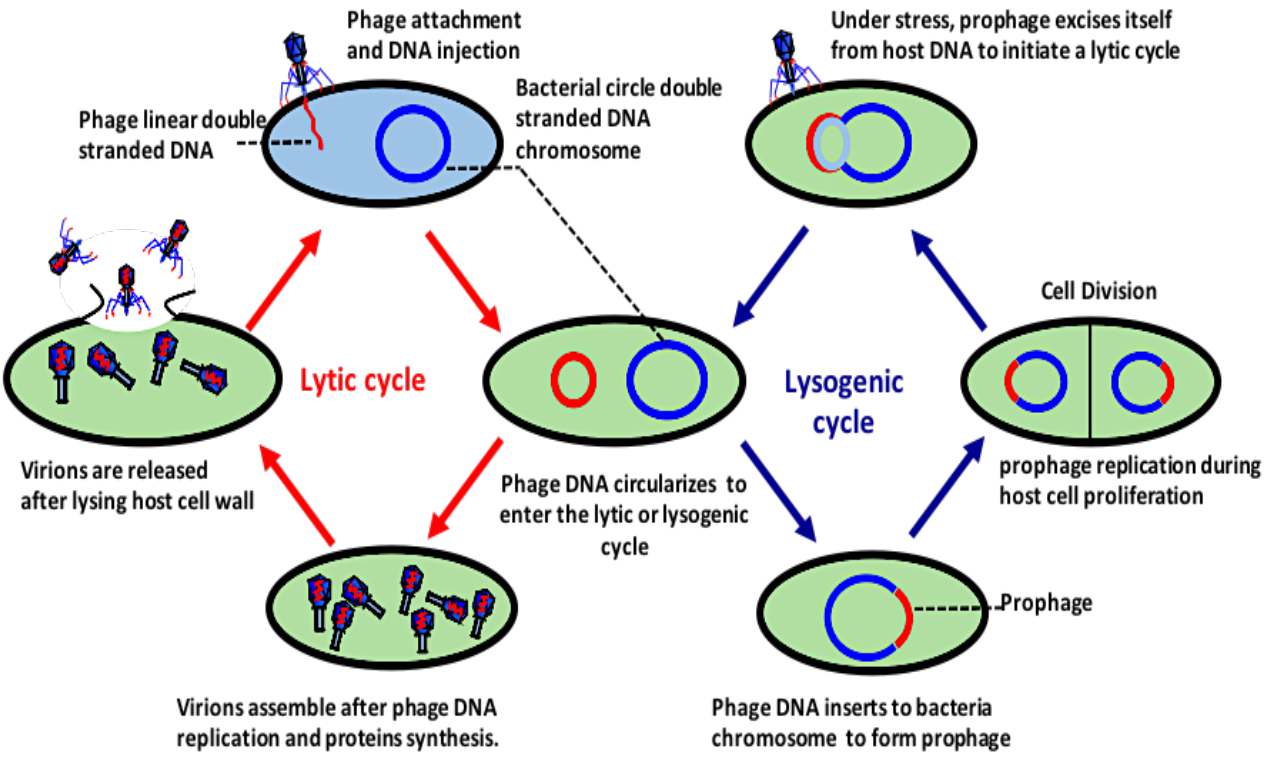


Figure 1.2: Replication cycle of λ phage.

The figure shows the lytic and lysogenic cycles of λ phage. When the lambda phage injects DNA into the host bacterial cell the ends join to form a circular DNA molecule. The bacteriophage can proliferate in *E. coli* by a lytic pathway (red), which destroys the cell, or it can enter a latent prophage state (blue). A lambda prophage can be induced to exit from the host chromosome and enter the lytic cycle if damage to the host cell happens. This figure is drawn by PowerPoint program.

1.4 Recombination system

During the replication cycle of λ phage, research showed its capability to drive a specific recombination system for repairing DNA double strand breaks or lysogenic use. The lambda phage system is related to the general recombination pathways in prokaryotes as described in the following section.

1.4.1 *E. coli* recombination system

Single-stranded or double-stranded DNA breaks are potentially lethal forms of DNA damage present in all organisms (Kuzminov, 1999). To repair these lesions and reduce their threats, organisms have developed numerous mechanisms. One of the fundamental processes is recombination (Didelot et al., 2012). Recombination is imperative also for the genetic divergence of bacterial genomes in many ways besides the housekeeping role it plays in repair (Rocha et al., 2005). Firstly, recombination permits the integration of homologous foreign DNA resulting from conjugation or transformation in bacteria (Smith 1991; Lorenz & Wackernagel 1994). Secondly, it assists adaptive mutations and eradicates harmful mutations (Otto and Michalakis, 1998) by permitting allelic recombination between closely associated strains (Feil, 2004). Thirdly, chromosomal instability can result from recombination between homologous sequences in the genomes of related bacteria (Hiom, 2009). In bacteria, the chromosome rearrangement rate correlates with the number of repeated segments in the genomes (Rocha, 2003). Fourthly, recombination is also involved in production of genotypic diversification, e.g., in pathogens owing to intra-chromosomal homologous recombination between large repeated segments (Finlay & Falkow 1997; Mehr & Seifert 1998; Rocha & Blanchard 2002).

Two types of recombination are recognized. One strategy, which has been used extensively in archaea, some bacteria and eukaryotes, is to re-join broken DNA ends, irrespective of the DNA sequence at the break in a process known as non-homologous end-joining (NHEJ) or non-homologous recombination. Nevertheless, this mechanism is perhaps mutagenic because one or more nucleotides are frequently absent at the break site (Hiom, 2009). A second strategy is homologous recombination, in which a genome portion is substituted by the suitable sequence from another genome (Didelot and Maiden, 2010), and allows more precise DNA breaks repair. This process uses, as a

template, the other undamaged copy of the chromosome made during DNA replication, to truthfully restore the genetic information at the repaired break. Furthermore, homologous recombination is considered the prevalent pathway in bacteria such as *E. coli* for repairing DNA breaks (Hiom, 2009). It plays a crucial role in all organisms for the repair of DNA damage of exogenous and endogenous origin (Kuzminov, 1999).

Homologous recombination itself involves two pathways: the RecBC pathway and the RecF pathway (Touchon et al., 2009). One of these pathways is used for recombination between DNA molecules with short regions of homology, whereas the other pathway proficiently stimulates recombination between DNA molecules with longer sections of homology (Watt et al., 1985). However, in *E. coli,* either the RecBCD or the RecFOR pathway may start homologous recombination (Rocha et al., 2005).

In *E. coli*, the homologous recombination occurs primarily through the RecBCD pathway of generalized recombination. Genetic analysis shows that recombination occurring through this pathway relies on the RecA, RecBCD, and single-stranded DNA binding (SSB) proteins (Clark & Margulies 1965; Howard-Flanders & Theriot 1966; Emmerson & Howard-Flanders 1967; Glassberg et al. 1979) (Figure 1.3).

RecA protein was shown to stimulate both the DNA strand exchange between complementary single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA) and the renaturation of homologous ssDNA. A clear structural and functional resemblance was found between RecA and RAD51 recombinases of bacteria and eukaryotes, respectively, which verify the idea that homologous pairing and strand exchange are mechanistically analogous in all organisms (Benson et al., 1994). RecA inactivation in many organisms was found to cause a spectacular increase in the sensitivity to all DNA-damaging agents used in laboratories (Rocha et al., 2005). The SSB protein was demonstrated to stimulate the ability of RecA protein to exchange complementary DNA strands *in vitro* (Kowalczykowski 1991; Kowalczykowski et al. 1994). It was found that mutations in the *recA* gene reduce recombination levels by approximately six fold (Clark and Margulies, 1965), while the *ssb* gene defects can drop recombination by five fold (Glassberg et al., 1979).

RecBCD is a multi-functional enzyme complex that contains three non-identical subunits, the RecB, RecC, and RecD polypeptides. It has many biochemical activities *in vitro* including ssDNA and dsDNA exonuclease, ssDNA endonuclease, DNA-dependent ATPase, and ATP-dependent DNA helicase (Spies and Kowalczykowski, 2005). In addition to its nonspecific nuclease activities, RecBCD enzyme can produce ssDNA fragments whose 3'-end terminates 4-6 nucleotides to the 3'-side of the χ recombination hotspot (Ponticelli *et al,* 1985; Taylor *et al,* 1985). This specific interaction is based on orientation, as RecBCD enzyme should come near the 3'-side of the χ sequence in order for recognition to happen (Taylor *et al,* 1985). Mutations in the *recB, recC*, or *recD* genes create modified RecBCD enzymes with various impacts concerning χ-stimulation of recombination. (Howard-Flanders & Theriot 1966; Emmerson & Howard-Flanders 1967).

The RecF pathway has an essential role in bacteria to repair DNA breaks and gaps. While this pathway is mainly implicated in the homology-directed repair of single-stranded DNA (ssDNA) gaps created during replication of a damaged DNA template, it can also replace RecBCD in dsDNA break repair (Hiom, 2009). In contrast to the RecBCD pathway, numerous components of the RecF pathway possess structural or functional counterparts in archaea and eukaryotes as well.

1.4.2 Mechanism of Recombination for DNA double-stranded breaks

In *E. coli*, the jointed molecules generated by RecA may be resolved through the action of the RecG helicase or via the RuvABC complex. Following strand exchange, postsynaptic steps occur that involve the creation of a viable recombinant molecule and cause the resolution of the recombination intermediates generated by RecA (Rocha et al., 2005). It was reported that the RuvAB protein enhances the branch migration, while the RuvC protein catalyzes the resolution of recombination intermediates known as Holliday junctions. It was found that the RuvABC proteins work as a resolvasome complex to track all along DNA, using the ability of a RuvC protein for the determination of cleavable regions within the DNA sequence as the DNA passes through the complex (Rocha et al., 2005).

During recombination in *E. coli*, DNA molecules synapsis (Pairing of homologous DNA) includes the homologous alignment of ssDNA with dsDNA to form a nascent heteroduplex structure (reviewed by Radding 1982). Compared with the presynapsis and the strand exchange, synapsis is a fast process (Wu et al. 1982; Kahn & Radding 1984). The first step of synapsis known as conjunction, which is independent of sequence homology, comprises the dsDNA invasion by ssDNA (Gonda and Radding, 1983). In the second step of synapsis, complementary sequences will undergo homologous alignment either by facilitated diffusion or by translocation of the DNA molecules in a tripartite complex with the RecA protein (Gonda and Radding, 1983). Prior to strand exchange, enzymes known as presynaptic enzymes permit the creation of a RecA filament, which is a complex of ssDNA covered with RecA particles, by making the DNA accessible to RecA (Rocha et al., 2005). Occurrence of strand exchange will create long heteroduplex joints between the DNA molecules (Wu et al., 1982).

DNA double-strand breaks (DSBs) are general lesions that arise in all cells (Motamedi et al., 1999). They can occur due to damage of DNA resulting from processing of arrested replication forks (Seigneur et al., 1998) and are assumed to arise from normal intermediates in DNA replication (Kuzminov, 1995). Because of the deleterious effects that the DSBs cause to the cells, multiple mechanisms are developed by cells for their repair (Motamedi et al., 1999). Homologous recombination may be the unique mechanism for double-stranded DNA repair (DSBR) *in E. coli* (Haber, 1999). This process involves four biochemical steps: processing of DNA molecules into appropriate recombination substrates; homologous pairing of the DNA partners and DNA strand exchange; extension of the emerging DNA heteroduplex; and resolution of the resulting crossover structure (Kowalczykowski et al., 1994).

Homologous recombination in *E. coli* is initiated based on either the RecBCD or the RecFOR pathway. Both pathways create a ssDNA molecule coated with RecA to facilitate the homologous molecule invasion (Kowalczykowski et al., 1994). RecBCD complex has been shown to be involved in the repair of DSBs. Two basic mechanisms of RecBCD-mediated recombination and DSBR in *E. coli* have been proposed including: a break–join pathway that operates depending on endonucleases called Holiday junction resolvases and a replicative pathway that needs DNA Pol III but is independent of resolvases.

These pathways have been suggested as alternate outcomes for strand-exchange intermediates (Motamedi et al., 1999). Approximately 50% of all host cell RecBCD-mediated recombination of phage lambda DNA results from a replicational recombination mechanism (Motamedi et al., 1999). Whereas, the RecFOR pathway, which participates in the repair of single stranded DNA (ssDNA) gaps, is associated with the early steps of homologous recombination-mediated DSBR including DNA end-processing and homologous pairing (Handa et al., 2009). Double strand break repair in *E. coli* was thought to be the consequence of breakage and reintegration of parental DNA molecules, and this process has been found to be mediated by the Holliday junction resolvases (Motamedi et al., 1999). By means of homologous recombination, the RecF pathway components were shown to act in a concerted manner to raise DSBR (Handa et al., 2009).

The main common step in all homologous recombination pathways is to recruit recombinase, i.e. RecA in bacteria. In cells, ssDNA is rapidly covered with SSB (single strand binding protein), which must be replaced in order to load RecA. This process is facilitated by RecF, RecO and RecR (Morimatsu and Kowalczykowski, 2003). It was found that a complex of RecO and RecR binds to DNA covered with SSB, replaces the SSB from DNA and enables RecA binding (Umezu et al., 1993). Then, after binding cooperatively to ssDNA, RecA will create a nucleoprotein filament with a distinct 5'–3' polarity along the DNA (Hiom, 2009). This filament, in turn, steers the pairing and exchange of homologous DNA molecules, Although the broken DNA ends need to be processed first to produce ssDNA tails where RecA can be loaded (Hiom, 2009).

It was found that the RecA, RecBCD and SSB proteins are necessary for homologous pairing between duplex DNA substrates, which depends on stimulation from the presence of a χ-sequence in the donor linear dsDNA (Dixon and Kowalczykowski, 1993) (Figure 1.3). In the RecBCD pathway, all the essential components including those involved in helicase, nuclease, and RecA loading are accumulated in a single holoenzyme (Seigneur et al., 1998). RecBCD binds to the ends of dsDNA causing unwinding and degradation of DNA until it encounters a χ site (Rocha et al., 2005).

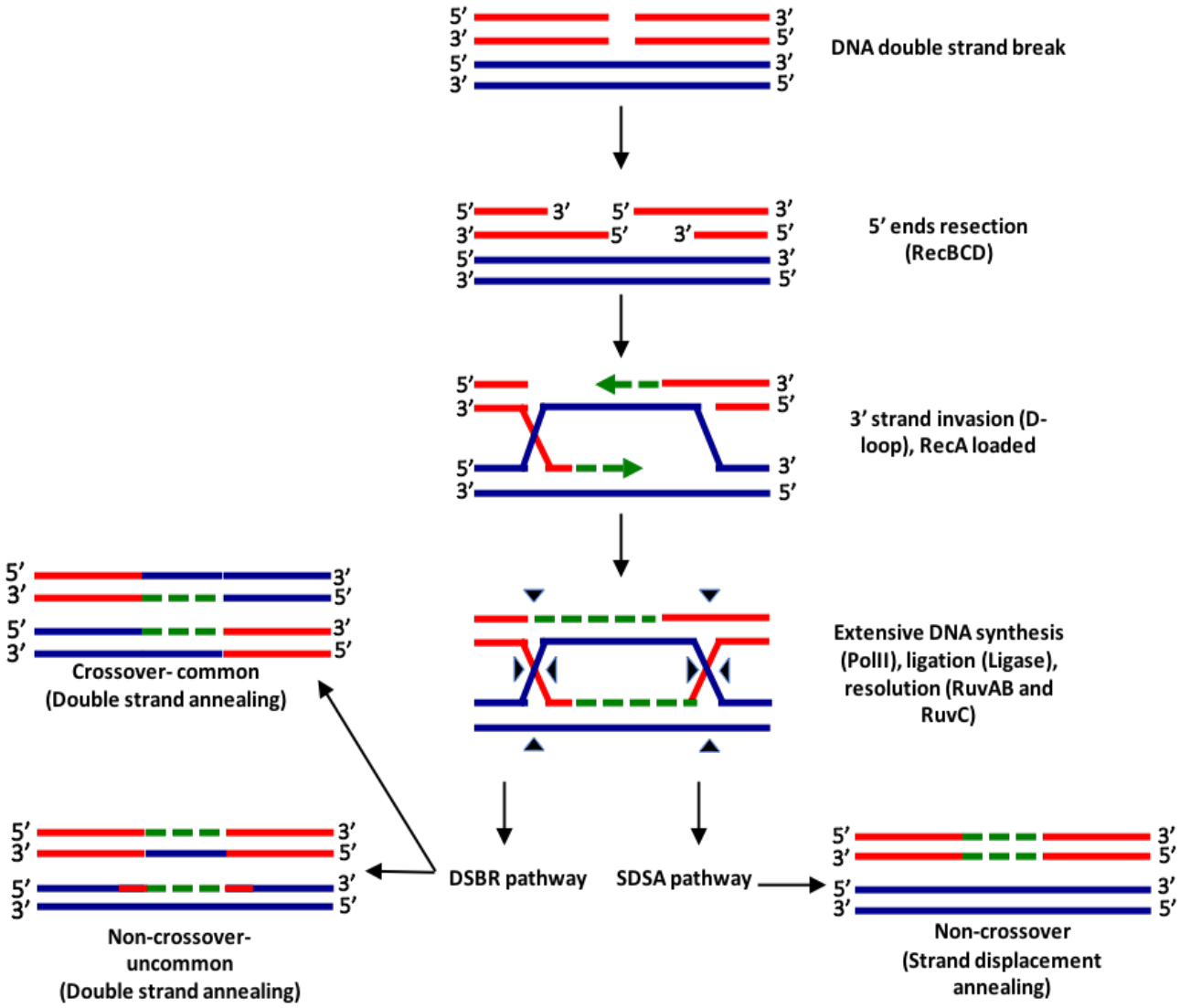


Figure 1.3: Mechanism of recombination for DNA double stranded break.

The figures shows a repair pathway of DNA double strand break and strand displacement annealing, which share the same initial stages, and then differ in the final stage. The SDSA (strand displacement annealing) result always gives a non-crossover product, whereas the DSBR (double strand break repair) always gives a crossover product. This figure is drawn by PowerPoint program.

1.5 Red pathway system of λ phage

The lambda phage contains a powerful recombination system (Red Pathway system) used in molecular biology for generating specific defined alterations in chromosome, plasmid or bacterial artificial chromosome (BAC) of *E. coli* genetically, such as insertion, deletion, and point mutations. It needs only 35 bp of homologous sequence on both sides of the desired change to start recombination (Thomason et al. 2007; Sharan et al.

2009). Although, there is uncertainty about the intermediates that are involved in the mechanism of this recombination system, the Red system has commonly been used over the past decade for modifying plasmid, BACs, and chromosomal targets of *E. coli* by using single-stranded DNA oligonucleotides (Mosberg et al., 2010). It was used for example in the production of lycopene via fast optimization of its biosynthetic pathway (Wang et al., 2009). Additionally, the Red Pathway system was used for linear double stranded DNA recombineering: to exchange chromosomal genes and knock-out gene function (Murphy 2000; Datsenko & Wanner 2000) to improve original cloning strategies; to remove 15% of the *E. coli* strain genome; to create a library of knocked out genes in *E. coli*; and to insert heterologous genes or complete pathway into the *E. coli* genome (Mosberg et al., 2010).

The Red pathway system in lambda phage performs distinct reactions of a homologous recombination that contribute to genomic variability and DNA repair. A main function of the recombination system is to repair double strand DNA breaks due to exposure to ionizing radiation, collapse of replication forks, restriction endonuclease incision or cleavage by terminase protein (Figure 1.4) (Takahashi and Kobayashi, 1990). The double-strand invasion and single-strand annealing pathways are distinct models for creating recombinant progeny. The initial study of the Red system assumed that the mechanism of this recombination occurs via strand invasion (Thaler et al., 1987). This assumption was rejected later due to the unreliability of strand invasion to be a leading mechanism in the absence of long homologous regions, and thus it was more likely to be under recA control. Moreover, recent research showed that the Red recombination mechanism tends toward strand annealing rather than strand invasion based on the consistent characteristics of recombination products. This mechanism targets the lagging strand during the replication process (Mosberg et al., 2010).

The mechanism of the Red system involves various λ-encoded proteins, which are functionally equivalent to the *E. coli* recombination proteins. For instance, phage lambda contains the so-called Red system, named after a partial recombination-deficient phenotype of the phage when the associated lambda genes are inactivated. The lambda Red system proteins Exo, Bet and Gam are analogous to the *E. coli* RecBCD enzyme, the λ Orf protein, which is necessary for early events of recombinational exchange, is

equivalent to the *E. coli* RecFOR system and works in its absence (Sawitzke & Stahl 1994; Poteete 2008; Rybalchenko et al. 2004), and the λ ninG resolvase (Rap) is an equivalent of the *E. coli* RuvC resolvase (Webb et al. 1997; Court et al. 2002).

Three main proteins (Gam, Exo, and Beta) are involved in the λ phage Red pathway system for carrying out double stranded DNA recombination. These proteins are located in the early left operon (PL promoter) in the tail fibre region of the lambda genome, and their expression is controlled by *cI* repressor (Poteete, 2001). A 12 kDa Gam protein is responsible for stopping linear double stranded DNA degradation of the lambda genome via deactivating the function of *E. coli* RecBCD and SbcCD nuclease proteins. A 24 kDa Exo protein is a lambda exonuclease protein (resection protein) that is responsible for revealing the single stranded DNA regions by binding and degrading the dsDNA at both 5' ends to produce 3' overhang ends (Szczepańska, 2009) (Figure 1.4), then a 28 kDa Beta protein anneals these single stranded ends at an homologous site in the host genome (Sawitzke et al., 2007).

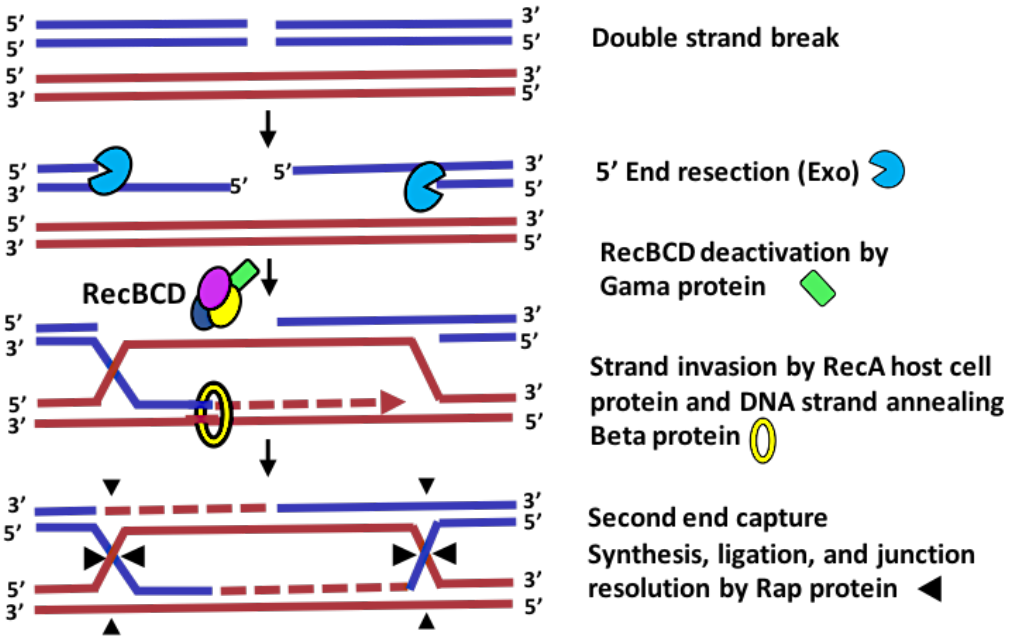


Figure 1.4: Red pathway system.

The repair pathway of DNA double strand break by the Red system proteins of lambda phage is shown. This system involves four main proteins that have specific roles during the repair process. The exonuclease, Exo, efficiently resects the 5' -termination strand of dsDNA, the Bet protein binds to ssDNA and initiates strand annealing process. The activities of the bacterial RecBCD enzyme are inhibited by Gam protein, and then the endonuclease protein Rap cleaves the DNA junction intermediates to form the final products. Host DNA is the blue lines and the red lines represent Lambda DNA. This figure is drawn by PowerPoint program.

During the λ replication cycle, the lambda phage will switch off the host cell recombination system and switch on the Red pathway system. A DNA double-strand break is generated in each λ progeny at the *cos* (the cohesive site). The repair mechanism for the double-strand end break includes interaction with homologous DNA by two pathways. Both suggest a start with an exonuclease-mediated processing of the double-strand end that recesses the 5‘end strand, while the 3’-end strand is maintained, thus producing a 3‘end single-strand overhang (Resnick 1976; Krasin and Hutchinson, 1981). The *E. coli* RecA protein binds to the processed 3’-ssDNA-overhang created by the lambda Exo protein. This stage permits the assembly of multiple RecA monomers, which form a presynaptic filament on ssDNA to ease invasion of a homologous duplex in the presence of a homologous partner ssDNA (Stahl et al., 1997). An alternative mechanism is available when a cut is present in each of two homologous duplexes and in this case, the λ system uses recombination with Exo-processed substrates via a RecA-independent single-strand annealing pathway (Figure 1.5). The lambda Bet protein is required as a λ-encoded ssDNA binding protein to anneal partner ssDNA (Figure 1.5) (Court et al., 2002).

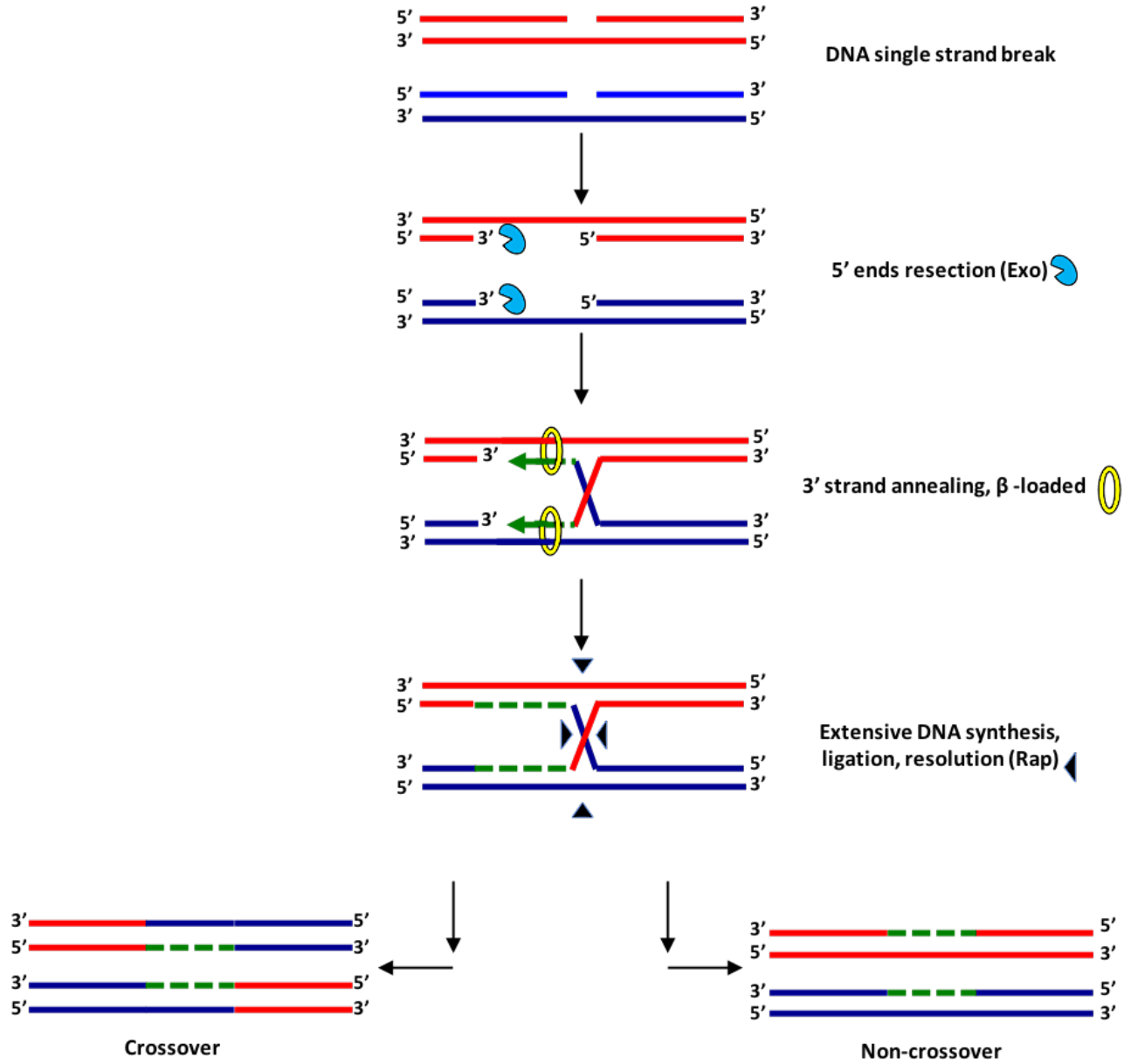


Figure 1.5: Mechanism of repair for two DNA duplex breaks.

A repair pathway for two DNA duplex breaks is shown, which shares similar initial and final stages as for repair of breaks using the RecA-independent ssDNA host protein but without the need for an invasion step. In this mechanism, the same Red pathway proteins are used (Exo, Bet, Gam and Rap). The final stage results in either a non-crossover product or a crossover product. This figure is drawn by PowerPoint program.

Since the annealing mechanism for binding complementary single strands is catalysed by a variety of proteins, the distance between the two interacting ends plays a very important role in determining the length of a generated hybrid DNA segment. For example, if the distance is 10 or 20 kb, annealing may produce a segment of hDNA longer than this formed by an invasion pathway (van de Putte et al., 1966).

1.6 NinR region

Comparative analysis studies of the genome organization in different lambdoid phages proposes that the *ninR* region establishes a genetic module containing nine short open-reading frames (ORFs) involved in DNA metabolism (Kriiger and Hobom, 1982). It was found that two of these ORFs are involved in the recombination pathways (Sawitzke and Stahl, 1992). The NinB protein induces an activity that replaces recombination defects in *E. coli* recFOR mutation (Tarkowski et al., 2002), whereas NinG (Rap) is a resolvase for branched DNA substrates (Sharples et al., 2004). Other Seven ORFs were found to encode uncharacterized function proteins (Court and Oppenheim, 1983).

1.7 Rap endonuclease (*ninG*)

A number of λ phage genes are expressed during the lytic cycle that participate in the recombination repair of DNA double-strand breaks, and suggest a splice recombinant as a consequence of nicking and ligation of a 3-stranded (D-loop) intermediate, which can initiate DNA replication (Figure 1.4) (Stahl et al., 1985). Although a splice–type product may be produced via nicking at specific points in the D-loop, a Holliday junction structure (a 4-stranded structure) may also be generated in the adjacent duplex DNA of the invading molecule during further strand exchange (Figure 1.3). These junction structures can be resolved by the introduction of symmetrically associated incisions (Figure 1.6)(White et al., 1997). Recent studies revealed that the lambda phage genome contains a gene called *rap* or *ninG*, located in the non-essential *ninR* region (Figure 1.8) (Sharples et al., 1998). This gene encodes a specific endonuclease, which targets the recombination intermediates generated by the Red system, and is found in several lambdoid phages including P22, H-19B, 933W and PS34 (GeneBank accession nos X78401, AF034975, AF25520 and AJ011580, respectively) (Sharples et al., 1998). Additionally, the previous studies on the mutated *rap* gene showed a hundred time

increase in RecBCD-dependent recombination between λ and a plasmid linked with the presence of Rap protein, which showed a bias toward the formation of splice recombinants (Stahl et al. 1985; Hollifield et al. 1987).

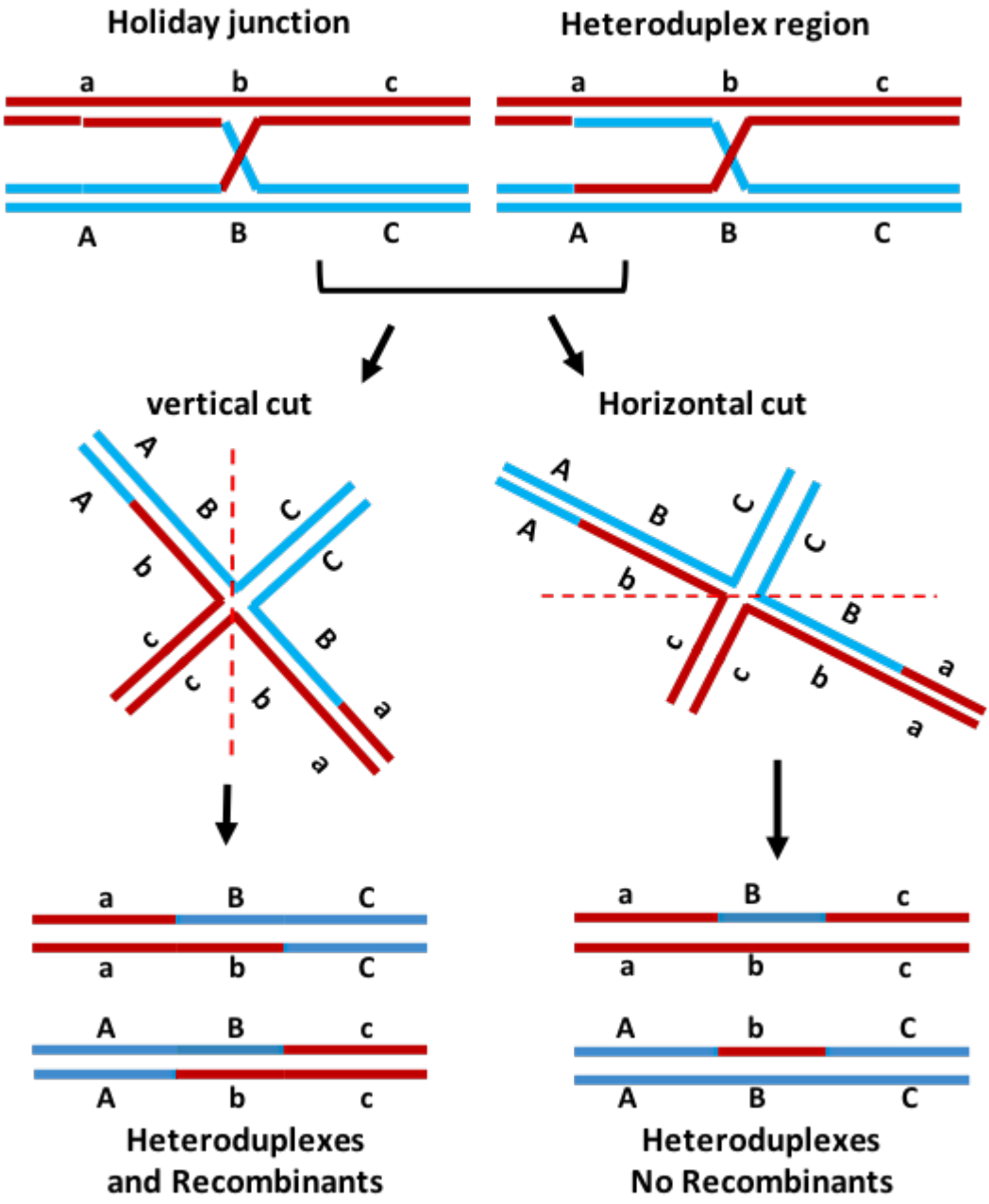


Figure 1.6: Resolution of DNA 4-stranded structure.

Figure shows the resolution steps for the DNA Holiday and Heteroduplex junctions. Different products are produced depending on what is the cut vertically or horizontally. This figure is drawn by PowerPoint program.

A comparison of the gene organization for *rap* and the *E. coli* lambdoid phage 82 *rusA* gene suggests the function of the Rap protein is perhaps similar to the RusA protein, which acts as a resolvase for DNA junction structures. However, there is no sequence similarity between their genes. RusA has a strong preference for cleavage of the Holliday junction structure (4-stranded junction) with an ability to bind to various branched DNA substrates (Mahdi et al., 1996). Rap can resolve 3-stranded junctions, such as D-loops, flaps and Y-junctions, as well as an ability for resolving Holliday junctions by symmetrically paired incisions like RusA (Sharples et al., 1998). This activity can generate splice recombinants by nicking at the branch point of Holliday and D-loop substrates in λ recombination in the presence of manganese ions (Sharples et al., 1998).

Rap protein was shown to bind and resolve various branched DNA structures, such as Holliday junctions, Y-junctions and flaps but, unusually for a resolvase, without affecting the angles formed between the duplex arms. Additionally, it was found that magnesium ions are required at the active site, and their presence may also influence junction conformations to a very limited extent (Sharples et al., 2004).

1.8 NinH protein

The *ninH* gene is situated within the ninR region between the genes of replication proteins O and P and the Q anti-terminator. Its location is adjacent to the lysis cassette, and directly downstream of *rap* (*ninG*) (Poteete et al., 2004) (Figure 1.9). It encodes a small protein of only 68 amino acids. The N-terminal region of the protein appears to have a putative helix-turn-helix (HTH), which shares significant homology with the DNA binding region at the C-terminal end of *E. coli* Fis (Factor for Inversion Stimulation).

Fis is one of the nucleoid-associated proteins (NAPs), which are involved in chromosome packaging to form the nucleoid in the cytoplasm of bacterial cells (De Vries, 2010; Browning, D. F., *et al,* 2010), and it is responsible for untwisting the supercoiled nucleoid structure (Postow et al., 2004). Therefore, Fis can expose DNA to transcription within the superstructure. The conformational shape of the chromosome is influenced by NAPs by bending and unbending the DNA based on the requirements of the cell (Luijsterburg et al. 2008; Dillon & Dorman 2010). The Fis-DNA crystal structure revealed contacts between residues Asn73, Gln74, Thr75, Thr87, Arg89 & Lys90 and the phosphodiester backbone plus contacts to the bases by Asn84 & Arg85 (Stella et al., 2010). A sequence

and predicted structure alignment of NinH protein with Fis (Figure 1.7), suggested all the DNA contacting residues of Fis seen in the Fis-DNA crystal structures were present in NinH protein with the exception of N84 in *E. coli* Fis, which corresponds to a serine in NinH predicted structure (Dr G. Sharples, personal communication).

The DNA binding activities of Fis influence numerous DNA transactions involving genome replication, chromosome separation, transcriptional regulation, cell division and site-specific recombination reactions such as lambda integration and excision (Esposito & Gerard 2003; Travers et al. 2001). Moreover, the highest stable complexes of Fis protein were seen in the interaction between the HTH with regions of DNA containing regulatory genes (Dorman, 2010). The relationship of the NinH protein to the Fis protein may suggest a similar accessory function in gene regulation by NinH, perhaps of the open-reading frames within the ninR region. Alternatively, NinH might replace Fis activities in genome packaging, phage transcription, and regulation of site-specific homologous recombination (Trotter and John, 2015).

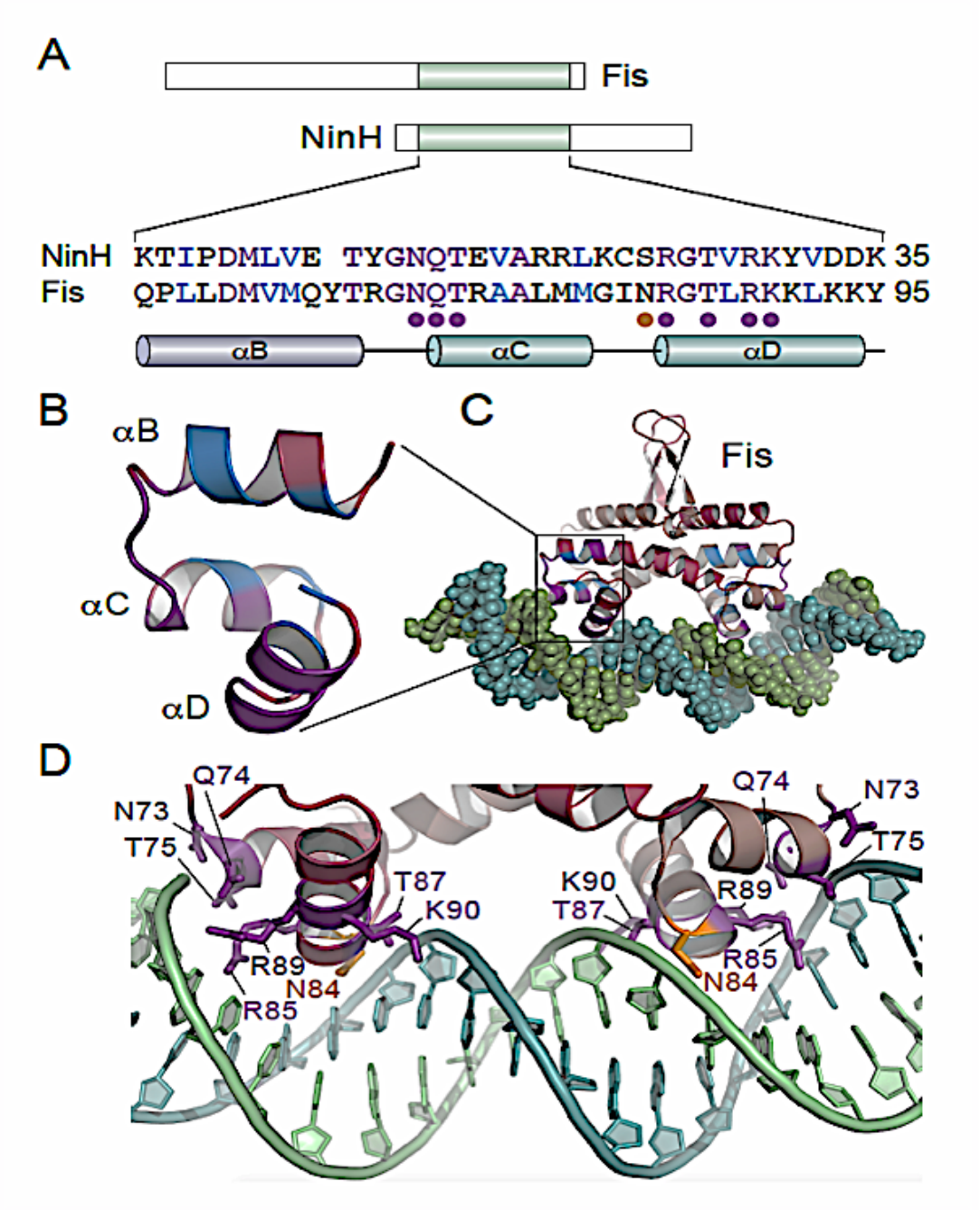


Figure 1.7: Sequence and structure alignment of λ NinH protein and E. coli Fis.

A) Sequence alignment between the 68 residue λ NinH and the 98 residue *E. coli* Fis. The region of similarity covers helix αB and the helix-turn-helix (αC and αD) binding motif of Fis. A filled circle below the Fis sequence indicates residues involved in DNA contacts. The identical residues are in purple, while the functional residues are blue. **B)** A modeled structure of NinH built by the Phyre2 server based on the *apo* structure of Fis (PDB 1FIP), homologous residues are colored as in part A. **C)** A structure of a Fis-DNA complex (PDB 3IV5) which shows the region homologous to NinH, binding to a DNA substrate. **D)** Residues in a Fis-DNA complex that are in contact with a DNA colored as in part A (Dr G.J. Sharples, personal communication).

1.9 Ea10 protein

The *ea10 (ssb)* gene is located in the phage genome between the *cIII* and *ral* genes and is under the control of the same promoter as other Red pathway genes. The gene encodes a 16 kDa Ea10 protein, which is the most plentiful protein to be expressed early after infection. It was believed that Ea10 protein is a single strand DNA binding protein (*ssb*) because of an apparent 60 kDa oligomeric form of the purified protein (Court and Oppenheim, 1983) and some fluorescent anisotropy data (Dr G. Sharples, personal communication), (Figure 7 & 8). Physiologically, an *ssb* product has a general impact on the host cell after lambda phage infection, for example the infection by a lambda *cro* mutant causes an extremely harmful influence on the host cell in the presence of *ssb* overexpression. These circumstances stop the growth of phage and host cell in 20 minutes and were named a Tro phenotype (Eisen et al., 1975). Although, the Tro inhibition affects growth by severely reducing the synthesis of the total RNA and protein, many host proteins such as gp*groE* and gp*dnaK*, which are necessary for phage growth and heat shock response, and the lambda gp*ssb* protein are still synthesized at a high level. It also involves many unidentified lambda elements. Deletion or point mutations in the *ssb* gene partially relieve the Tro effect (Georgiou et al., 1979).

It is not clear why the Ea10 protein product is involved in the Tro phenotype transaction. However, there are possible explanations for this. The direct or indirect interaction of Ea10 protein with replication proteins of the host cell, such as RNA polymerase or the ribosome or its initiation factors might lead to different ways to assemble **λ** proteins (Georgiou et al., 1979). The host transcription and translation processes are temporarily stopped under normal circumstances of early **λ** infection, before the **λ** late promoter starts the manufacturing of gene products. Alternatively, the host transcription and translation machinery must be returned to an active state when the **λ** phage is in a lysogenic state (Georgiou et al., 1979). It has been suggested that the role of Ea10 protein is to act as an essential key for making the early λ proteins with some host proteins, where the transcription and translation system is under phage control. These suggestions were strengthened via the initial experiments on i^{λ}c857*cro*27 infected cells tested at 40˚ for 90 minutes after infection that revealed the presence of Ea10 protein as one of a limited set of expressed phage-specific proteins (Oppenheim et al. 1977; (Georgiou et al., 1979).

Later research results revealed a significant effect of the *Ea10* product, which suggested that it might act as a protein regulator during a noticeable increase in the expression level of some of cellular proteins (gp*groE* and gp*dnaK)*, when the host cells were infected by *bio*252 mutants (*ea10*$^{+}$ and *ral*$^{+}$), but not after infection with *bio*10 mutants (*ea10*$^{-}$ and *ral*$^{-}$) (Drahos and Hendrix, 1982). Recently, it was found that the Ea10 protein is capable of targeting protein synthesis groups in *E. coli*, such as ribosomal proteins RpmA and RpsG, hypothetical protein (YIiL), YcbG (Cell division protein), and Primosomal replication protein (PriC) (Blasche *et al,* 2013). Additionally, a yeast two hybrid screen suggested that it might bind to the Rmf protein (Ribosome modulation factor), the CobB protein (an NAD-dependent deacetylase), and SoxS (Regulatory protein) (Blasche et al., 2013).

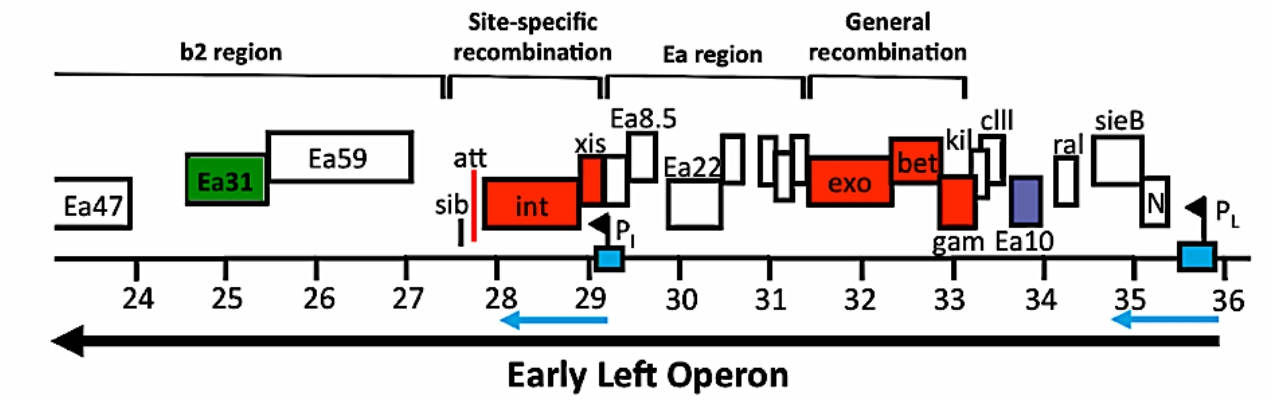


Figure 1.8: Map of the early left operon in λ phage.

Recombination proteins are found in the early left operon. The genes in this region are known to play a role in the Red pathway system. The *ea10* gene is under the control of the same promoter as the three main proteins (Exo, Beta and Gam) in double-stranded DNA repair. This figure is drawn by PowerPoint program.

1.10 Ea31 protein

In 1971, Hendrix identified a group of **λ** proteins of various sizes on polyacrylamide gels found during the infection stage of UV irradiated cells. The genes encoding these proteins are present in the *b* region within the lambda genome (Oppenheim et al., 1977) (Figure 1.8), and a study of the expression of the early b region proteins suggested the requirement for an active P_L promoter to initiate transcription (Nijkamp et al. 1971: Oppenheim & Oppenheim 1978). This transcription ends at the termination signal for the *ea47* gene. This termination can be overcome by the presence of the *gbN* gene product, which is the first gene transcribed from P_L promoter. (Gottesman et al. 1980; Honigman 1981). Deletion mutations were applied to the **λ** b region in order to determine the locations of early and late encoding genes based on their absence or presence after infection. This located the *ea31* gene between *ea59* and *ea47* genes within **λ**b map (Figure 8). The **λ** *b* region proteins were classified into two groups based on time of expression after infection (Early and Late).

Early proteins, included Ea31 (a 31 kDa protein, as well as Ea59 (a 59 kDa protein) and Ea47 (a 47 kDa protein), which corresponded to three open reading frames (Reeve and Shaw, 1979).

It was found that the purified Ea59 protein acted specifically as an endonuclease for supercoiled DNA and as an ATPase, using cofactors such as a single stranded DNA, ATP, and Mg^{2+} ions. However, its main role in the **λ** life cycle has not yet been discovered (Chowdhury et al. 1972; Rhoades and Meselson 1973).

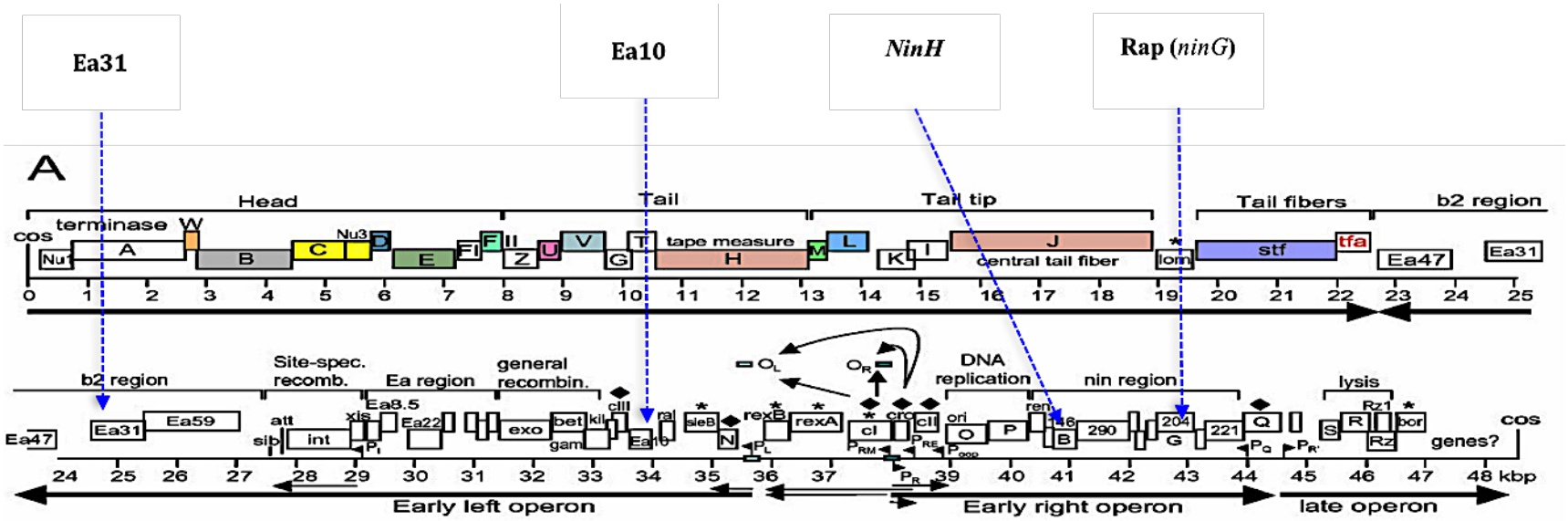


Figure 1.9: Genome map of Lambda phage.

It shows a Linear layout of lambda phage genome with major operons, promoter regions and capsid coding genes,

as well as the locations of Rap, Ea10, NinH and Ea10 genes (*Rajagopala, S. V.; Casjens, S.; Uetz, P. (2011).*

1.11 Aims of study

This project aims to crystallize and determine protein structures for Rap endonuclease, NinH, Ea10 and Ea31 proteins by using X-ray crystallography. A comparison of their structure to known models will indicate the possible mechanism of DNA binding, cleavage and bending of Rap and NinH proteins. Additionally, a similar analysis may reveal the possible function for Ea10 and Ea31 proteins.

2 Chapter 2. Materials and methods

2.1 Bacteria and plasmids

Many plasmids and competent cells were used throughout the whole study and from different sources table 2.1, 2.2.

Table 2.1: Plasmids and their sources used for cloning and expression experiments

1. Plasmids	2. Reference
3. pET28a	4. Invitrogen Cat # V430-01
5. pET21a	6. Invitrogen
7. pET24a	8. Invitrogen
9. pET22b	10. Invitrogen
11. pET16a	12. Invitrogen
13. PJoneX44	14. Prof. Jon Sayers' lab (Original manufacturer)

Table 2.2. The strains of competent cells used with their genotypes and sources

Organisms	Genomic type	Reference
DH5α™ Competent cell	F– Φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17 (rK–, mK+) phoA supE44 λ– thi-1 gyrA96 relA1	SLS Bioline cells cat # VX18265017
Silver efficiency competent cell	F' edeoR endA1 recA1 relA1 gyr96 hsdR17(rk , Mk) supE44 thi-1 phoA Δ(lacZYA argF)U169Ø80lacZΔM15 λ	SLS Bioline cells cat # BIO85026
BL21(DE3) pLysE competent cell	F– ompT gal dcm lon hsdS$_B$ (r_{B-} m_{B-}) λ (DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5])	SLS Bioline cells cat # BIO-85034
M72 competent cell	Δ(bio-uvrB) lacZ(Am) rspL Δ(trpEA2) (λ Nam7 Nam53 c1857 ΔH1)	
BL21-AI competent cell	F– ompT hsdSB (rB–, mB–) gal dcm araB::T7RNAP-tetA	Invitrogen™ cells cat # C607003

2.2 PCR primer design for cloning.

Primers were designed using either Snapgene software or manual calculation depending on the gene sequence obtained from http://genolist.pasteur.fr/Colibri/index.html with addition of sequences of restriction sites. The melting temperature, GC content and self-complementary features were calculated using oligo calculator (http://www.basic.northwestern.edu/biotools/oligocalc.html). The required primers were custom synthesized and provided by Eurofins (https://www.eurofinsgenomics.eu) (Table 2.3).

Table 2.3. Primers for Rap endonuclease protein used with different vectors (Ribosomal binding site (green), Restriction site (red), Start codon (blue), and Stop codon (brown))

Vector	Protein	Primer type	Primer sequence
pJonex44	P4	Forward (EcoRI)	aattGAATTCaggaggTATCTAatgAAGCCGCGTAGCTATTGG
		Forward (EcoRI+His)	aattGAATTCaggaggATACTAatgCATCATCATCATCATCATA AGCCGCGTAGCTATTGGATTAAGCAG
		Reverse (HindIII+ His)	aattAAGCTTttatcaGTGGTGGTGGTGGTGGTGTGCGGCTT CGCTACGGCTA
		Reverse (HindIII)	aattAAGCTTttatcaTGCGGCTTCGCTACG
	P3	Forward (EcoRI)	aattGAATTCaggaggTATCTAatgTGTAGCCCGGAATGTGGT
		Forward (EcoRI+His)	aattGAATTCaggaggATACTAatgCACCACCACCACCACCAC TGTAGCCCGGAATGTGGTACGAAG
		Reverse (HindIII+ His)	aattAAGCTTttatcaGTGGTGGTGGTGGTGGTGTGCGGCTT CGCTACGGCTATT
		Reverse (HindIII)	aattAAGCTTttatcaTGCGGCTTCGCTACGGCT
pET21a	P3	Forward (NdeI)	CGACATatgTGTAGCCCGGAATGTGGTACG
		Reverse (XhoI+ His)	CGACTCGAGTGCGGCTTCGCTACGGCTATT
		Reverse (XhoI)	CGACTCGAGtcaTGCGGCTTCGCTACG
	P4	Forward (NdeI)	CGACATatgAAGCCGCGTAGCTATTGGATT
		Reverse (XhoI+ His)	CGACTCGAGTGCGGCTTCGCTACGGCTATT
		Reverse (XhoI)	CGACTCGAGtcaTGCGGCTTCGCTACG

2.3 Agarose gel electrophoresis

All the DNA samples were analysed on 1% agarose gels. The mixture of agarose powder of 1% (w/v) in 1X Tae buffer (40 mM Tris-HCl pH 8, 20 mM acetic acid, 1 mM EDTA) was boiled to homogenize the solution-using the microwave, and then supplemented with 1 µg/ml Gel red (Sambrook et al. 1989). The solution was poured into a gel tank and the solid agarose gel then was covered with running buffer (1X TAE). DNA samples were mixed with 6X loading dye (Bioline). Hyper ladder (Bioline) was used as the molecular size marker. The gel was run at a constant voltage of 80 V for 60 minutes after loading samples and bands visualized under ultraviolet illumination.

2.4 PCR amplification of target genes from genomic DNA

PCR amplification was carried out to clone the target gene from its genomic DNA for incorporation into an expression vector (Going 2003). All the target genes were amplified in 25 µl volume in a 0.1 ml microfuge tube using a ^{3}Prime PCR machine. The genomic DNA, primers and Q5 polymerase kit 9 (Biolab) were kept on ice. PCR Mastermix solution was prepared by following the Q5 protocol in which 5 µl of Q5 reaction buffer, 0.5 µl dNTPs, 5 µl of High GC enhancer and 0.25 µl Q5 polymerase were combined and the final volume made up to 21.5 µl using autoclaved MilliQ water. The PCR samples were prepared by adding 1.25 µl of 5 pmol/µl concentration of each forward and reverse primers and 1 µl genomic DNA to the 21.5 µl master mix. The PCR was performed using the Q5 kit setting program (30 second Initial Denaturation 98 ˚C, and then 30-35 cycles starting with 98 ˚C for 5-10 sec, 10-30 sec at 50-72 ˚C, 20-30 sec at 72 ˚C, final extension 2 min at 72 ˚C and a final hold for 10 min at 4-10 ˚C). After the PCR step, 3 µl PCR sample was loaded on a 1% agarose gel at 80 V for 60 minutes to check the DNA yield. The PCR products were then purified using an ion exchange spin column (quick PCR purification kit from QIAGEN/UK as described in manufacturer manual). After cleaning up, the DNA products were also checked using an agarose gel to confirm the presence of the target products.

2.5 Plasmid DNA purification

Microgram quantities of plasmid DNA were generated for cloning use and protein overexpression. Transformed competent cells, which carry the plasmid of interest, were grown overnight on LB plate (10 gm Trypton (Oxoid), 5 gm Yeast extract (Oxoid), 10 gm NaCl (Fisher), 15 gm agar (Oxoid)/ 1Liter, supplemented with 100 µg/µl of appropriate antibiotics (Sigma)). A single colony was picked to inoculate 5ml LB broth. The overnight culture was then pelleted using centrifugation for 20 minutes at 4500 g (Sigma 3-16 K) and the pellet was stored at 20 ˚C. The frozen pellet was used to extract DNA plasmid using a QIAGEN plasmid mini prep kit as explained in manufacturers manual. The concentration of DNA plasmid was measured by using a Nano drop machine (IMPLEN). A 3 µl volume of extracted plasmid solution was loaded and checked on a 1% agarose gel.

2.6 Restriction digestion

DNA genes and vectors were digested using various restriction enzymes with appropriate buffers, which are manufactured by New England Biolab (NEB), following the user manual (1 µg of DNA of interest, 2 µl of reaction buffer, 1 µl of each restriction enzyme, completed to 20 µl by autoclaved MilliQ water). A single digest strategy was used to check the enzyme activity. The restriction enzymes target sequences were checked using the NEB cutter server (http://www.labtools.us/nebcutter-v2-0) to ensure that they did not cut somewhere else within the gene sequence (Table 2.4). The whole amount of DNA was loaded on agarose gel, and then extracted using a QIAGEN gel extraction kit (as mentioned in user's manual) was performed to purify the cut plasmid and to eliminate all the unwanted materials.

Table 2.4. Restriction enzymes used for gene restriction digestion

Enzymes	Source	Target Sequence 5'----3'
NdeI	Neisseria denitrificans	CA TATG
EcoRI	Escherichia coli	G AATTC
HindIII	Haemophilus influenzae	A AGCTT
XhoI	Xanthomonas holcicola	C TCGAG
BamHI	Bacillus amyloliquefaciens	G GATCC
NotI	Nocardia otitidis	GC GGCCGC
DraI	Deinococcus radiophilus	TTT AAA
EagI	Enterobacter agglomerans	C GGCCG

2.7 Ligation

The PCR product of the gene of interest and its vector were ligated using a T4 ligase kit (NEB manual protocol: 1 µl T4 ligase to 1 µg DNA) using a number of different ratios (2:1, 1:1, 1:2 and 1:3 plasmid to insert). All the ligation samples were kept on ice overnight for better activity of the T4 ligase. Two microliters of the ligated constructs were run on a 1% agarose gel against the cut and uncut plasmids to visualize the difference in size and to confirm the ligation results.

2.8 Transformation

The ligated DNA constructs were transformed into their appropriate competent cells such as DH5α (Table 2.2) for DNA recovery and expression. The transformation procedure was carried out by thawing 50 µl of the competent cells on ice, mixing them gently with 5 µl of a target DNA construct on ice for 30 minutes before exposing them to 40 seconds 42 ˚C. After the heat shock step, a further 2 minutes incubation on ice was used and then 1 ml of LB media was added to the samples and they were incubated for 60 minutes at 37 ˚C in a shaking incubator. The sample was centrifuged for 5 minutes at

4000 rpm and most of the supernatant was discarded. The pellet was re-suspended with the remaining supernatant, and then spread on LB plates. The plates were incubated overnight (Hanahan, 1983).

2.9 Colony screens

A colony PCR method was performed to confirm the ligation and presence of the target DNA. A single colony was picked from an overnight culture plate and suspended in 0.1 ml of sterile MilliQ water. A 30 µl volume of the suspended sample was heated for 10 minutes at 100 ˚C. A PCR reaction was set up using 5 µl of the sample solution, 1.25 µl volume of each appropriate primer, 12.5 µl polymerase Mastermix and 20 µl volume of autoclaved MilliQ water. This sample was run in a PCR machine using the same setting program from the initial DNA amplification stage. The PCR product was then examined to reveal the presence of the target DNA on a 1% agorse gel.

2.10 Sequencing investigation

The remaining 70 µl of the earlier 100 µl suspension was plated first on LB media and incubated overnight and then a single colony was used to inoculate 5 ml LB media. A glycerol stock of 1ml volume from this overnight culture was made and labelled for storage at -70 ˚C and the rest of the broth culture was used with a plasmid Miniprep kit protocol (QIAGEN) to extract the plasmids. The plasmid pellet was stored at -20 ˚C.

The extracted target construct was sent to GATC Biotech for sequencing, following the GATC guideline manual (5 µl of DNA (>100 ng/µl) mixed with 5 µl of appropriate forward or reverse primer) to analyse the gene sequence and ensure the absence of mutations. Following sequence analysis, plasmid samples were transformed into specific protein expression competent cell lines (BL21(DE3) pLysE, M72 or BL21-AI) using the same transformation method as mentioned in section 2.8.

2.11 Protein over expression

2.11.1 Small scale over expression

The transformed cells for protein expression were used in small scale cultures to check the capability of the cell to produce the target protein and to determine the optimum post induction condition, which yielded a high level of soluble protein. Transformed cells were plated first on LB agar and incubated overnight and then a single colony was used to inoculate 5 ml LB media. A 1 ml volume of overnight culture was used to inoculate 50 ml LB media in 250 volume flasks, which were used for better aeration condition, supplemented by an appropriate antibiotic at 100 μg/ml. The inoculated samples were incubated at 37 ˚C in a shaking incubator until log growth stage estimated from an OD_{600nm} of 0.6-0.8. A 1ml volume pellet of each sample was stored at -20 ˚C as a pre-induction sample. The cultures were then induced using various concentrations of appropriate inducers (0.1, 0.5, 1, and 2 mM) and under different temperatures (37, 30, 25, 18 and 13 ˚C). Pellets of 1 ml volume culture from different times of incubation after induction (3, 6, 18 and 48 hours) were stored at -20 ˚C to check the expression and the solubility on SDS PAGE later.

The transformed M72 cell line was cultured in 50 mL of LB, supplemented with 0.1 mM IPTG to repress the target protein and 100 μg/mL ampicillin, at 30 ˚C. Cells were grown to an OD_{600} = 0.6, then induced by increasing the temperature to 42 ˚C for 2-3 hours, and then returned back to 30 ˚C. 1 ml sample pellet was collected before induction and after 3, 6, and 18 hours of induction to check the expression and the solubility on SDS PAGE.

2.11.2 Large-scale overexpression.

The ideal expression conditions reported from small-scale overexpression assays were performed to scale up to large expression volumes of target protein. The target protein cells were plated overnight to obtain fresh single colonies, and then to inoculate 50 ml media. From the overnight culture (50 ml) a sample of 5 ml was used to inoculate a 500 ml volume of suitable culture media (LB, M9 minimal, 2xTY or 4xYT) in a 2000 ml flask, and then the optimized conditions were used to produce the proteins.

Finally, the cells were harvested by transferring the culture into 500ml centrifuge tubes, spinning down in an Avanti centrifuge (Beckman), using an F500 rotor at 8500 rpm for

30 minutes at 4 ˚C. The supernatant was discarded, and the pellet was re-suspended into 50 ml falcon tubes and centrifuged again at 5000 rpm for 15 minutes at 4 ˚C. Cell pellets were stored in -20 until needed.

2.12 Determination of protein expression by SDS-PAGE

2.12.1 Sample preparation

Stored pellets from 1 ml volume of culture were re-suspended in 300 µl of lysis buffer and then lysed on ice by sonication using pulses of approx. 15 kHz for 3×20 seconds in a Soniprep 150 at 3 micron amplitude. The suspension was centrifuged for 10 minutes at 45000 x g, the supernatant was transferred into 1.5 ml centrifuge tubes and 300 µl of water was added to re-suspend the debris.

2.12.2 Determination of protein concentration by Bradford assay.

The concentration of protein was analysed by using the Bradford method. A 1-10 µl volume of protein solution was added into a 1 ml cuvette and then diluted with a 0.8 ml MilliQ water and 0.2 ml Bradford reagent (Bio-Rad). The resulting absorbance change at 595 nm was used to measure the concentration of protein by applying the following formula: concentration of protein (mg/ml) = (OD_{595} x 15)/ volume of protein (µl) added to the cuvette) (Bradford, 1976).

2.12.3 SDS-PAGE preparation

SDS-Acrylamide gels were used to visualize expression and solubility of protein samples. The gel was prepared by following the standard laboratory protocol (Hames, 1998). A 12% acrylamide resolving gel was generally used in most cases except when the protein of interest had an expected molecular weight less than 10 kDa and a 15% acrylamide gel was used. A 12% acrylamide resolving gel contains 2.5 ml 30 % acrylamide, 2.35 ml 1M Tris-HCl pH8.8, 1.28 ml MilliQ water, 62.5 µl 10 % SDS, 62.5 µl 10 % ammonium persulphate (APS), and 6.25 µl tetramethylethylenediamine (TEMED). On top of the resolving gel, a 6% stacking gel was used which consisted of 0.75 ml 30% acrylamide, 0.47 ml 1M Tris-HCl pH6.8, 2.46 ml MilliQ water, 375 µl 10% SDS, 37.5 µl 10 % APS, and 3.75 µl TEMED. A 15 µg protein sample was mixed with 5 µl loading dye and 2 µl of 10x

reducing agent, and then incubated on a heating block at 100 °C for 5 minutes. Samples were then loaded on the SDS gel under running buffer, which contained 1 g/L SDS, 3 g/L Tris-HCl pH 8.8, 14.4 g/L glycine, and the gel was run at 80V for 15 minutes before increasing the power to 200V for 35 minutes.

2.12.4 Staining and de-staining for SDS gel

Gel was released from the gel plate and washed with distilled water. A typical SDS gel was stained for 2 hours either with instance blue stain solution (Expedeon/ protein discovery and analysis) followed by distilled water washing, or Coomassie blue stain solution (0.1% Coomassie blue, 50% Methanol, 10% Acetic acid). A Coomassie stained gel was de-stained overnight with de-stain solution, which contained 20% methanol and 10 % acetic acid (Hames and Rickwood, 1998).

2.13 Production of Selenium-Methionine (Sel/Met) protein

Sel/Met protein was produced by replacing the usual sulphur containing methionine residues within a protein structure with Seleno-Methionine during the protein expression process. A 1% starter overnight culture was used to inoculate 500 ml LB media at 37 ˚C in a 250 rpm shaking incubator until an OD_{600} of 0.6-0.8 was reached. This culture was then centrifuged at 3500 rpm for 20 minutes and the resultant cell pellet was re-suspended and washed once with M9 minimal media before being centrifuged again and then re-suspended with M9 minimal media. This was used to inoculate 6×500 ml of M9 media containing appropriate antibiotic. The M9 minimal media contained 10.5 g/L K_2HPO_4, 1 g/L $(NH_4)_2\,SO_4$, 4.5 g/L KH_2PO_4, 0.5 g/L Tri-sodium Citrate.$2H_2O$, and 5 g/L of glycerol, adenine, guanosine, thymine and uracil. The M9 media was also supplemented with 1 g/L $MgSO_4.7H_2O$, 40 mg/L Seleno-L–Methionine, 100 mg/L L-lysine, 100 mg/L L-phenylalanine, 100 mg/L L-threonine, 50 mg/L L-isoleucine, 50 mg/L L-leucine, and 50 mg/L L-valine. Using the same growth conditions for native cultures in LB media, the culture was induced after obtaining an OD_{600} of 0.6-0.8 with 0.5 mM final concentration of IPTG and 0.2% final concentration of L-arabinose at 37 ˚C, in a 250 rpm shaking incubator.

2.14 Protein Purification

A cell pellet was defrosted on ice and suspended in 5 volumes of lysis buffer, which contained 50 mM Tris-HCl pH8, and in the case of 6xHis-tagged protein an additional 0.5 M NaCl was required in the buffer. The suspended sample was lysed on ice by sonication (pulse of approx. 18kHz for 3×20 seconds with a Soniprep 150 at 18-micron amplitude). The soluble proteins in the sample were separated by centrifugation in a JA-20 Avanti centrifuge (Beckman), using a JA 20.50 rotor at 20000 rpm (60000 xg), for 20 minutes at 4 ˚C. Supernatant was applied immediately onto chromatography purification columns after measuring the protein concentration by the Braford method (Bradford, M., 1976). A number of different purification methods and columns were used in these studies based on the binding affinity of proteins to the columns.

2.14.1 DEAE FF column

Supernatant was loaded onto a DEAE Hi-Trap Ion Exchange Fast-Flow 5 ml column (GE healthcare), which was pre-equilibrated in buffer A (50 mM Tris-HCl pH8) at 4 ml/min. Protein was eluted with a gradient of buffer B (buffer A plus 0.5 M NaCl) in 100 ml. The concentration of alternate 2.5 ml collected fractions was checked using the Bradford reagent (Bradford, M., 1976) and the peak fractions were run on an SDS gel to check the presence and the purity of the target protein. Peak fractions containing the highest protein concentrations were combined and concentrated before the next purification step.

2.14.2 Resource Q column

A 6 ml Resource Q column (GE healthcare) was required for the Ea10 purification due to a high level of contaminants seen after the DEAE purification column. The purified protein from the DEAE column was diluted 2-fold with buffer B (50 mM Tris-HCl pH8 + 1 M NaCl). The sample was loaded onto a Resource Q column, which was pre-equilibrated in MES buffer pH 6.5. The protein was then eluted with a gradient of 0-1 M NaCl in 60 ml, at a flow rate of 5 ml/min collecting 2.5 ml fractions. Peak fractions were analysed on SDS-PAGE gel to check the presence and the purity of the target protein.

2.14.3 Q-HP column method

Alternatively, a new purification strategy for purifying Ea10 protein was applied. Immediately after lysis and centrifugation, the cell free extract from the Ea10 protein culture was applied onto a 5ml Q-HP column (GE healthcare). The Q-HP column was pre-equilibrated to 50 mM Tris-HCl pH8 at 2.5 ml/min. The protein was eluted from a gradient of 0-1 M NaCl in 100 ml in the Tris buffer. The concentration of alternate 2.5 ml fractions was checked using Bradford assay (Bradford, M., 1976), and the peak fractions were checked on an SDS-PAGE gel.

2.14.4 His-trap column Method

For samples with His-tagged target proteins, the CFE was loaded onto a 5 ml Ni-His Trap column (GE healthcare). The column was pre-equilibrated to 50 mM Tris-HCl pH8 at 2.5 ml/min and 0.5 M NaCl. The protein was eluted from a gradient of 0-0.5 M NaCl and 0-0.5 M Imidazole over 100 ml in 50 mM Tris-HCl buffer pH 8. The concentration of alternate 2.5 ml fractions was checked using Bradford assay (Bradford, 1976) and the peak fractions were checked on an SDS-PAGE gel.

2.14.5 Heparin column method

The CFE was applied onto a 5ml Hi-Trap Heparin HP column (GE healthcare), which has proven affinity for DNA-binding proteins, coagulation factors, lipoproteins, and protein synthesis factors. The column was pre-equilibrated in buffer A (50 mM Tris-HCl pH8) at a flow rate of 2.5 ml/min. After washing the column with 20 ml of buffer A, protein was eluted from a gradient of 0-0.5 M NaCl in buffer A over 100 ml. A gradient over 75 ml from 0.1 to 1.0 M NaCl was used for the Rap protein. The concentration of alternate 2.5 ml fractions was checked using the Bradford method (Bradford, M., 1976) and the peak fractions were checked on an SDS-PAGE gel.

2.14.6 Gel filtration.

Following either ion exchange, Ni-His Trap or heparin columns, the best fractions as assessed by SDS-PAGE were merged and concentrated to a volume of 1 to 2 ml using Vivaspin concentrator at 4000 ×g. The protein was then applied onto a Superdex^{TM} 200 gel filtration column pre-equilibrated with 50 mM Tris-HCl pH 8 + 0.5 M NaCl at a flow rate of 2.5 ml/min. The peak fractions were checked on an SDS-PAGE gel.

2.15 Nuclear Magnetic Resonance test (NMR)

The purified protein samples were diluted to contain lower salt in a buffer of 10 mM Tris-HCl pH 8.0, 50 mM NaCl, and 10% D_2O (1 ml at 0.5 mM concentration of protein), and then sent for analysis by NMR (**nuclear magnetic resonance centre / MBB department / University of Sheffield**) to study the conformational nature of the protein sample.

2.16 DTNB assay for Cysteine accessibility investigation

The purpose of this assay was to determine the presence and accessibility of cysteine residues in the proteins for heavy atom derivatization, based on the interaction between any available cysteine sidechain sulfhydryl group and DTNB (5,5'-dithio-bis-[2-nitrobenzoic acid] also known as Ellmans reagent) . The compound was detected by its absorbance at 412 nm (Ellman, 1959).

The assay was performed using either native buffer A (0.1 M sodium phosphate, 1 mM EDTA pH 7.3), or denaturing buffer B (6 M GuHCL, 0.1 M Sodium phosphate, 1 mM EDTA pH 7.3) along with DTNB reagent (10 mM in buffer A). A sample of 0.1 ml of DTNB reagent was added to 0.9 ml of buffer (A, or B) and used to set the baseline at 412 nm. To the cuvette, 50 µl of protein solution (5-50 mg/ml concentration) was added and mixed by inversion, then incubated for 1-2 minutes at room temperature. An absorbance measurement was taken at 412 nm to calculate the amount of the DTNB reagent that had reacted with protein, using the extinction coefficient of DTNB reagent (14150 M^{-1} cm^{-1}).

Concentration of TNB in µM / concentration of protein in µM = amount of reduced sulfhydryl group in the protein.

The number of readily available cysteines in the protein is calculated from the free sulfhydryl groups in the native buffer A, whereas the total number of cysteines not forming disulphide bridges is measured by the number of free sulfhydryl groups in the denaturing buffer B.

2.17 Protein Crystallization.

2.17.1 Initial crystallization trials

The highly pure protein at a range of different concentrations > 6 mg/ml was tested for crystallization using a wide range of commercial screening conditions including the JCSG, MPD, PACT, Morpheus, Classic and Ammonium sulphate screens (Molecular Dimensions), via the sitting drop vapour diffusion method using 96 well MRC2 crystallization plates on either a Hydra II or Mosquito robot. The crystallization plates were incubated at either 17 ˚C or 7 ˚C. For the first week, trays were checked daily for crystal formation and weekly thereafter.

2.17.2 Crystal optimization

Optimization of conditions that produced protein crystals was carried out by varying buffer pH, concentration of salt, and precipitant concentration. The hanging drop vapour diffusion technique was employed, which included mixing 1:1, 1:2 and 2:1 ratio of the purified protein with crystallization solution on a siliconized coverslip before sealing the coverslip with paraffin oil over a reservoir of the crystallization solution. Plates were then incubated at the same temperature as the initial crystallizations. Optimization trays were checked daily for crystal formation during the first week and weekly thereafter.

2.17.3 Siliconizing coverslips

Hanging drop optimisations were set out using siliconized coverslips, which have a hydrophobic surface for the formation of drops. A volume of 5 ml dimethyldichlorosilane solution (siliconizing solution) was poured in a small beaker. A large glass petri dish containing 50-100 coverslips was placed above the beaker. All were placed in a large bell jar connected to a water pump for 5 minutes to form a vacuum in the bell jar before closing the inlet valve. The vacuum was maintained for a further 30 minutes. Siliconized coverslips were polished using lens tissues and dusted with compressed air to eliminate debris particles on the glass surface. Polished coverslips were then stored in petri dish until used.

2.17.4 Cryoprotection solutions

Crystals were cryo-protected before being sent for X-ray analysis. Cryoprotection solutions contained 10-30% $^{v}/_{v}$ ethylene glycol or low molecular weight PEGs. A cryo-protectant of 10 – 30% $^{v}/_{v}$ glycerol was used in conditions not containing PEGs.

2.17.5 Mounting crystals

For data collection, crystals were taken from the crystallization solution drop, and mounted into appropriate loops. Loops varied in size and were selected so that the size of the loop matched the size of the crystal. The clear film covering the well was removed to expose the crystal drop under a light microscope. A small volume of cryo-protection solution was placed in the adjacent small well and the loop was used to extract a single crystal and transfer it immediately into the cryo-protection solution. The crystal in the loop was removed again to place it quickly into either liquid nitrogen directly or a nitrogen gas stream at 100 K. Crystals in conditions which contained a high percentage (> 25 % $^{w}/_{v}$ or $^{v}/_{v}$) of either PEG or MPD were looped and stored directly into liquid nitrogen.

2.17.6 Data collection and processing

Crystals were shipped in liquid nitrogen for collecting diffraction datasets at the Diamond Synchrotron, UK. The initial test images were carried out using 4 images at 45° intervals with 0.10˚ oscillation per image, and 0.1 sec exposure and typically 40% Transmission to obtain a first estimate of the crystal parameters and a data collection approach using Mosflm or EDNA software components. The strategy for collecting data usually involved collecting 1800 or 3600 images with a 0.10° oscillation per image. The data in this thesis were processed automatically at the Diamond synchrotron to determine a space group, resolution and cell dimensions. Three separate Xia2 pipelines based on the XDS, Xsacle and Aimless programs and one FastDP pipeline were used for indexing, integrating, scaling, and merging the data.

2.17.7 Merging scaled reflections

In order to obtain an overall value of averaged intensity for each reflection from the crystal lattice, the indexed and scaled, equivalent reflections were merged following the rules of the crystal symmetry. A quality determination of the merging process was also

performed by using the calculations of the agreement in intensities of measurements for the high-resolution bin, which helps to determine a suitable high-resolution cut-off of the dataset, for the low-resolution bin, which should be higher quality than the high-resolution dataset, and for all of the data, which can be useful for determining if a too high symmetry space group has been assigned. The most common quality indicator for merging is the linear merging R_{merge}:

$$R_{merge} = \frac{\sum_h \sum_{i=1}^{N} |I_{(h)i} - \bar{I}_{(h)}|}{\sum_h \sum_{i=1}^{N} I_{(h)i}}$$

Where h is each reflection hkl, N is the number of reflections merged, I is the intensity of the specific reflection and $\bar{I}$ is the averaged intensity of each reflection hkl. The R_{merge} assesses the agreement between the averaged intensity measurements of a set of symmetry related reflections and the individual intensities of each reflection by measuring the spread of independent measurements of the intensity of the reflection. It was found that the R_{merge} value is possibly not reliable because of the effect of collecting large numbers of reflections giving an apparent but incorrect increase in its value. Therefore, the R_{merge} measurement must be adjusted using a specific factor of N/(N-1) to give values more independent of multiplicity. The R_{meas} value provides a more reliable consistency of the individual metrics (Diederichs and Karplus, 1997).

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{j=1}^{n} |I_{hkl,j} - \langle I_{hkl} \rangle|}{\sum_{hkl} \sum_j I_{hkl,j}}$$

The linear R-value will increase as the multiplicity increases, as the redundancy of the data is not taken into account. A precision merging R_{value}, R_{pim} is preferred in these cases, given by:

$$R_{p.i.m.} = \frac{\sum_{hkl} \sqrt{\frac{1}{n-1}} \sum_{j=1}^{n} \left| I_{hkl,j} - \langle I_{hkl} \rangle \right|}{\sum_{hkl} \sum_{j} I_{hkl,j}}$$

Where the 1/(N-1) term accounts for many cases of the same reflection being measured. For that reason, R_{pim} is much lower usually than Rmerge, and can often be lower than 0.1 for the low-resolution bin. R_{pim} can also be useful in determining the amount of anomalous signal in a dataset by comparing the value with R_{anom}, where R_{anom} is analogous to signal, and R_{pim} is analogous to noise, and a value of R_{anom} /R_{pim} = 1.5 indicates a substructure solution is likely achievable (Rupp, 2010).

The quality of the data was also assessed by considering the overall completeness of the data, the multiplicity of measurements and the mean ratio of the intensity divided by its standard deviation (Mean I/σI) in resolution bins from low to high resolution. The combination of these metrics allowed decisions to be made about data quality and resolution limits.

2.17.8 Structure determination

Initial map and model were determined by experimental phase estimates using multiple SHELX jobs (C, D and E) to analyze early phases with minimal user input by altering the number of heavy atom sites, solvent content and or space group enantiomorphs. Alternatively, the phases were estimated by molecular replacement with homologous models using the PHASER program from the CCP4 suite. Repeated rounds of model building and refinement were carried out using the COOT and REFMAC programs, respectively, from the CCP4 suite (Emsley et al. 2010; Murshudov et al. 1997). Any other programs used in the model building process are mentioned throughout this thesis.

2.17.9 Structure validation

The Molprobity server was used to validate target structures after all the structure features had been accounted for in the electron density map and the rebuilding and refinement process had finished. Model coordinates were uploaded to the server and a validation report generated, which analysed bond angles, bond lengths, Ramachandran

angles, and sidechain rotamers (Chen et al., 2010). Hydrogen atoms were included in the model for this process and sidechains were flipped and rotated to produce appropriate hydrogen bonding in the COOT program. Models were then refined in REFMAC program to improve the Molprobity scores and the model compared with all other models, and particularly those of equivalent data resolution, in the protein database.

2.18 Heavy atom derivatization

For the Ea10 protein, crystals were treated with different heavy atoms with an aim to produce a scattering source to generate initial estimates of the phasing information from a simpler heavy atom substructure.

2.18.1 Iodine crystal soak

The idea of iodination in the crystal is to solve the protein structure using the phasing information from the effect on X-ray scattering of the interaction between iodine atom and tyrosine residues of the protein. Iodine crystals were prepared by dissolving 1 g of potassium iodide in 1 ml distilled water. 1 ml of concentrated HCl was added to this iodine solution and mixed thoroughly, and then 20 ml of 3% Hydrogen peroxide was added to the solution to precipitate the iodine crystals to the bottom of the flask. The solution was poured through filter paper and the iodine crystals were collected with further washing and allowed to dry on the filter paper (https://sciencing.com/extract-iodine-potassium-iodide-5640020.html). A small iodine crystal was introduced into grease next to a drop containing the protein crystal drop on a coverslip within the hanging drop method. The crystallization well was then re-sealed and incubated overnight at 14 °C to enable the iodine to sublimate and enter the crystallization drop (Figure 2.1).

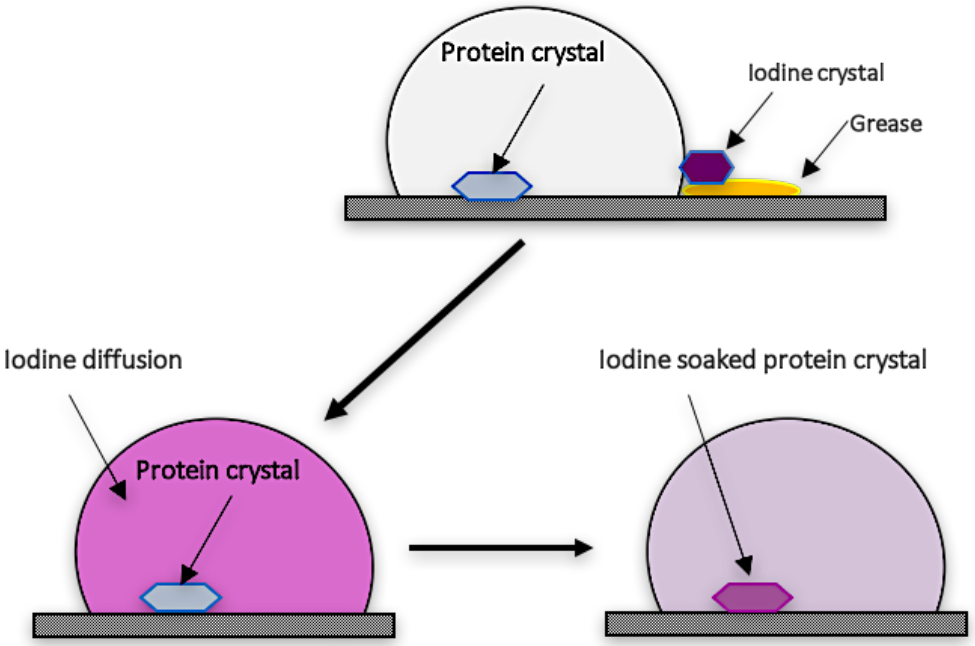


Figure 2.1. Process of Iodine crystal soak.

Shows absorption cycle of Iodine atom into protein crystal during 24 hour

2.18.2 Tantalum bromide cluster soak

A tantalum bromide compound (Ta_6Br_{12}) was used as a heavy atom derivative in order to obtain a strong anomalous scatter based on its electron–rich structure The method was performed using a micro-needle to add a microgram amount of Tantalum bromide powder into the crystallization drop containing protein crystals (Figure 2.2). The crystals turned green as an indicator of soaking with the tantalum bromide compound (Dahms et al., 2013).

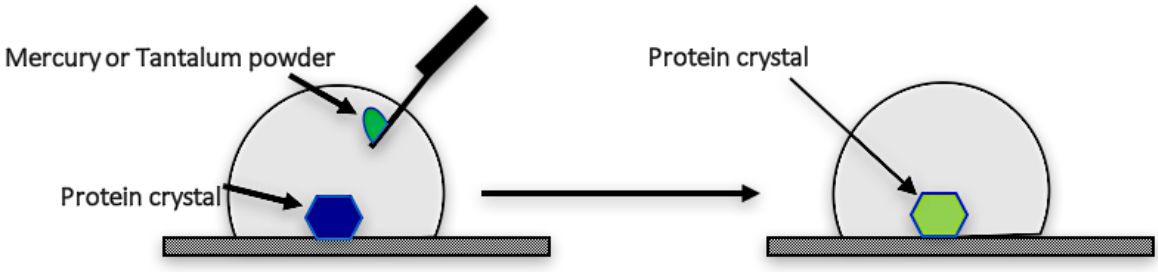


Figure 2.2. Process of Tantalum crystals soak

Shows absorption stages of Tantalum bromide into protein crystal during 24 hour incubation

2.18.3 Mercury derivatization

Ethyl-mercury Phosphate compound (EMP) was used as a heavy atom derivative because of the high affinity of mercury compounds for interaction with cysteine. The same technique was performed to add EMP powder into the crystallization drop as was used for tantalum bromide. In addition, a buffered EMP solution was added to the protein solution in a 2:1 molar ratio of EMP to protein prior for setting up crystallization trials in an effort to produce heavy atom derivatization for phase determination and possibly also to influence and improve general crystallization (Ramagopal et al., 2005).

2.19 DNA binding assay:

DNA binding assays were performed to investigate the ability of the Ea10 protein for DNA interaction based on suggestions from fluorescent anisotropy data (unpublished data, Dr. Gary Sharples, University of Durham). Assays were carried out using varying types and sizes of DNA substrates (Table 2.5).

Table 2.5. DNA substrates used in binding assays with Ea10 protein

DNA sequence	Number of nucleotides
CACCAA	6
GCTAGC	6
AGCTTGCTAAAGTATCT	17
AGCATGTCCTAGCAAGC	17
CAGATACTTATGACGTA	17
CTACGTCATGGACATGC	17
CAGATACTTGGACATGC	17
TTTGGTCTAACTTTACCGCTACTAAATGCC GCGGATTGGTTTCGCTGAATCAGGTTATTA	60
TAATAACCTGATTCAGCGAAACCAATCCGC GGCATTTAGTAGCGGTAAAGTTAGACCAAA	60

2.19.1 Thermo-shift assay

A protein Thermal Shift assay was used to measure the change in protein stability using a fluorescent protein-binding dye SYPRO Orange (Invitrogen, Carlsbad, CA) with the addition of different DNA substrates. A sample of 1 μg of purified protein per assay on a 96 well plate was used in high-throughput fashion. The samples were gradually heated from 25°C to 95°C while changes in fluorescence at 570 nm were monitored with a real-time PCR machine (Mx3005p from Stratagene, La Jolla, CA) following excitation at 470 nm. Protein samples, at varying pH of Tris-HCl buffer containing NaCl with or without DNA substrates (at various concentrations) in a reaction volume of 50 μl, were transferred into a 96-well PCR plate (semi skirted, STARLAB International GMbH, Hamburg Germany) in the RT-PCR device. Each protein was first examined without DNA to establish production of a strong signal at the lowest possible concentration of protein. The samples were heated at 1 °C per 45 sec, from 25 °C to 95 °C. The fluorescence intensity was measured at regular intervals of 1 °C (Vedadi et al., 2006). The resulting data were analyzed using specially developed spreadsheets in Excel and GraphPad Prism programs (Niesen et al., 2007).

2.19.2 Electrophoretic mobility shift assay (EMSA)

In order to test protein-DNA binding by retardation of DNA samples, a native-acrylamide gel was prepared following the same strategy as in section 2.12.3 but without addition of SDS. It contained a 12% resolving gel and a 6% stacking gel. A 1:2 molar ratio mixture of the protein (15 μg) to DNA substrate was loaded on the native gel with 5 μl of Bromothymol blue dye and control samples (Schägger et al., 1994). The prepared samples were loaded under running buffer, which contained 0.02% Coomassie stain (without SDS), 3 g/L Tris-HCl pH 8.8, 14.4 g/L glycine, and run initially at 80 V for 15 minutes, then at 200V for 35 minutes (Novex 2012). The gel was stained either with instant blue stain solution, or Coomassie blue stain solution (0.1% Coomassie blue, 50% Methanol, 10% Acetic acid) (Hames and Rickwood, 1998).

2.19.3 Tryptophan fluorescence assay

This assay was carried out in order to determine the capability of protein to bind to DNA, in order to assess the interference between the tryptophan and DNA substrate.

The fluorescence peaks were determined for 2 ml of the protein at 2 µM and 1 µM of DNA ligand in a 2:1 molar ratio in a 50 mM Tris pH8 buffer using a quartz cuvette in a fluorescence spectrophotometer (Varian/ Cary Eclipse). Fluorescence intensity at 340 nm (20 nm slit width) was measured following excitation at 280 nm (5 nm slit width). This data was analysed using Graphpad prism program.

3 Chapter 3. Rap endonuclease, NinH and Ea31 results

Summary

No previous crystallographic observations have been made of the Rap endonuclease, NinH and Ea31 structures. This chapter describes the protein purification and crystallization for these three proteins, as well as a re-cloning strategy for two short versions of Rap Protein.

Full length Rap protein was readily expressed and purified, but no crystals were formed after many crystallization trials. Two alternative versions of Rap protein were cloned in different vectors and cell lines by removing either the N-terminal 30 or 70 residues in an effort to find a form of protein, which might be crystallized. Similar over-expression and purification processes were used on these versions. The longer truncated version was degraded quickly. However, crystallization trials were set down for both versions. No crystals had formed at the time of writing this thesis, despite many attempts to improve the crystallization.

NinH and Ea31 proteins were expressed with a His- tag or an MBP-tag, respectively, because of solubility problems. Purified samples of both these proteins were set up in many crystallization trials. Crystals were not observed. Tag cleavage for these proteins was also carried out to increase chances of crystallization. However, it was not possible to purify them to a suitable level of purity to set up crystallization trials. Therefore, these experiments were suspended because of time constraints.

3.1 Rap endonuclease results

3.1.1 Full length Rap endonuclease protein

3.1.1.1 Prediction of Rap endonuclease structure

The initial stage of the investigation of the Rap endonuclease structure was to find Rap homologues and a predicted structure of the protein. The result was a predicted model from the Phyre2 server. A model corresponding to residues 86-141 (28% of total sequence) of the Rap protein was created with 90.8% confidence using gve2 hnh2 endonuclease from the deep-sea thermophilic bacteriophage Geobacillus virus E2 (pdb entry: 5H0O) as the template. The structure of Rap was built as two fragments by Phyre2 that showed a distinct helix at the N-terminal end of the first fragment followed with two anti-parallel beta strands in the second fragment (Figure 3.1).

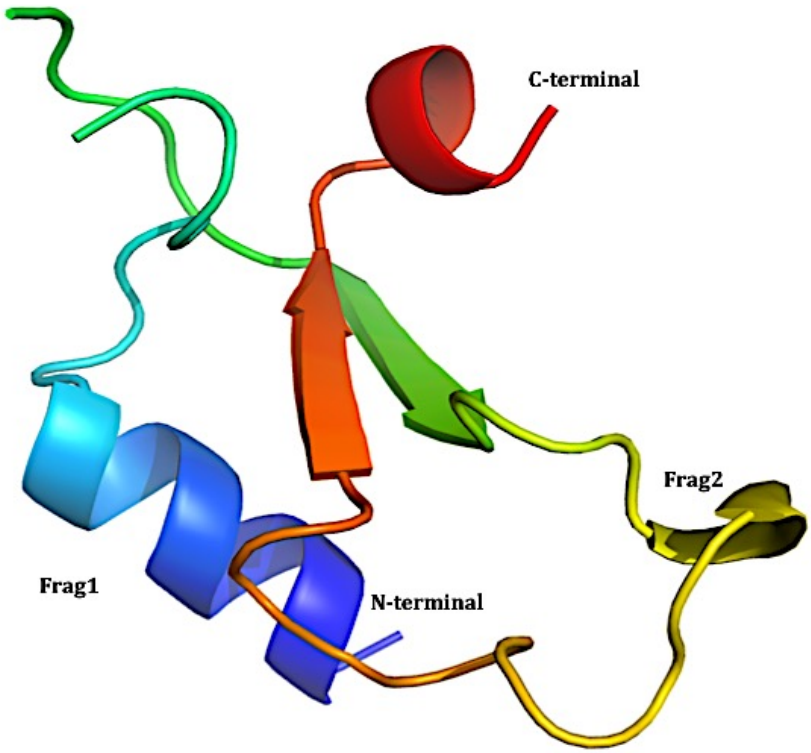


Figure 3.1. Phyre2 prediction of Rap endonuclease structure

The best predicted structure of Rap endonuclease from the Phyre2 server is shown in cartoon form coloured by rainbow format N- to C-terminus. The model is based on the gve2 hnh2 endonuclease (pdb entry 5H0O) but gives only a very limited coverage of the overall Rap sequence (28%).

3.1.1.2 Rap cloning background

The Rap endonuclease gene was cloned in the laboratory of Dr. Gary Sharples (University of Durham) into a pET22b vector using NdeI and XhoI restriction sites (Figure 3.2). The pET22b vector has a powerful system for cloning and expression of target genes controlled by a strong bacteriophage T7 system of transcription and under *Lac* repressor control and contains an ampicillin resistance marker. The recombinant plasmid was transformed by the heat shock method into *E. coli* competent cells (BL21 AI) for protein expression, which contain a chromosomal insertion of the gene encoding T7 RNA polymerase (T7RNAP) into the *ara*B locus of the *ara*BAD operon, thus the T7RNAP expression is regulated by the arabinose-inducible *ara*BAD promoter. This property makes the BL21 AI cell very useful for expression of genes that might be toxic to the other BL21 strains (Laitinen et al., 2005). Additionally, the *Lac*UV5 and *Lac*I genes are present from the bacteriophage DE3 system lysogenized into the genomic DNA of the host cell. The expression was induced by the addition of IPTG (Isopropyl-ß-D-thiogalacopyranoside) to a growing culture for inducing the LacUV5 promoter to direct transcription of the T7 RNA polymerase, which in turn transcribes the target gene in the plasmid (Studier and Moffatt, 1986). The Heat Shock method was used to transform the construct into BL21 AI competent cells.

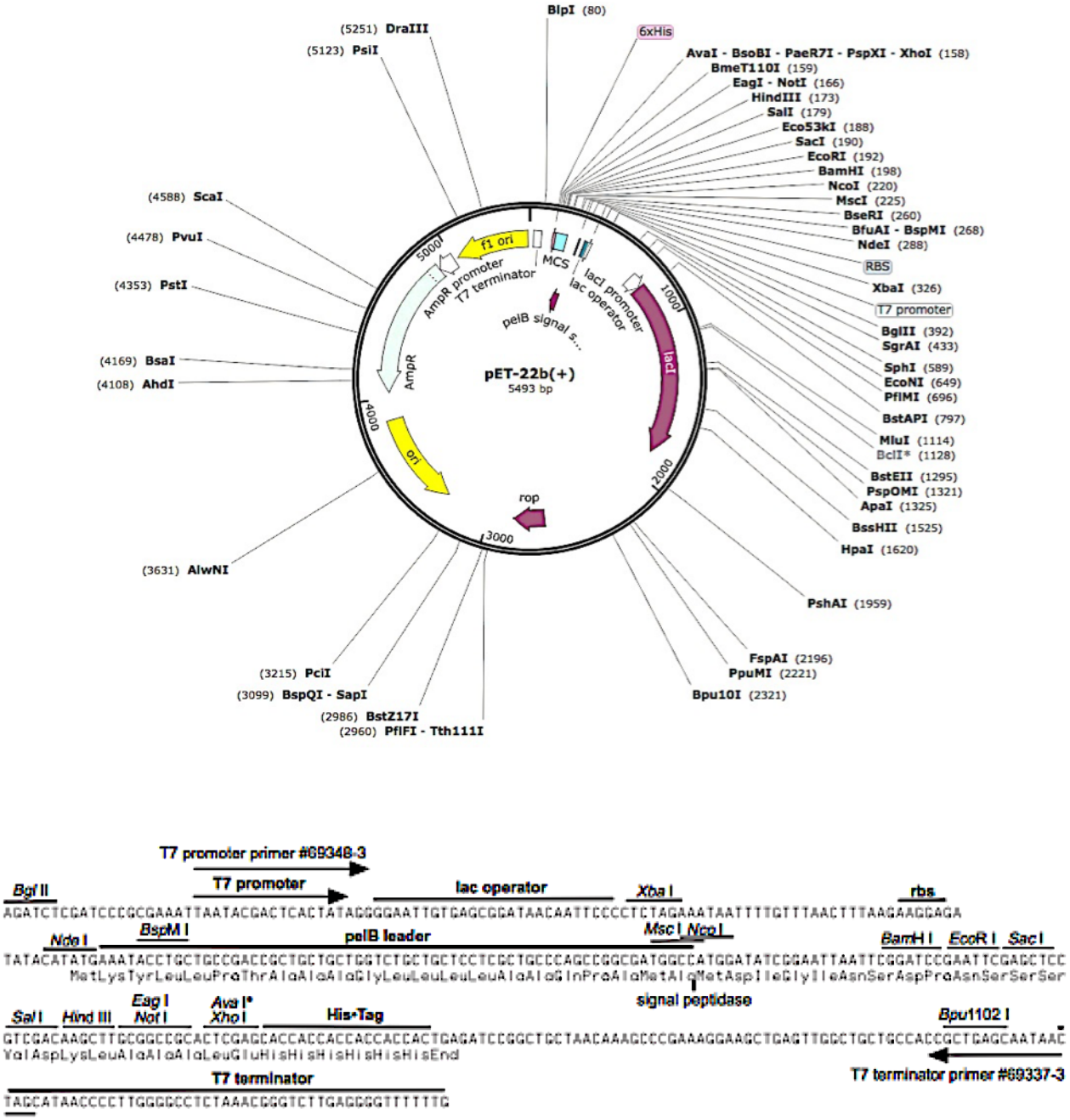


Figure 3.2. pET-22b vector map

Rap gene was cloned into a pET22b vector containing an Ampicillin resistance marker and Lac Operon system using NdeI and XhoI restriction sites. Image from Snap gene.

3.1.1.3 Colony screening

The BL21 AI cells containing the Rap constructs were plated out on LB agar plates supplemented with 100 µg/mL ampicillin, to obtain single colonies. Six colonies were selected to perform sequencing to make sure no mutation had occurred within the gene sequence. The plasmids were extracted from the colonies using the Qiagen Plasmid Miniprep Kit. Sequencing of the construct was performed by the GATC sequencing company (Section 2.10). Clear sequencing reads were obtained for four colonies. Sequence alignment with the standard λ phage *rap* sequence in NCBI GenBank came up with 100% identity using Blastn and Blastp.

3.1.1.4 Protein Expression

3.1.1.4.1 Small Scale Expression and Solubility

One of the four colonies containing the correct sequence of *rap* was selected for small-scale expression studies. Induction of the Rap protein from its pET22b vector was carried out (section 2.11.1). Varying concentrations of IPTG (0.1, 0.5, and 1 mM) and temperatures (16°, 25°, 30°, and 37 °C) were tested to compare their levels of Rap expression. 1 mL samples were collected before induction, and 3 hours, 6 hours, and 18 hours after induction, and analysed on a 12% acrylamide SDS gel to check the expression and the solubility.

The small-scale expression of Rap protein was optimized at 37 °C, using 0.5 mM IPTG and 0.2% L-arabinose, and at 18 hours after induction. A large band was found on the gel after induction, between the 21.5 kDa and 31 kDa markers, likely corresponding to the 24 kDa size of Rap endonuclease. This band was not present in the pre-induction sample, revealing a high level of expression of Rap. The majority of this protein was also found in the supernatant of the cell sample, indicating a high level of solubility.

3.1.1.4.2 Large Scale expression

Large-scale overexpression of the Rap protein was carried out using the optimised conditions for small-scale expression, using 6x 500 mL of LB media. Samples were collected before induction and 18 hrs after induction and analysed on SDS-PAGE (Figure 3). All cells were harvested 18 hrs after induction (section 2.11.2). 7 g of cell pellet was collected from 3 litres of growth media.

A similar level of expression of Rap protein can be seen for large scale overexpression compared with small scale expression. The high level of solubility suggested that it would be feasible to do purification (Figure 3.3).

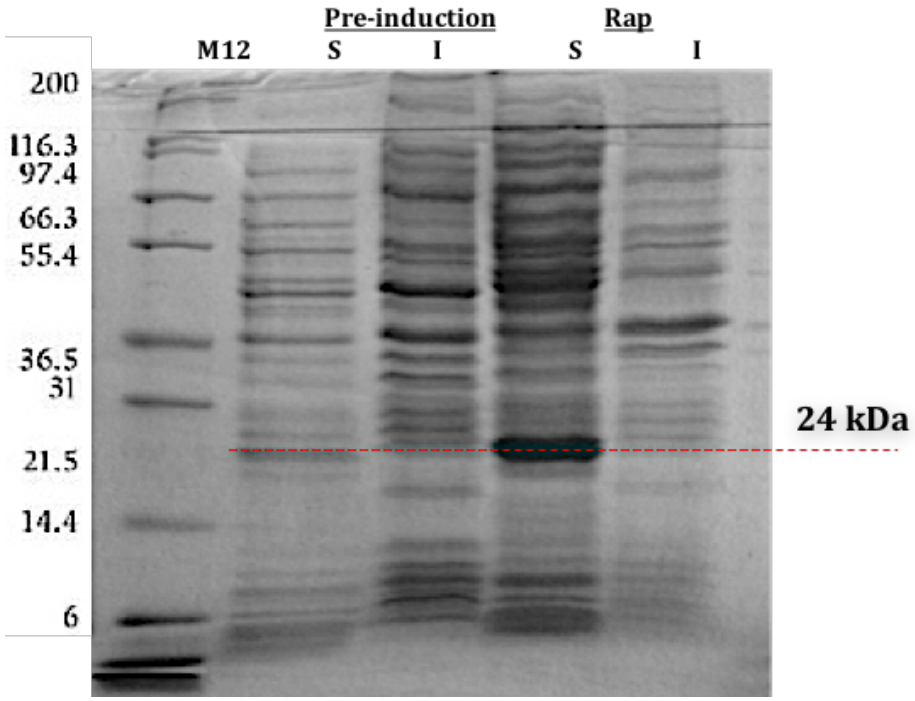


Figure 3.3. Overexpression and solubility.

A 12% SDS-PAGE gel showed a high level of soluble expressed Rap protein, which has an expected MW. around 24 kDa, as assessed by a Mark12 ladder. 15 µg was the amount of total protein in the sample of supernatant used for the SDS gel and judged by a Bradford assay. The protein was expressed at 37 °C, 0.2% L-Arabinose, and 18 hours incubation. The cells were induced with 0.5mM IPTG after OD600 0.6. S= soluble, I= Insoluble.

3.1.1.5 Protein purification

The Expasy protparam server (http://web.expasy.org/protparam/) results on the Rap protein sequence showed that the exact molecular weight of 203 amino acids was 23984.3 Dalton, and the theoretical Isoelectric point (PI) is 9.76, which makes the Rap protein more likely to be a positively charged protein in pH 8 buffer. The predicted Extension Coefficient was 38960 $M^{-1}cm^{-1}$.

The cell free extract was loaded on a 5 ml Heparin Sepharose Fastflow column (GE healthcare) (Section 2.14.5). The Rap protein fractions were confirmed on SDS-PAGE. The results showed the presence of many bands below the expected molecular weight of Rap protein (24 kDa) that might be due to a degradation of the Rap protein or some protein contaminants also bound to the column (Figure 3.4). The ones containing the Rap protein were pooled and concentrated down to 1 ml for gel filtration purification.

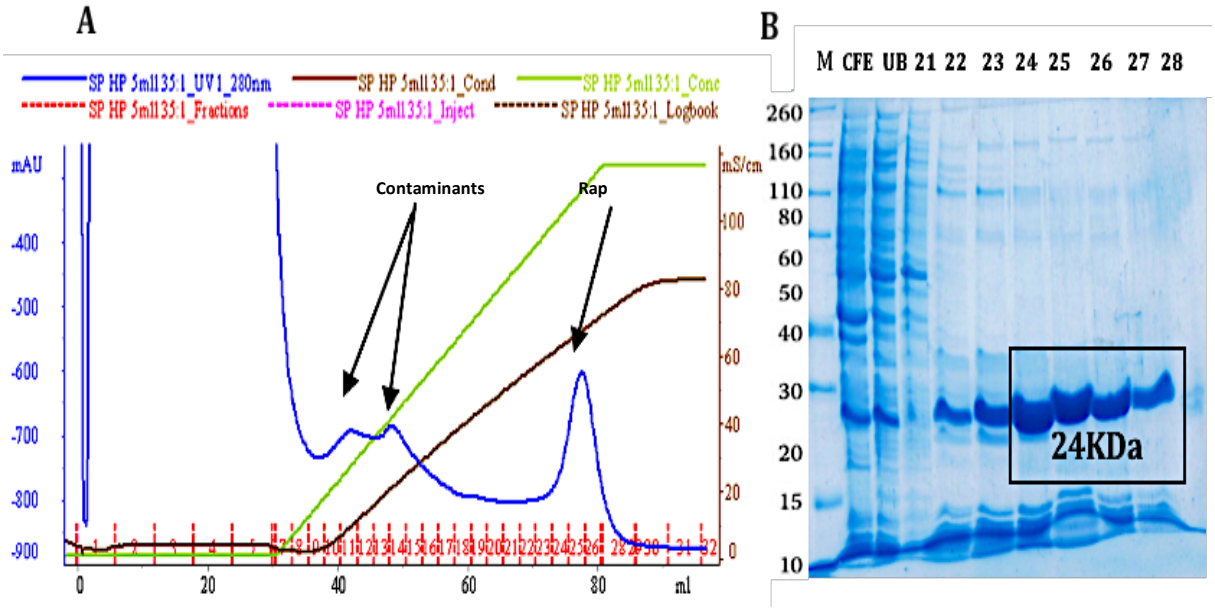


Figure 3.4. Heparin purification for Rap endonuclease protein sample.

A) Chromatogram of the Rap protein eluted from a 5 ml Heparin-HP column. The blue line represents the UV absorption at 280 nm, green line represents the salt ingredient, and the brown line refers to the conductivity. The results showed that the Rap protein eluted in fractions 24-27, which were pooled and used for the next purification step. B) The SDS gel shows the Rap purified protein after elution from a Heparin-HP column. Rap protein (24 KDa) appears in the lanes 7-10. 15 µg of protein was loaded in each lane. M= Mark 12, CFE= cell free extract, UB= unbound proteins.

The pooled fractions were then applied to a 1.6x60 cm Superdex^{TM} 200 gel filtration column (GE healthcare) equilibrated with 50 mM Tris-HCl pH 8.0, 0.5 M NaCl at a flow rate of 1 ml/min. 2 ml fractions were collected and the protein concentration for each determined using the BioRad protein assay. The purification was confirmed using SDS-PAGE with samples from the purification step (Figure 3.5). The result suggests that the Rap protein may be forming a dimeric assembly based on the chromatogram calculations for the main peak, using the following equation: Kav = (Ve-41)/75, where Ve= the eluted volume, 41= Vo= void volume, 75= V_t= total column volume, and Kav represents the elution volume function of a molecule for the column. The Ve in this case for the highest peak on the chromatogram is 78.31, giving a Kav of approx. 0.5. From the Kav, the molecular weight of the protein was then estimated by comparison to protein standards plotted on a logarithmic scale. A molecular weight of approx. 40 kDa was calculated for the peak, close to the expected 48 kDa size of a Rap dimer. Additionally, there was also a small peak following the major peak, which was estimated to have a molecular weight of around 20 kDa, indicating that this peak likely corresponds to the monomeric form of Rap, which has an actual size of 24 kDa. It is possible that the monomer of Rap underwent degradation that led to cleavage of around 4 kDa from the protein, thus a dimer formed from two cleaved Rap proteins would lead to a molecular weight of around 40 kDa. However, the calculations are estimations and can be affected by the shape and properties of a protein such as affinity for the column matrix material.

The fractions with more than 20% of the maximum protein concentration were pooled (Figure 3.5). The protein concentration of the resulting solution was determined, and the sample was used for setting crystallization trials. The remaining sample was stored at 4^{o}C (up to 1 week) or -70°C (for long-term storage) until needed.

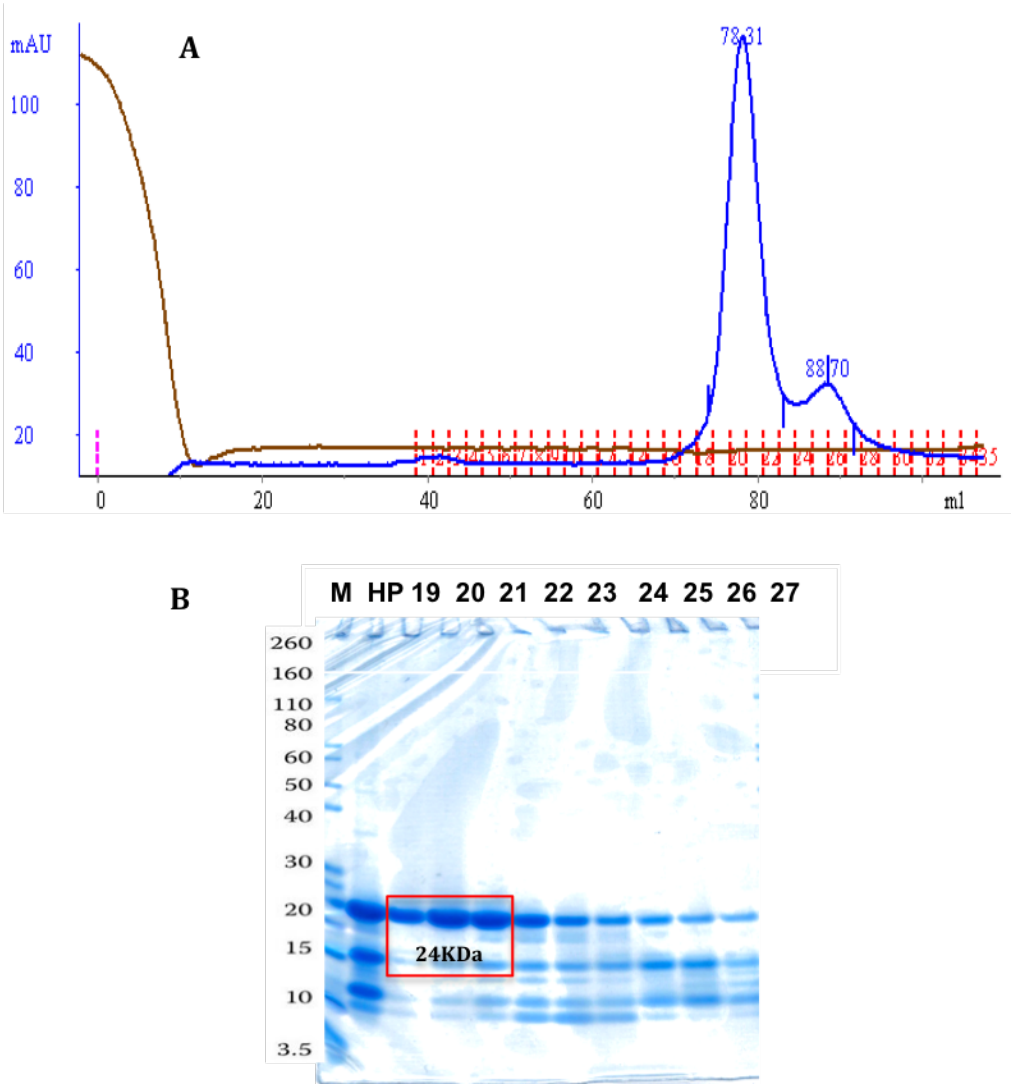


Figure 3.5. Gel-filtration for Rap endonuclease protein sample.

A) Chromatogram of Rap protein eluted from the final gel filtration on a Superdex-200ml column. in fractions 19-21 at a peak elution volume of 78.3 ml, which suggested formation of a dimer based on the chromatogram calculations. This peak was followed with a small peak at 88.7 ml elution volume, which corresponds to a monomeric form. The fractions 19-21 were combined, concentrated and used for crystallization. The blue line represents the UV absorption at 280 nm. **B)** SDS PAGE gel shows the purified Rap protein after elution from the Superdex-200 column. The majority of the Rap protein appears in lanes 3-5, with fewer contaminants. 15 µg of protein was loaded in each lane. M= Mark 12, HP= Heparin column protein.

3.1.1.6 Crystallization trials

The purified Rap protein was concentrated to 3 different concentrations (6, 18 and 26 mg/ml) using a 1 mL 10 kDa MWCO (Molecular Weight Cut Off) Viva spin column. The buffer was also exchanged into 5mM Tris-HCl, pH 8 and 0.1 M NaCl using a Zepa column.

The different concentrations of Rap were then used to set up crystal trials using various screening kits (PACT, MPD, JCSG, Ammonium Sulfate, Proplex and Classic). Trials using 96-well sitting drop plates containing the different crystallization conditions from each screen were set up by the Hydra robot (Section 2.17). Any crystals observed in the trials were tested for X-ray diffraction at the Diamond synchrotron, but no successful crystallization conditions were obtained for the protein

The trials were monitored for more that 12 months on a regular basis, but no crystals for the purified full-length Rap protein have yet been obtained.

3.1.2 Generating truncated versions of Rap endonuclease.

A new experiment was designed to re-clone the protein to remove a potentially disordered region at the beginning of the protein that might be stopping protein crystallization. The predicted secondary structure found using the PSIPRED Protein Sequence Analysis Workbench server (http://bioinf.cs.ucl.ac.uk/psipred/), showed a long coil in the first 30 residues of the N-terminal end, which might affect the crystallization of Rap protein due to formation of a mobile loop. Moreover, there is a very repetitive sequence at the N-terminal end of the protein found from approx. residue 39 to 70 of the protein, which includes a high percentage of R, K and E residues (Figure 3.6). Regions of sequence with little variation in types of residue are often associated with higher flexibility and may also account for problems with crystallization. Thus, two truncated versions of the Rap protein were made; P3 (full length Rap protein without the N-terminal 30 residues), and P4 (Rap protein without the N-terminal 70 residues).

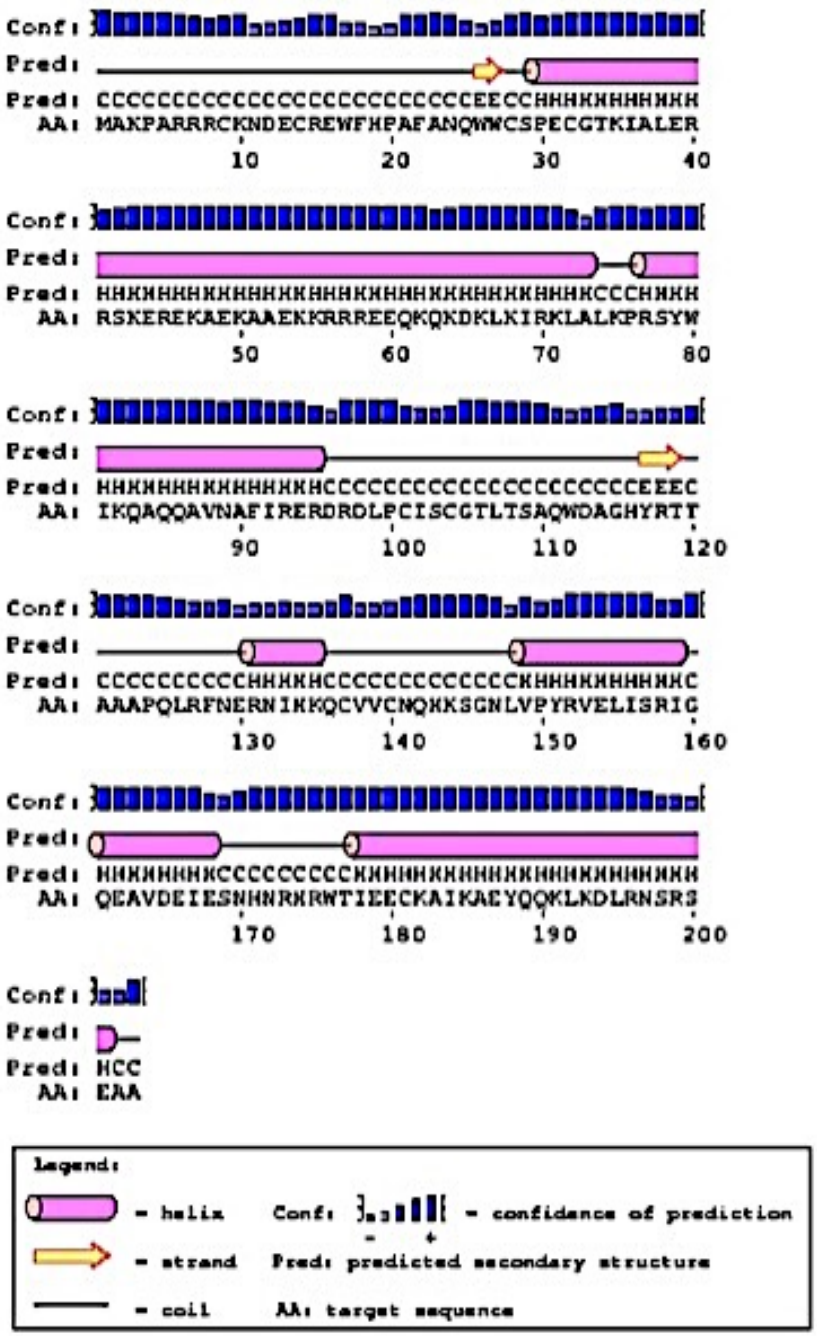


Figure 3.6. Predicted secondary structure.

The predicted secondary structure of Rap protein full-length sequence via PSIPRED Protein Sequence Analysis server showed a potential long coil (residues 1-30) and a very repetitive sequence (residues 39-70) at the N-terminal end. These were suspected to be potential reasons for the problems in crystallizing the protein.

3.1.2.1 Prediction of New Versions of Rap endonuclease structure

A predicted model of Rap P3 was produced using the Phyre2 server (Chapter 2) comprising 67 residues coverage (38%) of the total protein sequence (173 residues), while the predicted model of Rap P4 showed 53% coverage (71 residues) of the total protein residues (133 residues). Both protein models were given a 98-99 % confidence value by the program using of the structure of gve2 hnh2 endonuclease from the deep-sea thermophilic bacteriophage Geobacillus virus E2 as a template (pdb entry: 5H0O). The model structures of the truncated Rap proteins were built as two fragments by Phyre2 and showed a distinct helix at the N-terminal end of the first fragment followed by two anti-parallel beta strands in the second fragment, which are very similar to the full-length version (see 3.2.1), except the new versions appeared to show three distinct helices (Figure 3.7).

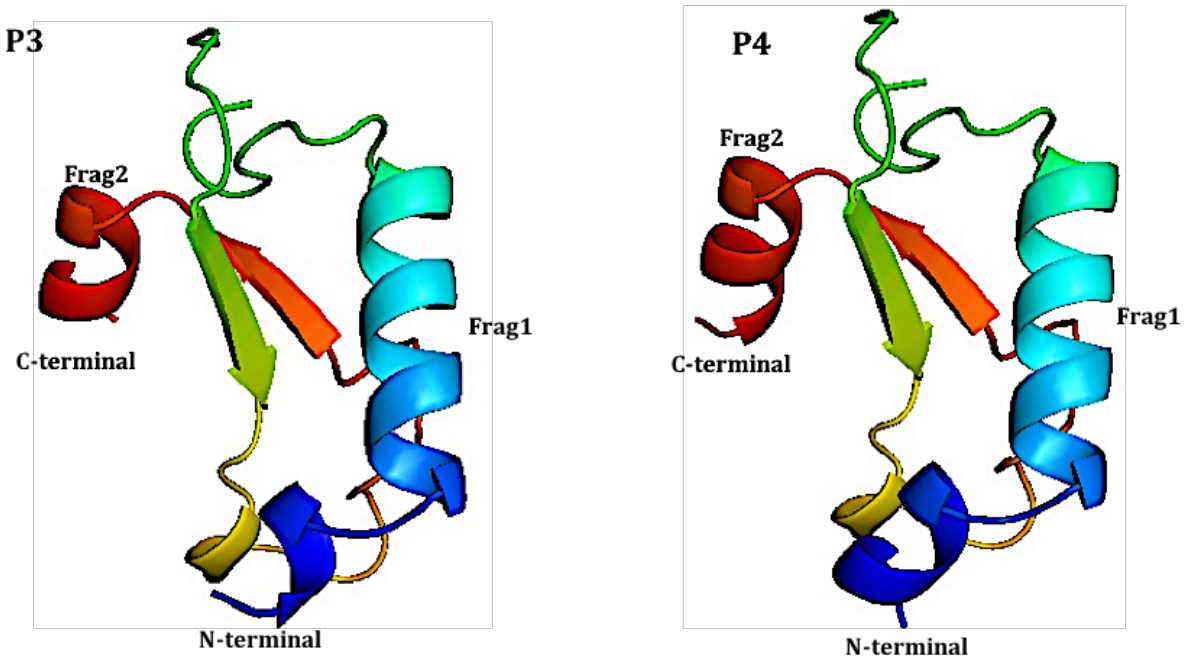


Figure 3.7. Prediction of truncated Rap endonuclease structures.

The predicted structure of the truncated versions of Rap endonuclease (P3, P4) modeled by the Phyre2 server. The predicted models were given a high level of confidence for both P3 and P4 but only covered 38% and 53% respectively of the total protein sequence, using a template of gve2 hnh2 endonuclease from the deep-sea thermophilic bacteriophage Geobacillus virus E2 (pdb: 5H0O).

3.1.2.2 Molecular cloning

The truncated versions of the *rap* gene were amplified using the vector for the full-length *rap* gene as the source of template DNA (Section 2.4). The Q5 High fidelity polymerase Kit was used to make the amplification of multiple gene fragments quicker (section 2.4). The primers were designed using the SNAPGENE program and supplied by Eurofins. The vectors used during this study with the primers details are listed in Table 2.3 (section 2.2). All the primers were designed to have a large consensus sequence with the target gene due to a high range of mutations produced within the PCR product sequences during the use of the normal consensus sequence primers.

3.1.2.3 Protein expression systems

Two different expression systems were used to clone the truncated versions of the *rap* gene to try to obtain a high level of Rap expression with a high solubility level. A *LacI* expression system (pET21b vector), and a Heat Shock expression system (pJonex4 vector) were used.

3.1.2.3.1 pET21b vector

The pET21b vector uses an IPTG system and is controlled by the *LacI* promoter, similar to pET22b, except that it is slightly smaller in size and does not have a signal sequence for potential periplasmic location. While, it carries T7-tag that is 280 amino acids fusion tag (Novagen, 1999).

3.1.2.3.2 PJONEX4 vector

The truncated Rap protein constructs P3 and P4 were sub-cloned into the pJonex4 vector (provided by Prof Jon Sayers, University of Sheffield). A pJonex4 vector depends on a heat-shock system to express the target protein via a heat inducible promoter (PLλ), which is under the control of a transcription repressor (cI857) (Sayers and Eckstein, 1991).

3.1.2.4 Cloning.

3.1.2.4.1 Gene amplification

The amplification of the genes for the truncated Rap proteins (P3 and P4) was carried out using the Q5 High fidelity polymerase Kit in the PCR reaction. DMSO was supplemented to improve the production of the target genes with reduced background contaminants, due to the high GC content of the Rap gene (Winship, 1989). The specific primers for each gene were used to obtain three PCR products of each protein with versions that contained N- or C-terminal locations for a six-histidine tag or no histidine tag to examine the effect on purification or crystallization.

The PCR amplification program was set up according to the Q5 polymerase protocol (Section 2.4). The temperatures were modified to 98 ˚C for 1 min for denaturation temperature, and 65 ˚C for 35 sec for annealing temperature.

The PCR products were checked on agarose gel (Figure 3.8), and then purified using agarose gel electrophoresis and a QIAquick® Gel Extraction Kit or PCR Purification Kit (Qiagen) by standard protocols (Section 2.4).

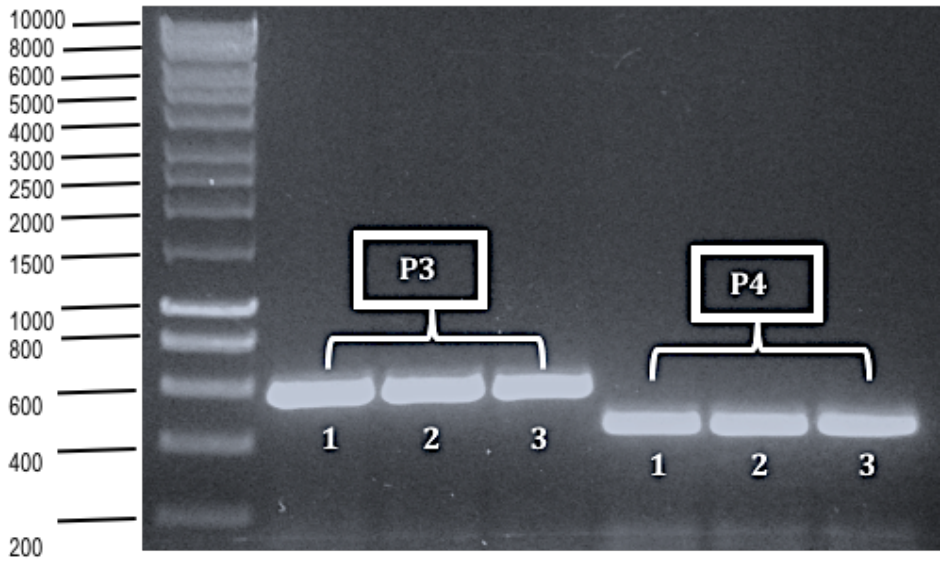


Figure 3.8. PCR amplification for P3 and P4 of Rap.

Agarose page shows the PCR products of the new truncated constructs of Rap endonuclease protein cloned with different vectors: P3= Long truncated protein (531bp). P4= Short truncated protein (390bp). (1) pET21b N-Terminus His-tag, (2) pET21b C-Terminus His-tag, (3) pJONEX4 C-Terminus His-tag.

3.1.2.4.2 Vector production and restriction digest

Miniprep was performed using the QIAprep® Miniprep kit (Qiagen) on the Silver Efficiency and *E. coli* M72 competent cell stocks, which contained the pET21b and pJONEX4 vectors, respectively, to increase the quantity of vectors and extract them from their respective cells for initiating cloning. The purified PCR products and plasmids were then digested using appropriate restriction enzymes (Table 2.4, section 2.6). Restriction digest was carried out for 30-45 minutes at 37 ˚C. Both vectors and inserts were analysed on an agarose gel to check the digestion process (Figure 3.9). They were then purified from the gel using the QIAprep® gel extraction kit (Qiagen) (Section 2.6).

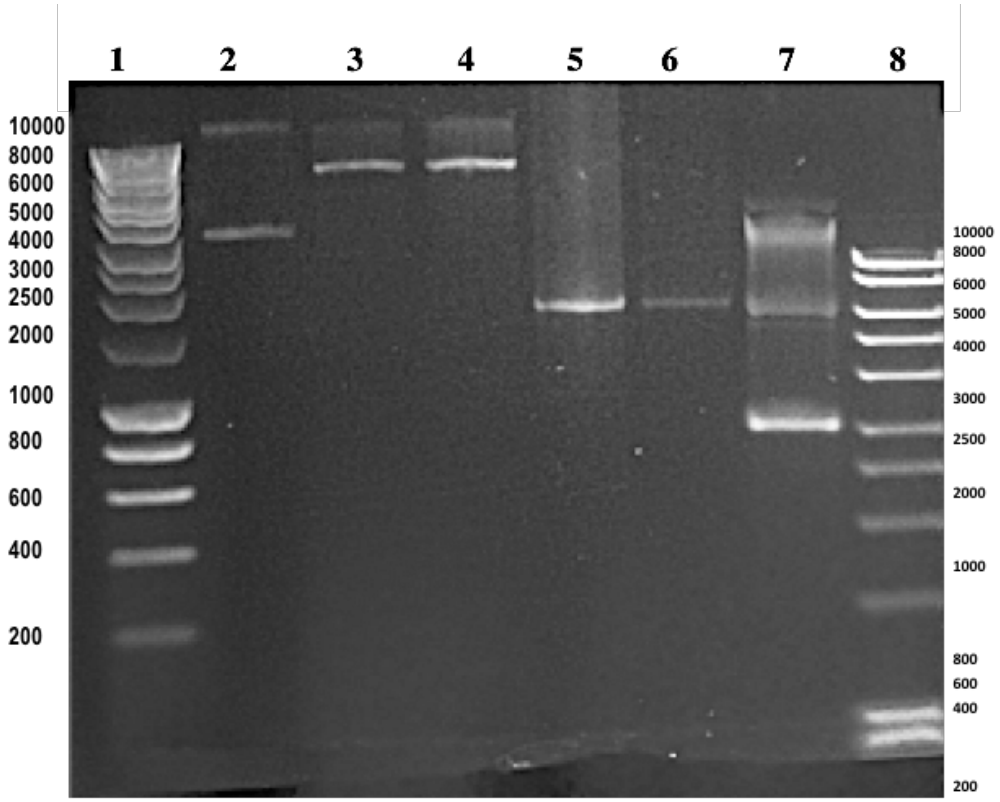


Figure 3.9. Restriction digestion for pJONEX and pET21b.

Agarose gel shows the digestion results of the pET21b and pJonx4 vectors using appropriate restriction enzymes depending on the site present in the primer sequence vectors (EcoRI and HindIII for pJonX4, and NdeI and XhoI for pET21b): (1 and 8) Hyper ladder, (2) Uncut pET21b, (3) Cut pET21b, (4) Cut pET21b control, (5) Cut pJoneX4 control, (6) Cut pJoneX4, (7) Uncut pJonx4.

3.1.2.4.3 Transformation and Colony screening

Both the vector and the insert were ligated together using T4 DNA ligase (NEB) and then transformed into their respective competent cells (pET21b construct into BL21 (DE3), pJONEX4 into M72). Cells then were plated on LB agar supplemented with 100 µg/ml Ampicillin for selection and incubated overnight, at 37 ˚C for the BL21 cells, and at 30 ˚C for the M72 cells (Section 2.8). Colonies were screened by PCR using the original primers for specific gene amplification (Table 2.3, section 2.2), to detect for the presence of the insert (Figure 3.10).

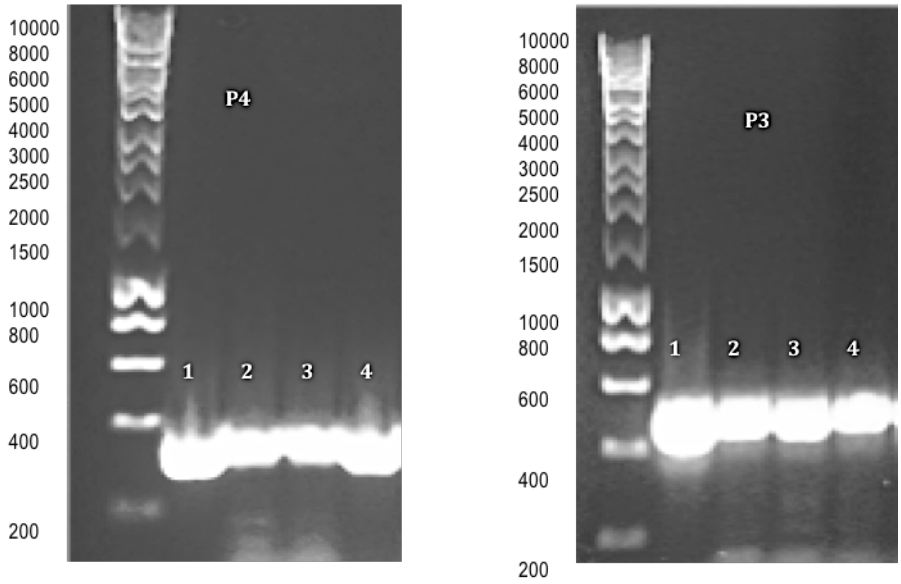


Figure 3.10. Colony screening of truncated Rap products (P3, P4).

Images show the results of the colony screen assay that confirmed the presence of the truncated products (P3 and P4) of the *rap* gene in the pET21b vector (1 and 2), and pJonex4 vector (3 and 4).

3.1.2.4.4 Gene Alignment

DNA samples from the constructs were recovered by miniprep, sequenced (GATC-Biotech) and compared to the expected sequence from the *rap* full-length gene. Alignment of the cloned sequences using Blastn and Blastp revealed 100% identity to the *rap* gene. The selected pET21a constructs were transformed into BL21 DE3 cells to express the proteins.

3.1.2.5 Protein Expression

3.1.2.5.1 Small Scale Expression and Solubility

One of the cloned colonies for each Rap construct was selected for small-scale expression studies.

3.1.2.5.1.1 pET21a constructs

The expression levels of the proteins (P3, P4) in BL21 (DE3) cells were compared at various concentrations of IPTG induction (0.1, 0.5, 1 and 2 mM), and tested for a range of incubation times after induction using the same strategy as employed with the full-length Rap protein (see section 3.1.1.4). The results showed a very low level of expression for both proteins, which was not satisfactory to run large-scale expression. Therefore, the experiment on these constructs was suspended.

3.1.2.5.1.2 pJONEX4 constructs

Small-scale expression of the Rap P3 and P4 proteins in M72 cells was carried out with a similar strategy as above except in adding 0.1mM IPTG for inhibition of the target protein under shaking incubation at 30˚ C before and after induction, which was accomplished by a heat shock strategy (2 hours incubation at 42˚ C) (Section 2.11.1). A similar protocol was used to check the expression and the solubility as used for full length Rap protein.

The small-scale expression of P3 and P4 proteins were optimized at 30˚C, using 0.1 mM IPTG, then 3 hours heat-shock induction at 42 ˚C, followed by 3 hours incubation after induction for P3 and 6 hours incubation for P4 at 30˚C again. Samples were analysed by SDS-PAGE and large bands were found on the gel after induction, at approx. 20 kDa, likely corresponding to the size of Rap P3, and at 15 kDa, likely corresponding to the size of Rap P4. These bands are not present in the pre-induction samples, revealing a high level of expression of Rap. The majority of this protein can also be found in the supernatant of the cell sample, indicating a high level of solubility (Figure 3.11).

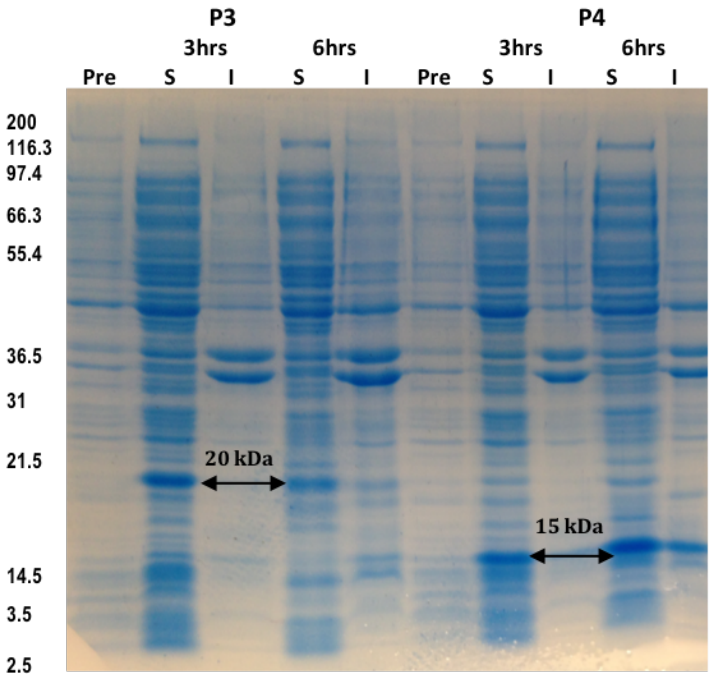


Figure 3.11. Small scale Overexpression and solubility for P3 and P4 truncated Rap proteins.

A 12% SDS PAGE gel showing the level of expression of Rap proteins. The results revealed that 3 hours incubation after induction gave the best expression level of P3 (20 kDa), while 6 hour incubation was the best for P4 (15 kDa). 15 µg was the final concentration of the supernatant loaded on the SDS gel. Pre= pre-induction, S= soluble, I= Insoluble

3.1.2.5.2 Large Scale expression

Large-scale overexpression of the Rap proteins was carried out using the optimized conditions for small-scale expression, using 6x 500 mL of 4YT media for each protein (P3, P4) supplemented with 1 ml of 1000x trace metals mix solution for a high density growth (Studier, 2005). Due to using a very rich component media and the short incubation periods, the induction happened when OD_{600} = 2 in order to grow a high density of cells for better total protein content. Samples were collected before induction and after 3 hours for P3 and 6 hours incubation for P4, and then analysed by SDS-PAGE (Figure 3.12). Around 10 g of cell pellet was harvested from 3 liters of growth media with each protein.

A comparable level of expression of Rap protein can be seen in the large scale over-expression compared with small scale expression. The high level of solubility looked suitable to do purification (Figure 3.12).

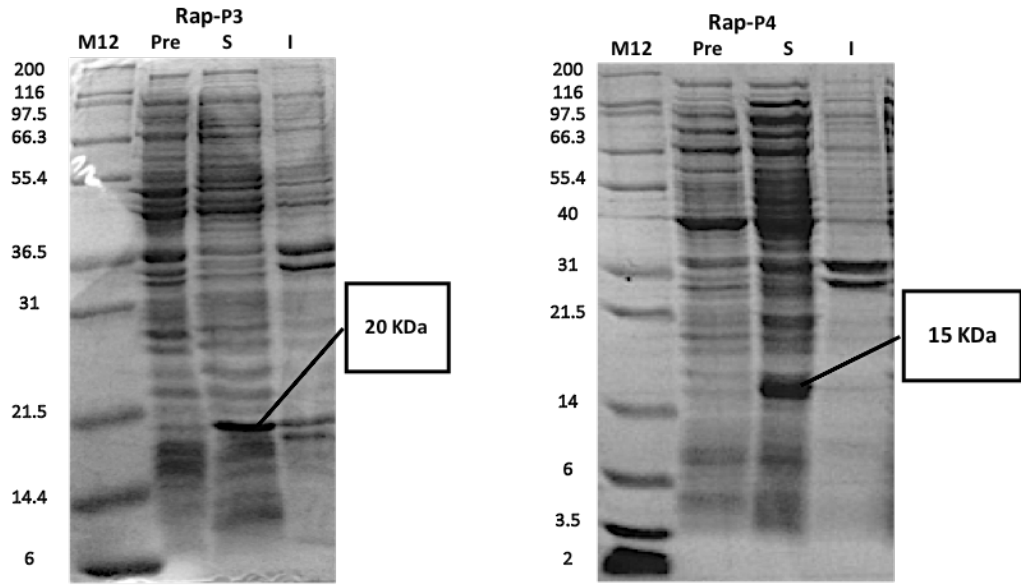


Figure 3.12. Large scale over-expression of P3 and P4 truncated Rap proteins.

12% SDS PAGE gel shows the soluble protein expression of P4 (15kDa) and P3 (20 KDa) before and after induction with appropriate incubation periods at 30 ˚C. 15 µg of protein is loaded in each lane. Pre= Pre-induction, S= soluble, I= Insoluble

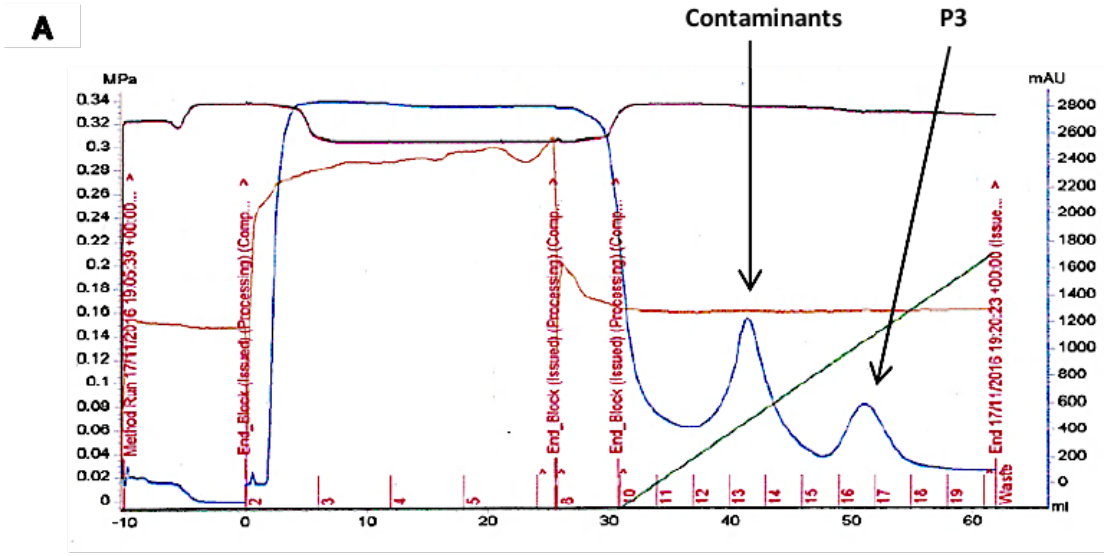
3.1.2.6 Protein purification

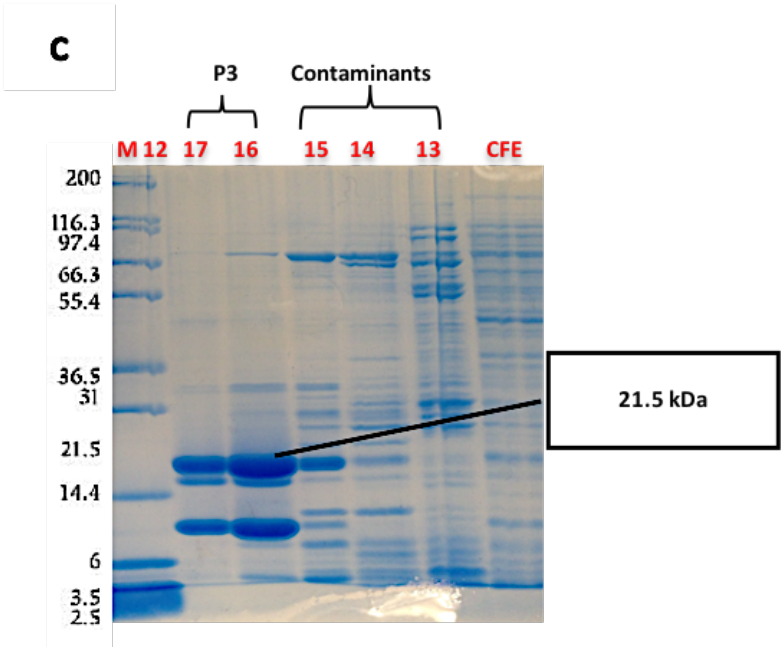
The sequences of the truncated Rap proteins (P3, P4) were analysed via the Expasy protparam server to provide the basic solution parameters, shown in Table 3.1 (http://web.expasy.org/protparam/).

Table 3.1. Basic properties of the truncated Rap proteins.

Protein	P3	P4
Size (kDa)	20.7 (21.5 6x His^+)	15.121 (15.9, 6x His^+)
Length (Amino Acid)	177	130
Isoelectric point (PI)	9.73	9.25
Extinction Coefficient	1.085	1.502
Predicted overall net charge in pH8 buffer	Positive	Positive

The tagged P3 and P4 proteins were purified by an efficient two-steps purification based on the C-terminal His tag. Firstly, a His-Trap Ni-HP 5 ml column purification (GE healthcare™). The results showed that both P3 and P4 protein proteins were eluted with 0.25 M Imidazole, and that this chromatography stage gave a high level of purity of P4 protein compared with P3 protein, which had many contaminants in its fractions based on the SDS-PAGE analysis (Figure 3.13).





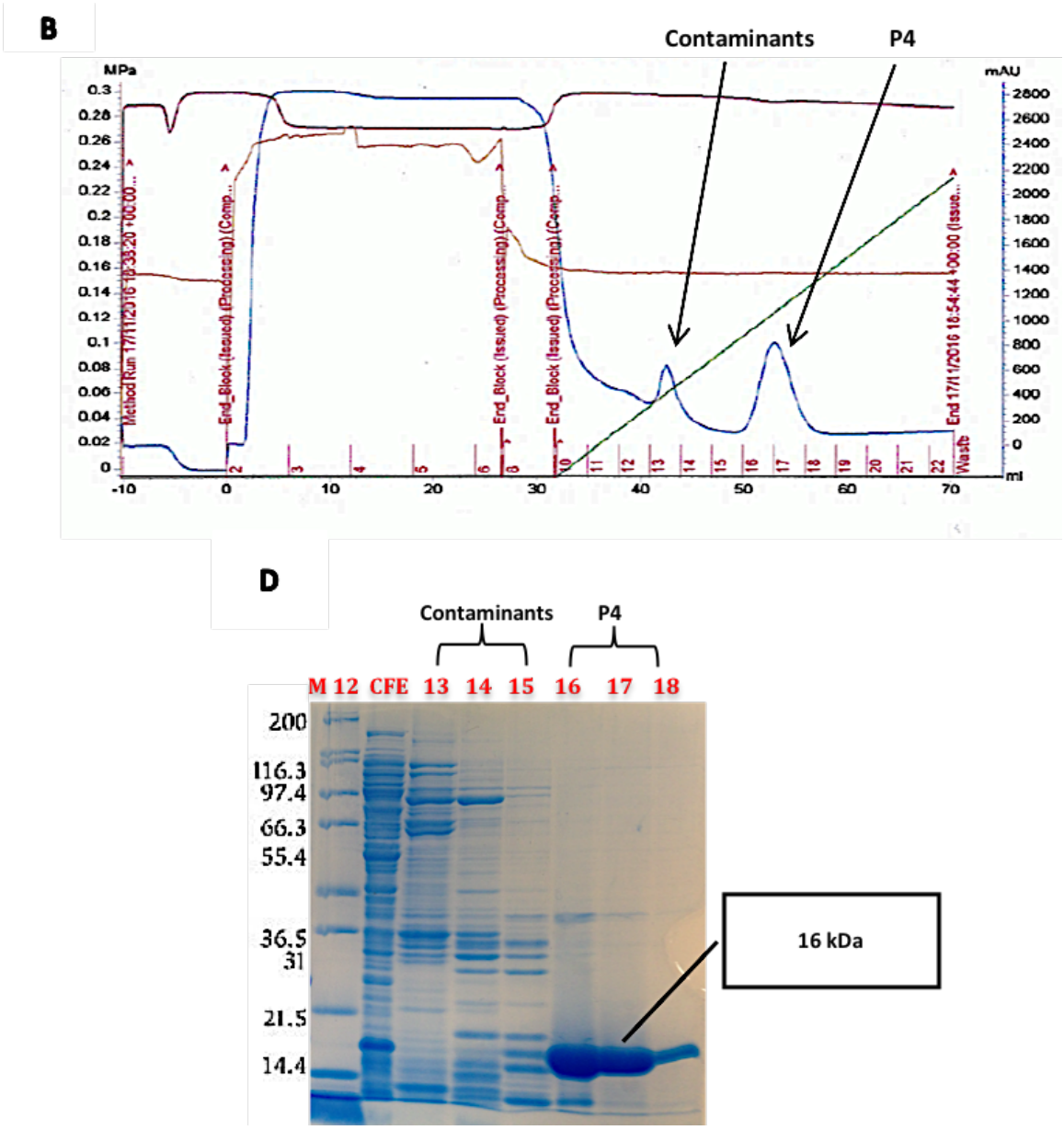
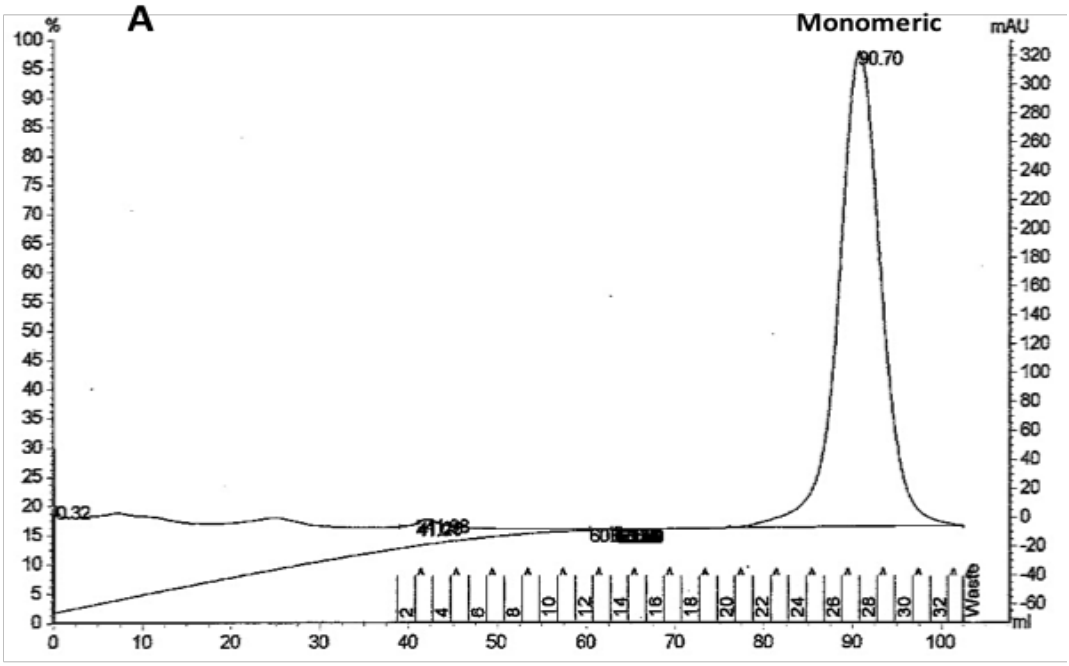
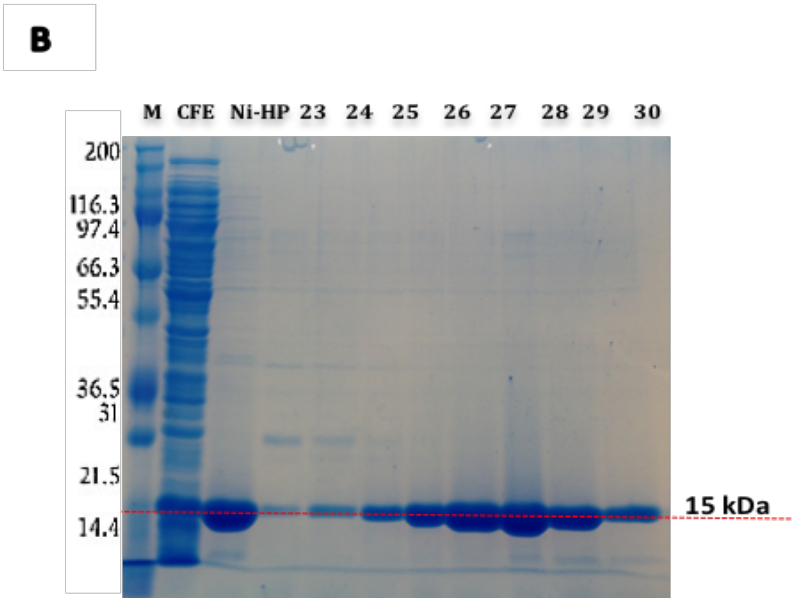


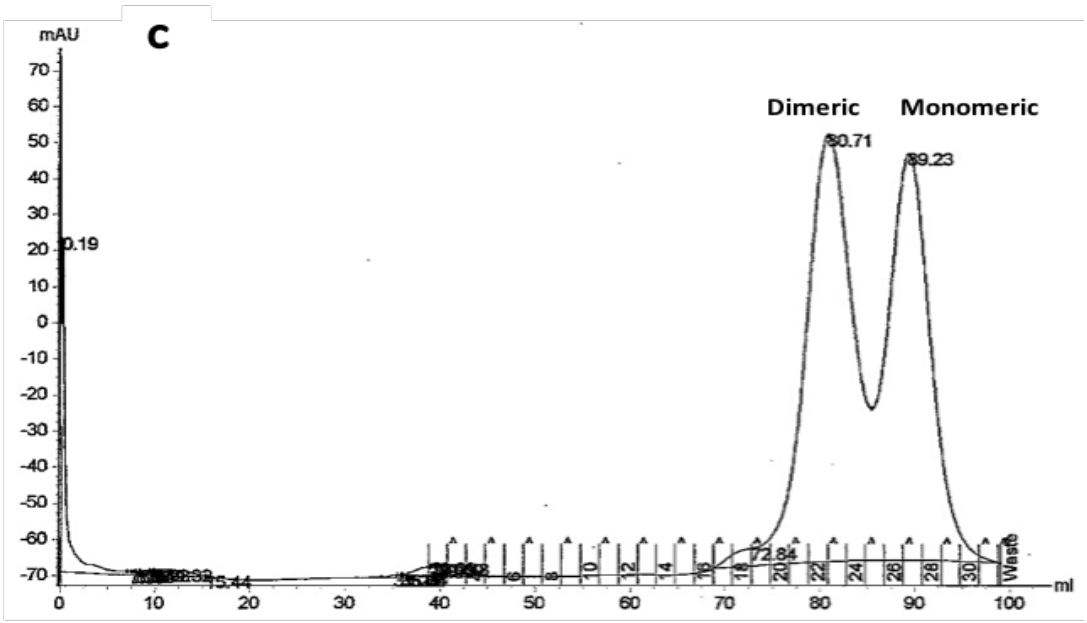
Figure 3.13. Ni-HP column purification column for the truncated Rap proteins.

A, C) The chromatograms illustrate the purification of P3 and P4 proteins on a Ni-HP 5 ml column. The results revealed that both proteins share similar characteristics in that two peaks were eluted at 43 and 55 ml, which correspond to contaminants and target proteins respectively. Both truncated Rap proteins were eluted at approximately 0.25 M Imidazole concentration in the second peak. **B, D**) SDS PAGE analysis of the fractions eluted from the Ni-HP columns with P3 and P4. The expected molecular weights for P3 and P4 (21.5 and 16 kDa, respectively) and the fraction numbers are shown. The P4 protein appeared to have approx. 95% purity, while the P3 protein seemed to be degrading and shows indications of possible protease cleavage. 15 µg of the protein was loaded in each lane.

A gel filtration column was used as a second step of purification. The fractions containing each truncated target protein from the Ni-HP column were merged and concentrated down to a volume of 1 ml with an approximate concentration of 4 mg/ml using 10000 kDa (MWCO) Viva spin column, and then loaded on a Superdex 200ml Gel filtration column (GE healthcare™) (Figure 3.14).







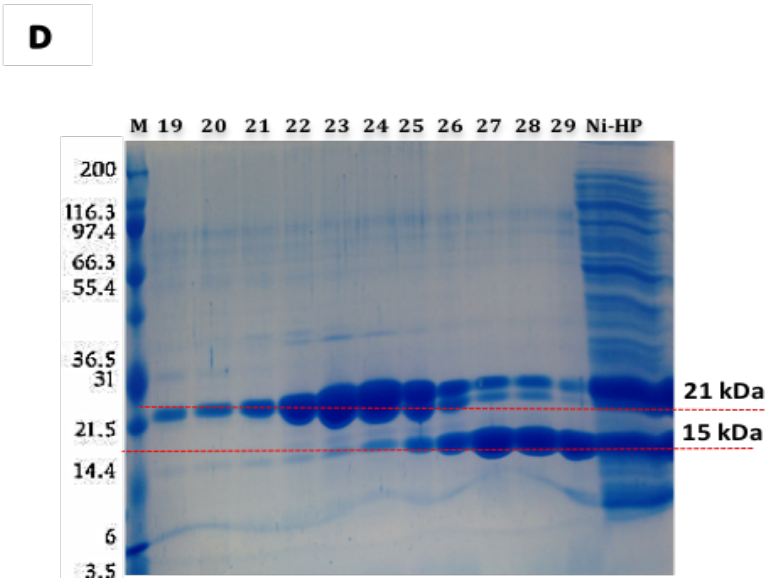


Figure 3.14. Superdex-200 gel filtration column purification for the P3 and P4 truncated Rap proteins.

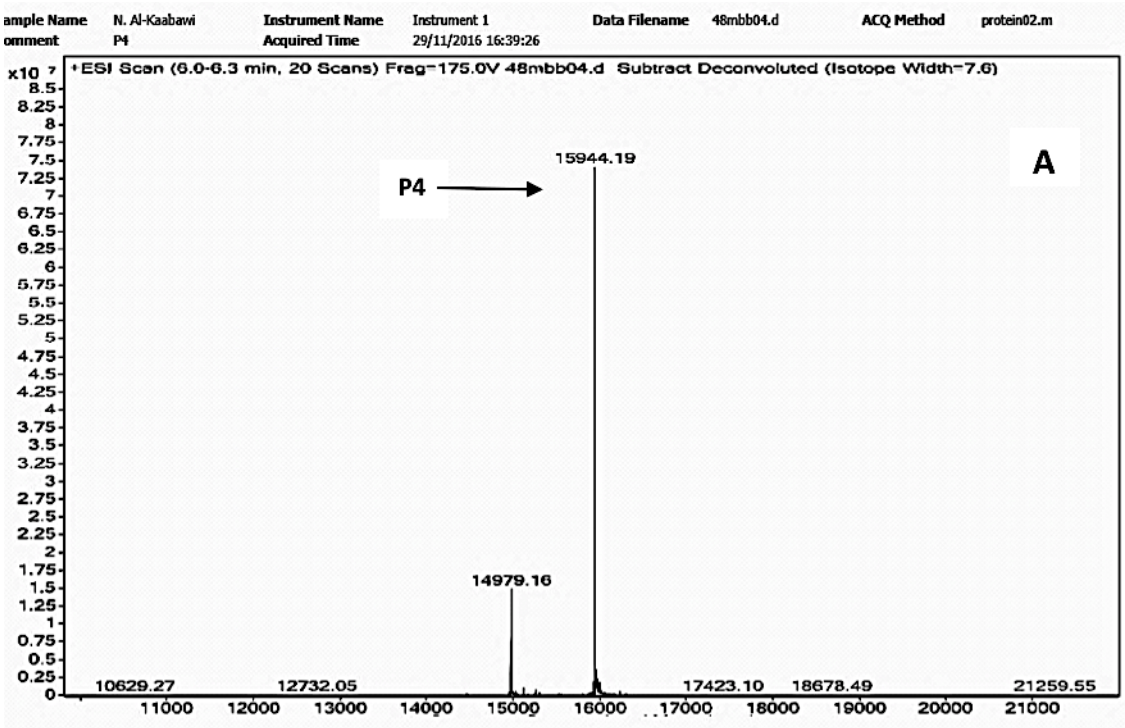
The chromatograms show the elution of P3 and P4 proteins from a Superdex-200 gel filtration column. **A)** A single peak of P4 protein was observed in fractions 23-30, which eluted at 90.7 ml corresponding to a molecular weight of 16 kDa, the size of one monomer. **C)** The P3 protein eluted in two peaks after 80 ml (fractions 19-22) and 89 ml (fractions 26-29), which correspond to proteins of molecular weight of 32 and 16 kDa, respectively. The 32 kDa size was equivalent to a dimeric form of the protein, while 16 KDa was less than the expected size of a monomer. **B)** and **D)** SDS-PAGE analyses of the P4 and P3 elution from the Superdex^{TM}200 column, respectively. Fraction numbers, molecular weight markers (M), cell free extract (CFE) and the sample loaded from the previous column (Ni-HP) are shown as labels on the lanes. The gel in (**D**) shows the fractions from the two eluted peaks of the P3 protein sample. 15 µg of the protein was loaded in each lane.

The P4 protein formed a single peak, which eluted after 90 ml, corresponding to a K_{av} of 4.2, and equivalent to a 16 kDa molecular weight. This size was consistent with the estimated monomeric size of P4 protein. By contrast, there were two peaks formed for the P3 protein that eluted at 80 and 89 ml, which corresponded to K_{av} values of 4.5 and 4.2, respectively, and calculated masses of 32 kDa and 16 kDa, respectively. The 32 kDa molecular weight is consistent with the anticipated dimeric form of P3 protein, while the 16 kDa is less than the actual size of the P3 (21.5 kDa). However, it is more likely to be a monomer as shown in figure 3.14.

3.1.2.7 Analysis of mass spectrometry for P3 and P4 proteins

Samples of both the purified P3 and P4 truncated Rap proteins were sent for Mass-spectrometry analysis (Chemistry Department, University of Sheffield) to check the exact molecular mass.

The results showed that the P4 protein sample was exactly the anticipated molecular weight (15944 Da) including the C-terminal His-tag (Figure 3.15 A). The analysis of the P3 sample gave a range of masses. Major peaks in the spectrum were seen at 1511, 16304, and 20703 Da. It was not easy to identify a peak corresponding to the size of the P3 protein with a C-terminal His-tag (21522 Da). However, it is possible that the 20.7 kDa peak may belong to the P3 protein after loss of the C-terminal His-tag (Figure 3.15 B).



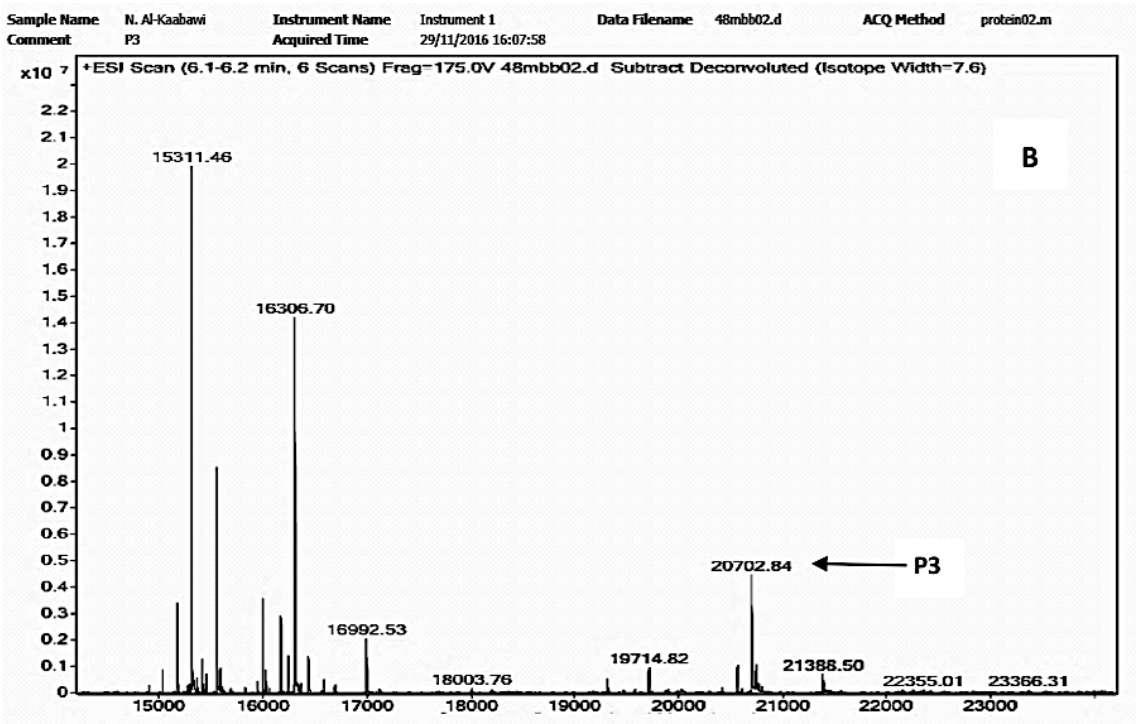


Figure 3.15. Mass-spectrometry analysis of samples of P3 and P4 truncated Rap proteins.

A) The spectrum of the P4 protein shows a peak of protein of exactly the anticipated molecular weight (15944 KDa). **B)** The spectrum of the P3 protein shows a number of peaks, which correspond to various sizes. This spectrum also shows three major peaks at 1511, 16304, and 20703 Da, where the third peak (20703 Da) is exactly the size of the P3 protein without the C-terminal His-tag.

3.1.2.8 Crystallization trials for P3 and P4 proteins

The purified P4 from fractions 25-30 from the gel filtration column was concentrated to 8 mg\ml. Similarly, fractions 19-23 from the gel filtration purification of P3 were merged and concentrated to 7 mg/ml. The proteins were buffer exchanged into 5 mM Tris-HCl pH8 with 100 mM NaCl using Zeba columns. The proteins were not fully stable in this buffer and kept precipitating, especially the P3 protein. However, the purified proteins were used to set up crystallization trials in 96 well sitting drop plates with 6 different screens of conditions as preformed for full length Rap protein (see section 3.1.1.6). No crystals have been observed yet.

3.2 NinH protein

3.2.1 Prediction of NinH structure

The initial stage of the investigation of NinH was to find homologues and a predicted structure for NinH. The Phyre2 server results gave a predicted model analysed with 22 residues or 34% coverage of the protein sequence and a 91.9% confidence value for a model based around a DNA/RNA-binding 3-helical bundle protein of the FIS-LIKE family (pdb entry: 1ETO). Thus, the model for the structure of NinH showed three distinct helices at the N-terminal end (6^{th} -36^{th} residue) in the shape of a triangle, which share 26% sequence identity with a FIS-LIKE family member, followed by a 16 residue helix and then ending with a four residue helix at the C-terminal end (Figure 3.16).

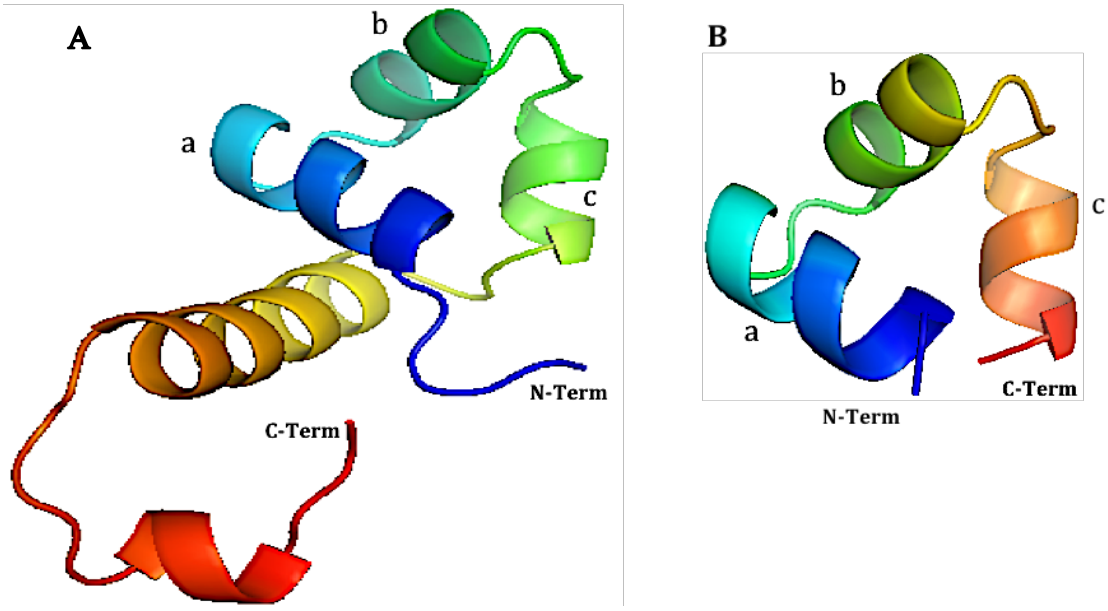


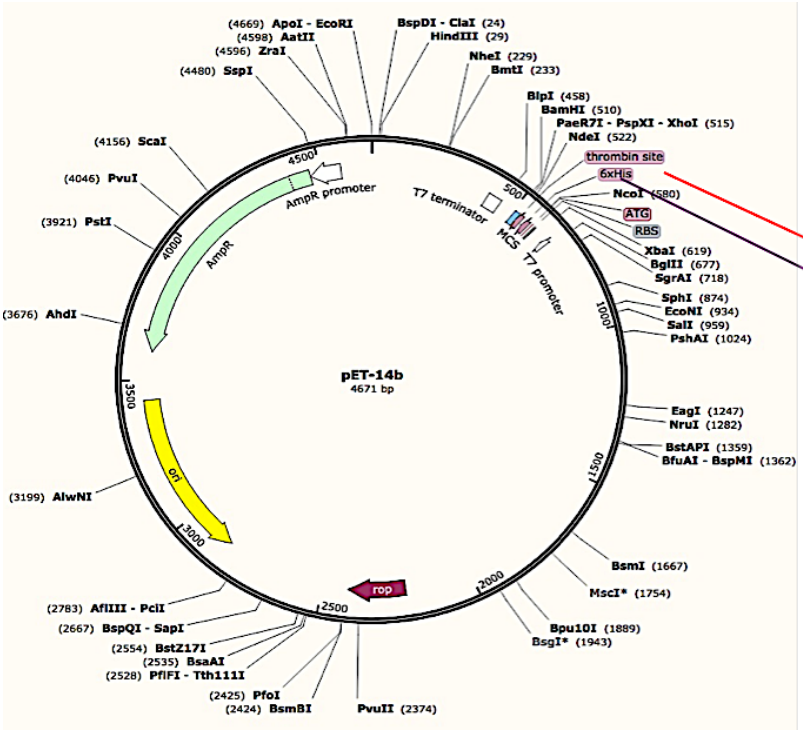
Figure 3.16. Predicted NinH endonuclease structure.

The predicted structure of NinH protein modeled and analyzed by the Phyre2 server is shown in (**A**). A 91.9% confidence value was assigned to the prediction that covered 34% of the protein sequence (22 residues), and the model was based on a DNA/RNA-binding 3-helical bundle protein of the FIS-LIKE family (pdb: 1ETO) shown in (**B**).

3.2.2 NinH cloning background

The *ninH* gene was cloned in the laboratory of Dr. Gary Sharples (University of Durham) into a pET14b vector, which allowed for the potential of an N-terminal histidine tag (Figure2). The pET14b vector contains an ampicillin resistance marker and relies on a *Lac* operon system in the BL21-AI competent cells. Cloning used NcoI, NdeI and BamHI restriction sites to produce His-tagged and wildtype versions. The protein was cloned to have an N-terminal His-tag with a cleavage property using a thrombin site.

The pET14b vector is very similar to pET22b (see 3.1.1.2), which involves a strong cloning and expression system for targeting genes, controlled as well by a bacteriophage T7 system of transcription and translation and is under *Lac* control. The recombinant plasmid was transformed into the *E. coli* host cells (BL21-AI), which contain a chromosomal copy of the T7 RNA polymerase gene, *LacUV5* promoter and *LacI* gene that are a bacteriophage DE3 system lysogenized into the genomic DNA of the host cell. The induction system involved addition of an appropriate concentration of IPTG (Isopropyl-ß-D-thiogalacopyranoside) added to a growing culture for inducing the LacUV5 promoter to direct transcription of the T7 RNA polymerase, which in turn transcribes the target gene in the plasmid (Studier and Moffatt, 1986). The Heat Shock method was the method followed to transform the construct into the BL21 AI protein competent cells (Figure 3.17).



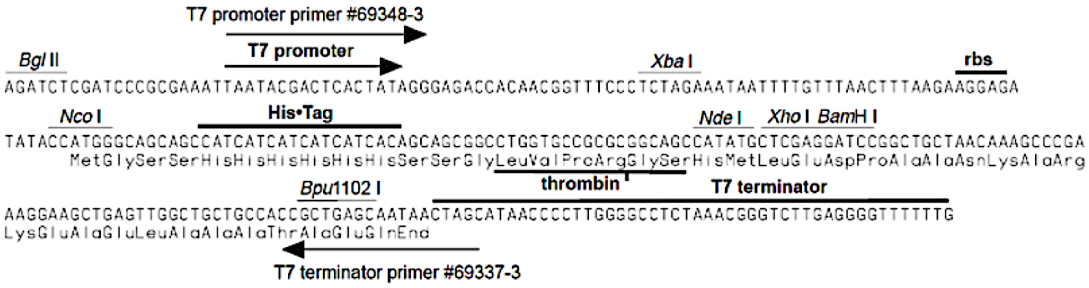


Figure 3.17. pET-14b vector map.

The *ninH* gene was cloned into the pET14b vector containing an ampicillin resistance marker using NdeI and BamHI restriction sites to provide the protein with an N-terminal His-tag for optional cleavage property at a thrombin site as shown. This figure was drawn by Snapgene program.

3.2.3 Colony screening

The BL21 AI cells containing the recombinant vector were plated out on LB agar plates supplemented with 100 μg/mL ampicillin, to obtain single colonies. The same investigation strategy used with the *rap* gene was followed to ensure the sequence accuracy and the absence of mutations, by picking several colonies and doing colony screening (see section 2.9). The constructs of confirmed colonies were isolated using the Qiagen Plasmid Miniprep Kit, and then sequenced by the GATC sequencing company (Section 2.10). Clear sequencing reads were obtained for a few colonies. Sequence alignment with the standard λ phage *ninH* sequence in NCBI GenBank came up with 100% identity using Blastn and Blastp.

3.2.4 Protein Expression

3.2.4.1 Small Scale Expression and Solubility

A colony containing the correct sequence of *ninH* was selected for small-scale and large-scale over-expression tests, which were the same as used for the expression tests on Rap protein (Section 3.1.2.5). Varying concentrations of IPTG under different temperatures of incubation (37, 30, 25, and 16 °C) were tested to compare their levels of protein expression. 1 mL samples were collected before induction, and at 3, 6, 18, and 48 hours after induction, and analysed on a 12% acrylamide SDS gel. The small-scale expression of NinH protein was optimized at 16 °C, using 1 mM IPTG and 0.2% L-arabinose, and at 48 hours incubation after induction (Figure 3.18). A large band was found on the gel after induction, between the 14 kDa and 6 kDa markers, likely corresponding to the 10 kDa size of N-terminal His-tag NinH protein. This band was not present in the pre-induction sample, revealing a high level of expression of NinH protein, and it was also found in the supernatant of the cell sample, indicating a high level of solubility. However, there was no sign of expression at all for constructs of the wild-type protein. Therefore, the plan was to carry on with the 6xHis tagged product and cleave it from the His-tag later to obtain the wild-type version for crystallization use.

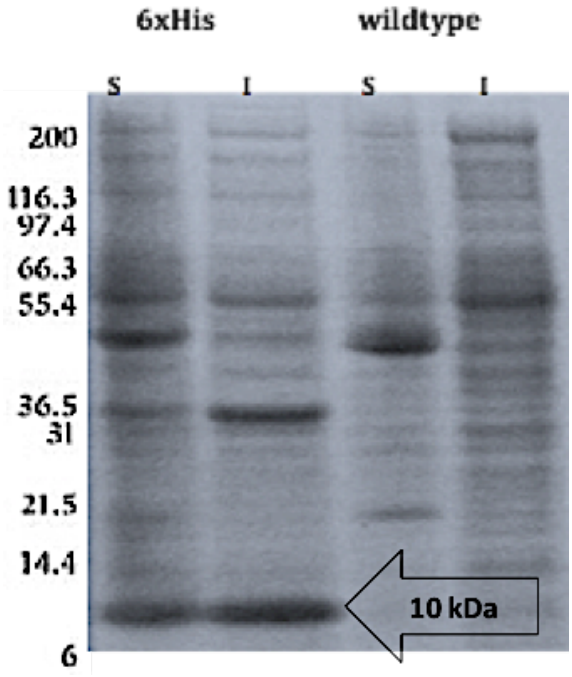


Figure 3.18.Overexpression and solubility for NinH.

A 12% PAGE-SDS gel showed a reasonable level of soluble expressed N-terminal 6xHis tagged NinH, at around 10 kDa based on a Mark12 ladder, compared with insoluble sample and the samples of the wild-type version of the protein. A sample of 15 µg supernatant protein was loaded on the SDS gel. The protein was expressed at 16 °C, 0.2 % L-Arabinose, and 48 hours incubation after induction. The cells were induced with 1mM IPTG when the OD_{600} was 0.6. S= soluble, I= Insoluble.

3.2.4.2 Large Scale expression

Large-scale culture growth and overexpression of the NinH protein was carried out using the optimized conditions for the small-scale expression, using 5x 500 mL of LB media in 2 L flasks. Samples were collected before induction and 48 hrs after induction at 16 °C and analysed on SDS-PAGE. All cells were harvested 48 hrs after induction. 11 g of cell pellet was collected from 2.5 L of growth media.

3.2.5 Purification of 6xHis NinH protein

The Expasy ProtParam server results for the 6xHis tagged NinH protein sequence showed that the exact molecular weight of 88 amino acids involving 6xHis tag and a thrombin cleavage site sequence was 10044.5 Dalton, and the predicted isoelectric point (pI) is 9.8, which makes the NinH protein more likely to be a positively charged protein in pH 8 buffer. The Extinction Coefficient was 8480 $M^{-1}cm^{-1}$.

The purification method for NinH protein was exactly the same as used for 6xHis-tag short-length Rap proteins (Section 3.1.2.6), except that the 6xHis-tag NinH protein was eluted with 0-0.7 M Imidazole over a 75ml gradient and collected in 2.5 ml fractions at a flow rate of 2.5 ml/min. The fractions from the His-Trap Ni-HP column were checked by SDS-PAGE, which showed that this chromatography stage accomplished almost a 95% level of protein purity. The eluted 6xHis NinH fractions were combined, which gave a 15 ml volume at 2.5 mg/ml concentration. Thus, a 37 mg sample of NinH purified protein was collected from 5 g cell paste (Figure 3.19).

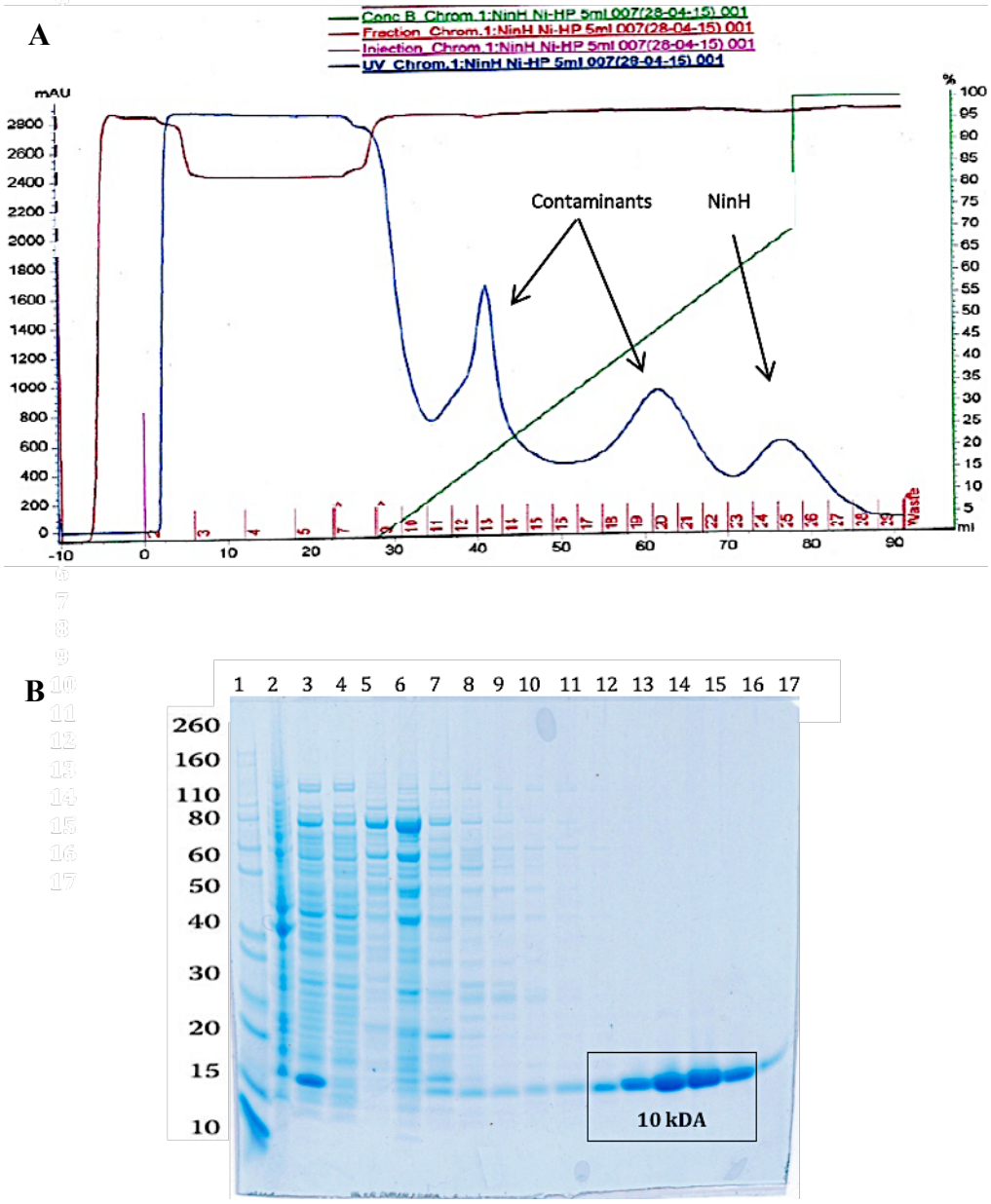


Figure 3.19. Purification strategy for 6xHis-tag NinH protein sample.

A) The chromatogram shows the NinH protein eluted from a His-Trap Ni-HP 5 ml column (GE healthcare™ 200) purification step. The blue line represents the UV absorption at 280 nm and the brown line refers to the conductivity. The results showed that the protein was eluted in fractions 23-27 (15 ml volume 2.5 mg/ml concentration, 37 mg total). B) The fractions were run on the SDS PAGE gel, which revealed approximately 95% purity of 10 kDa NinH protein in the lanes 12-16. 15 µg of the protein was loaded in each lane.

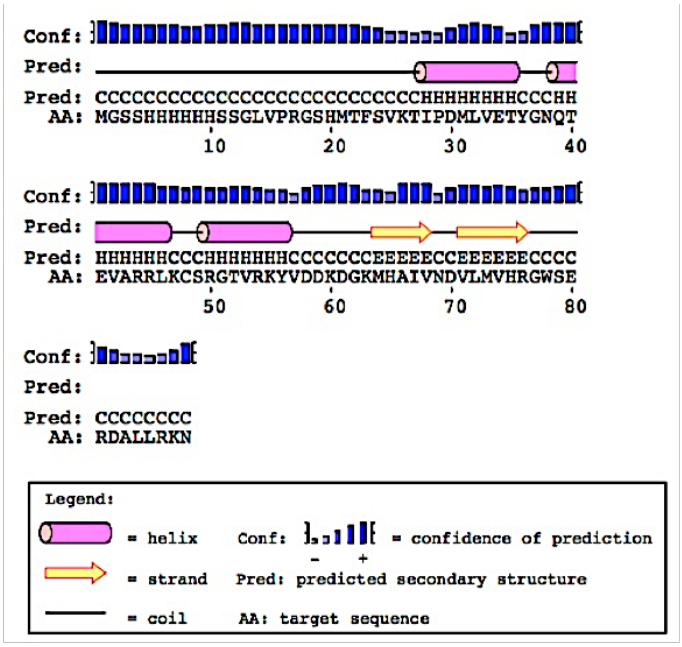
3.2.6 Crystallization trials for 6xHis NinH protein

The purified NinH fractions were concentrated using a 50 mL 10 kDa Vivaspin column. The buffer was also exchanged into 5mM Tris-HCl, pH8 and 0.1M NaCl using a Zepa column.

For crystallization trials, 96-well plates containing the different crystallization conditions were employed, with protein samples at different concentrations (11, 16, and 21 mg/ml). NinH protein was then used to set up sitting drop crystal trials using various screening kits (PACT, MPD, JCSG, Ammonium Sulphate, Proplex and Classic). Each screen was set up by the Hydra robot. No crystals were observed.

3.2.7 Predicted secondary structure of 6xHis-tag NinH protein:

The results from a PSIPRED Protein Sequence Analysis Workbench server (http://bioinf.cs.ucl.ac.uk/psipred/) analysis revealed that the 6xHis tagged NinH predicted secondary structure contains possible loops or unstructured regions as expected in the first 25 amino acids of the N-terminal end formed mainly by the 6 histidine and thrombin site sequences, and also at least 12 residues at the C-terminal end. These disordered parts might be the reason for not crystallizing the 6xHis NinH because of problems aligning these flexible regions with each other and forming crystal lattices (Figure 3.20).



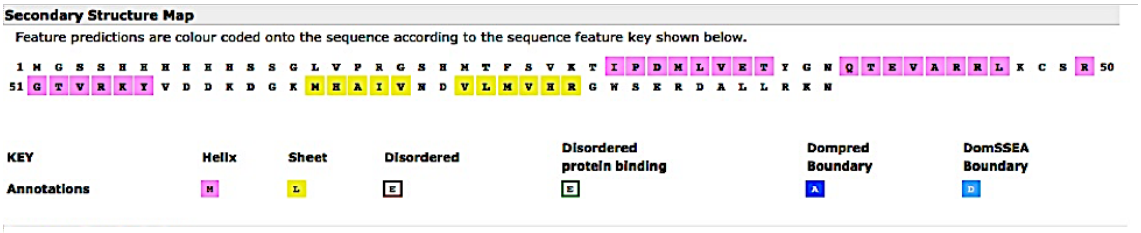


Figure 3.20. Predicted secondary structure for NinH protein.

The predicted secondary structure of 6xHis-tag NinH protein sequence via PSIPRED Protein Sequence Analysis server showed a long coil (1-25) at the N-terminus end due to the histidine and thrombin sequence, and a coil (76-88) at the C-terminal end. These could be the reason for failure to crystallize the protein.

3.2.8 Excluding the 6xHis-tag from NinH protein

Since the long coil at the N-terminal end was suggested as a strong possibility for not crystallizing the protein, a volume of 1.4 ml, 10.3 mg/ml 6xHis tagged NinH protein was used for experiments to remove the tag through addition of 90 µl of thrombin protease and 200 µl of cleavage buffer (100 mM NaCl, 2.5 mM CaCl2, 50 mM Tris-HCl pH 8). The NinH and thrombin mixture was then dialysed against the same buffer overnight at either room temperature or at 4 ˚C. The samples were centrifuged at 20000 x g for 3minutes after transferring from the dialysis membrane tube into a 2 ml centrifuge tube.

A small pellet was formed on the bottom of the tube that reflected the activity of the protease. The supernatant was collected in a volume of 1.6 ml, 5.5 mg/ml protein (9 mg in total) and a sample analysed by SDS-PAGE, which revealed a successful result of cleavage (Figure 3.21). The sample was diluted with 10 ml of 50 mM Tris-HCl pH8 to apply on a 1 ml Heparin column (Section 2.14.5). The results of the chromatogram were not clear enough to determine the elution peak of the target protein, but it was confirmed via a 15% acrylamide SDS-PAGE that showed the protein eluted in fractions 17-19 with its proper size (Figure 3.22).

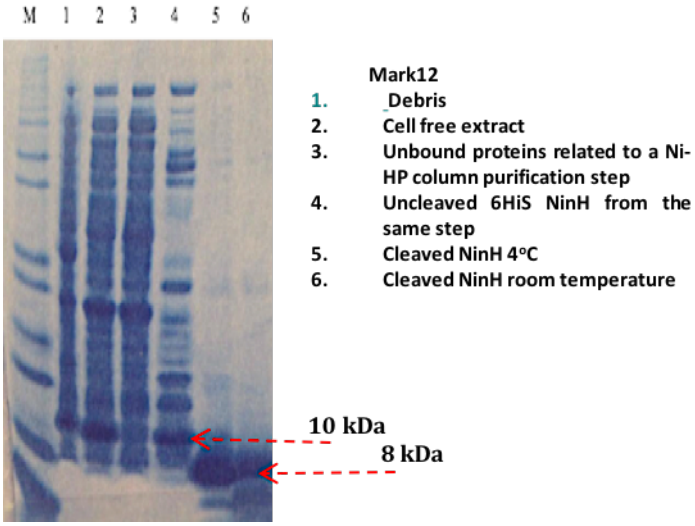


Figure 3.21. 6His-tag excision of NinH protein.

A 15% SDS gel showed the 8 kDa cleaved NinH protein compared with 10 kDa 6His-tag NinH protein, using thrombin protease and 18 hours dialyzing either at room temperature or 4 ˚C, which gave a similar result.

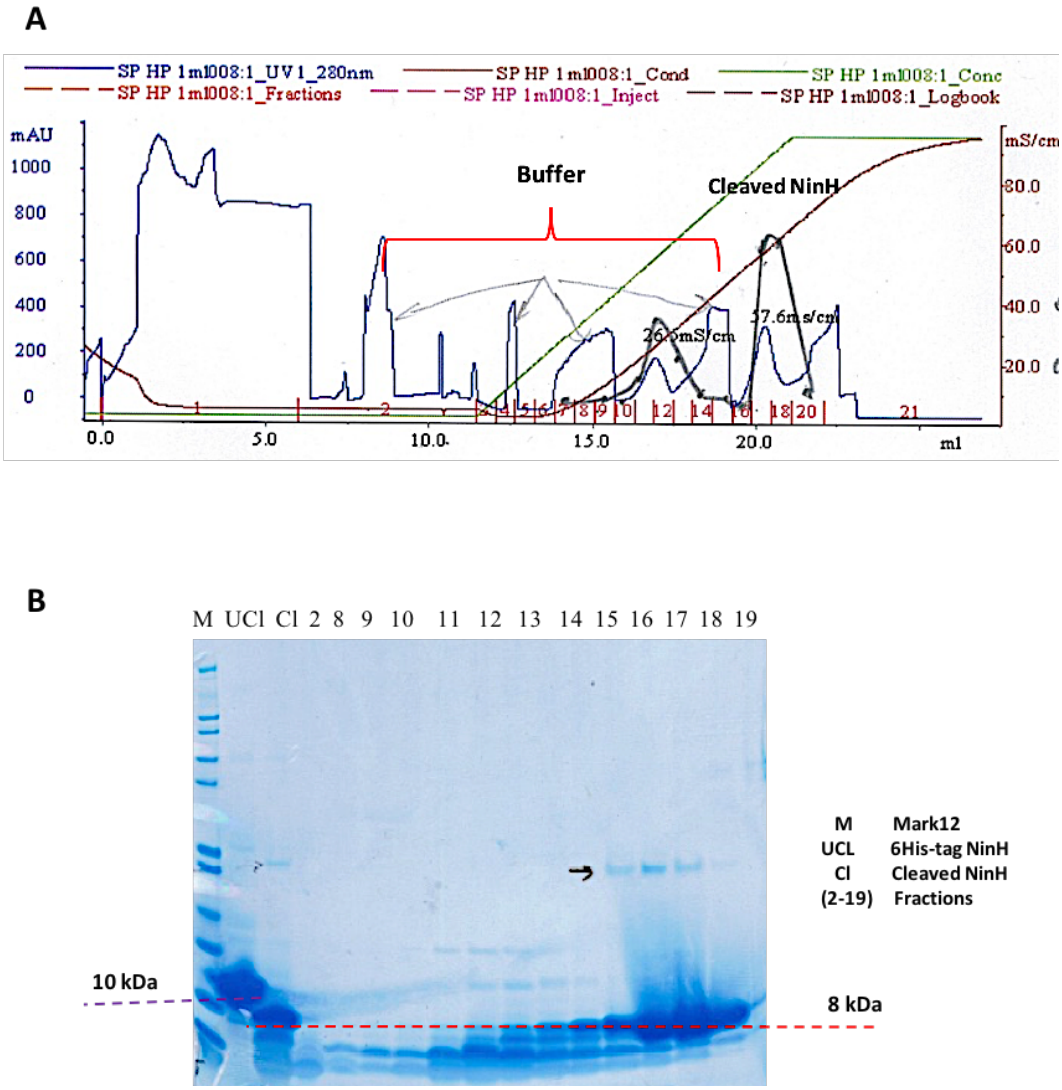


Figure 3.22. Cleaved NinH protein purification.

A) Chromatogram shows many peaks via His-Trap Ni-HP 1 ml column (GE healthcare™ 200) purification step. These peaks refer to many contaminants fractions 2-16, which may be related to the degraded particles of protein, thrombin protease, and the buffer components. **B)** The SDS PAGE shows that the purified NinH protein was eluted in fractions 17-19. These fractions were combined and concentrated down to volume of 160 µl, 8 mg/ml to set up crystallization trials.

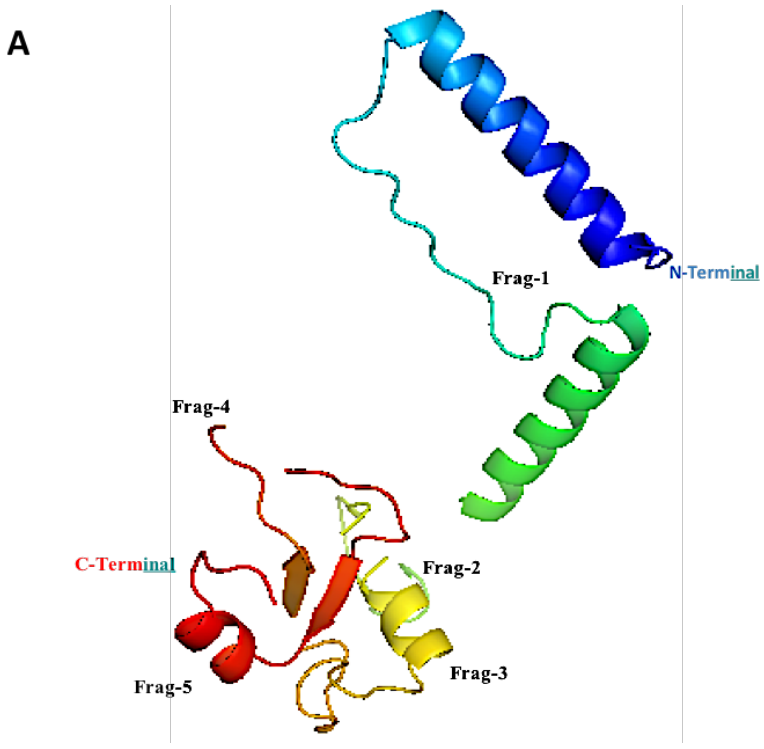
3.2.9 Crystallization trials for cleaved NinH protein

The purified NinH protein was used in a buffer of 50 mM Tris-HCl, pH 8 and 0.5 M NaCl because of instability problems, where the protein was seen to precipitate with a lower concentration salt buffer. Different crystallization conditions were screened, with an 8 mg/ml protein sample of NinH protein using various screens (PACT, MPD, JCSG, Ammonium Sulphate, Proplex and Classic). Each screen was set up by the Hydra robot. No Crystals were observed.

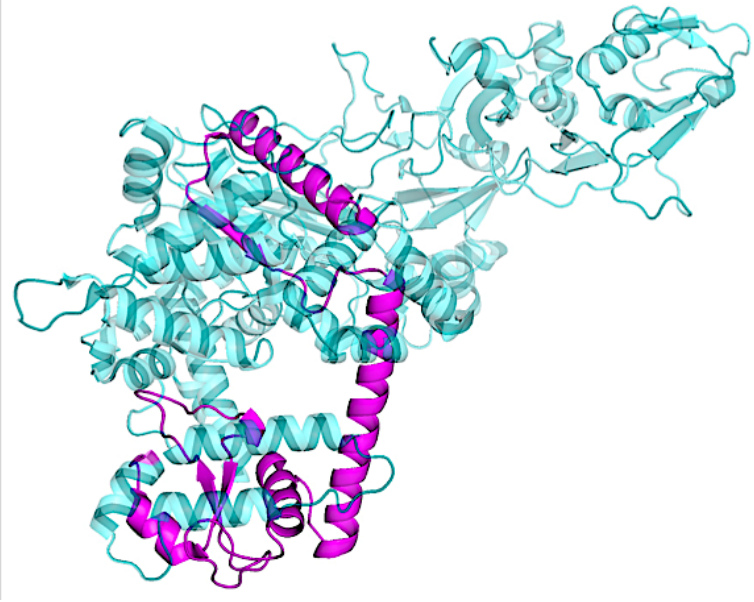
3.3 Ea31 protein

3.3.1 Prediction of Ea31 structure

A predicted model for Ea31 protein was produced via the Phyre2 server with 134 residues coverage equivalent to 45% of the protein sequence and a 98.9% confidence value using a homologue of the type ii-c cas9 enzyme from *Actinomyces naeslundii* (hnh endonuclease domain protein, pdb entry: 4OGC) as a template. The structure of Ea31 was built as five fragments, which share 22% sequence identity with the type ii-c cas9 enzyme. It also showed two distinct long helices at the N-terminal end of the first fragment followed by two more short helices in fragments 3 and 5. Additionally, anti-parallel beta strands were formed in the fourth and fifth fragment (Figure 3.23).



B



C

```
Ea31      ------------------------------------------------------------MKKL    4
cas9      ATDRIMRQTKISSLKTWWEEADSEQRGAMIRYLYEDPTDSECAEIIAELPEEDQAKLDSL    420
                                                                     :..*

Ea31      PLPAR--TYSE-MLNKCSEGMMQI------------NVRNNFITHFPT---FLQKEQQYR    46
cas9      HLPAGRAAYSRESLTALSDHMLATTDDLHEARKRLFGVDDSWAPPAEAINAPVGNPSVDR    480
           ***   :**.  *.  *: *:              .* :.:     :    : : .  *

Ea31      ILSSTGQLFTYDRTHPLEPTTLVVG----NLTKV--------KLEKLYENNLRDKNKPAR    94
cas9      TLKIVGRYLSAVESMWGTPEVIHVEHVRDGFTSERMADERDKANRRRYNDNQEAMKKIQR    540
           *. .*: ::  .:    * .: *     .:*.           .: *::* .  :*  *

Ea31      TY------------YDDMLVSSGEKCPFCGDIGQ--TKNIDHFLPIAHYPEFSVMPINLV    140
cas9      DYGKEGYISRGDIVRLDALELQGCACLYCGTTIGYHTCQLDHIVPQAG-PGSNNRRGNLV    599
           *              * *  .*  * :**      * ::**::* *  *  .    ***

Ea31      PSCRDCNMGEKGQVFAVDEVHQAIHPYI-DKDIFFREQWVYANFVSGTPGAISFYVE---    196
cas9      AVCERCNRSKSNTPFAVWAQKCGIPHVGVKEAIGRVRGWRKQTPNTSSEDLTRLKKEVIA    659
            *. ** .:..  ***   : .*     .: *   . *   .  :.: .   :  *

Ea31      ----------------CPANWRQEDKHRALH-HFK-----------------LLNIANRY    222
cas9      RLRRTQEDPEIDERSMESVAWMANELHHRIAAAYPETTVMVYRGSITAAARKAAGIDSRI    719
                        . *  :: *: :   :                     .* .*

Ea31      RLEAGKHLSEVITQRNSFV-----KVIRKYSSTATFQQLQSEFIEANLKPIIDLNDFPNY    277
cas9      NLIGEKGRKDRIDRRHHAVDASVVALME-ASVAKTLA------ERSSLRGEQRLTGKEQT    772
          .* . *  .: * :*:  *      ::.  * : *:        .:.*:    *..  :

Ea31      WKRVMYQCLANSED--FFRGI---------------------------------------    296
cas9      WKQYTGSTVGAREHFEMWRGHMLHLTELFNERLAEDKVYVTQNIRLRLSDGNAHTVNPSK    832
            **:   . :.  *.  ::**
```

Figure 3.23. Analysis for the predicted structure of Ea31 protein.

A) The predicted structure of Ea31 protein was modelled and analysed by the Phyre2 server using a template from the crystal structure of the type ii-c cas9 enzyme. The predicted structure was built in several fragments, which showed distinct helices at the N-terminal end followed by two anti-parallel beta strands at the C-terminal end. **B)** A crystal structure of the type ii-c cas9 enzyme from *Actinomyces naeslundii* (hnh endonuclease domain protein, pdb entry: 4OGC) in cyan shows the identical part (purple) with the predicted model of Ea31. **C)** The sequence alignment for Ea31 protein shares a 22% identity with the type ii-c cas9 enzyme.

3.3.2 Ea31 cloning background

The *ea31* gene was cloned into a pMALp2 vector, which contains an ampicillin resistance marker and Lac Operon system, using EcoRI and BamHI restriction sites in the laboratory of Dr. Gary Sharples (University of Durham) (Figure 3.24).

The pMALp2 vector offers a high efficiency strategy for expressing and purifying the target protein after cloning downstream of a *malE* gene, which encodes a maltose binding protein (MBP) tag. The MBP tag is designed to bind strongly to amylose that can help with one step purification. The expression system of MBP in this vector is controlled by a *tac* promoter which is controlled by the *lacI* gene via the encoded lac repressor (Duplay et al., 1984).

The vector pMALp2 provided means to solve the solubility issues of Ea31 protein, following earlier unsuccessful experiments to obtain a soluble product from the wild-type and His-tagged versions. The MBP-tagged protein could be purified straightaway using an MBP affinity column, and then cleaved from MBP using a specific protease Factor Xa cleavage sequence (lle-Glu-Gly-Arg), which lies upstream of the Xmnl site (Duplay et al., 1984) (Figure 3.24).

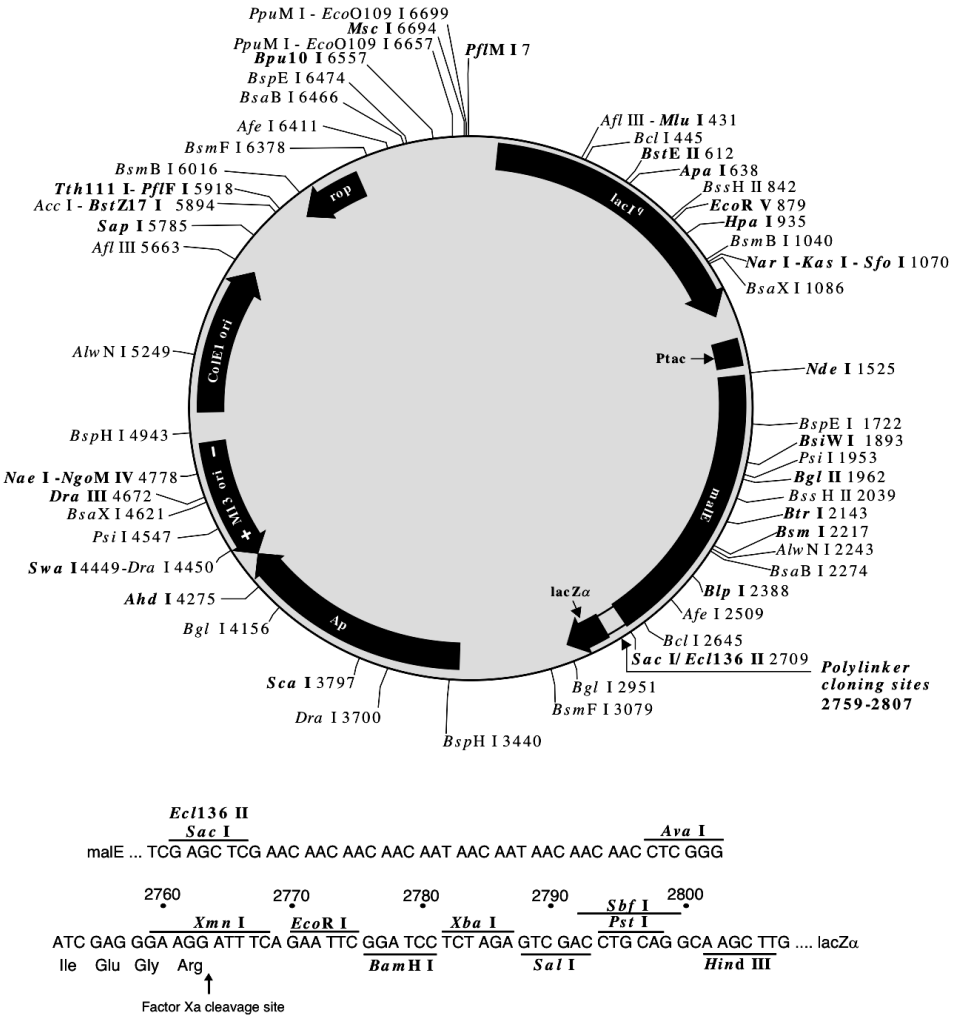


Figure 3.24. pMALp2 vector map.

NinH gene was cloned into a pMALp2 vector containing an ampicillin resistance marker and using EcoRI and BamHI restriction sites to provide the protein with a N-terminal MBP-tag controlled by *tac* and *lac* promoters with optional cleavage at a of Factor Xa site (Dr. Gary Sharples unpublished data).

3.3.3 Predicted secondary structure of Ea31 protein:

The Ea31 predicted secondary structure via the PSIPRED Protein Sequence Analysis Workbench server (http://bioinf.cs.ucl.ac.uk/psipred/) was in agreement with the Ea31 predicted structure via Phyre2 server, and demonstrated that the Ea31 predicted secondary structure likely contains many long disordered regions in the middle of its structure (residues 50-74, 101-144 and 149-177) (Figure 3.25). These regions may contribute to the insolubility problem, since it was found that disordered proteins tend

to be more likely to aggregate compared to stable and compactly folded proteins (Churion and Bondos, 2012).

The MBP protein is a highly soluble protein and plays an important role as a solubilizing agent by having similar properties to chaperone proteins. MBP contains a large hydrophobic region on its surface in the form of cleft, which gives the property of an interaction site to combine with target polypeptides. Additionally, the cleft of the MBP protein has the capability and flexibility to form different conformations to engage with many various proteins. Through fusion to its target protein, MBP induces the folding capability of the protein by blocking the self-association of the protein of interest (Costa et al., 2014).

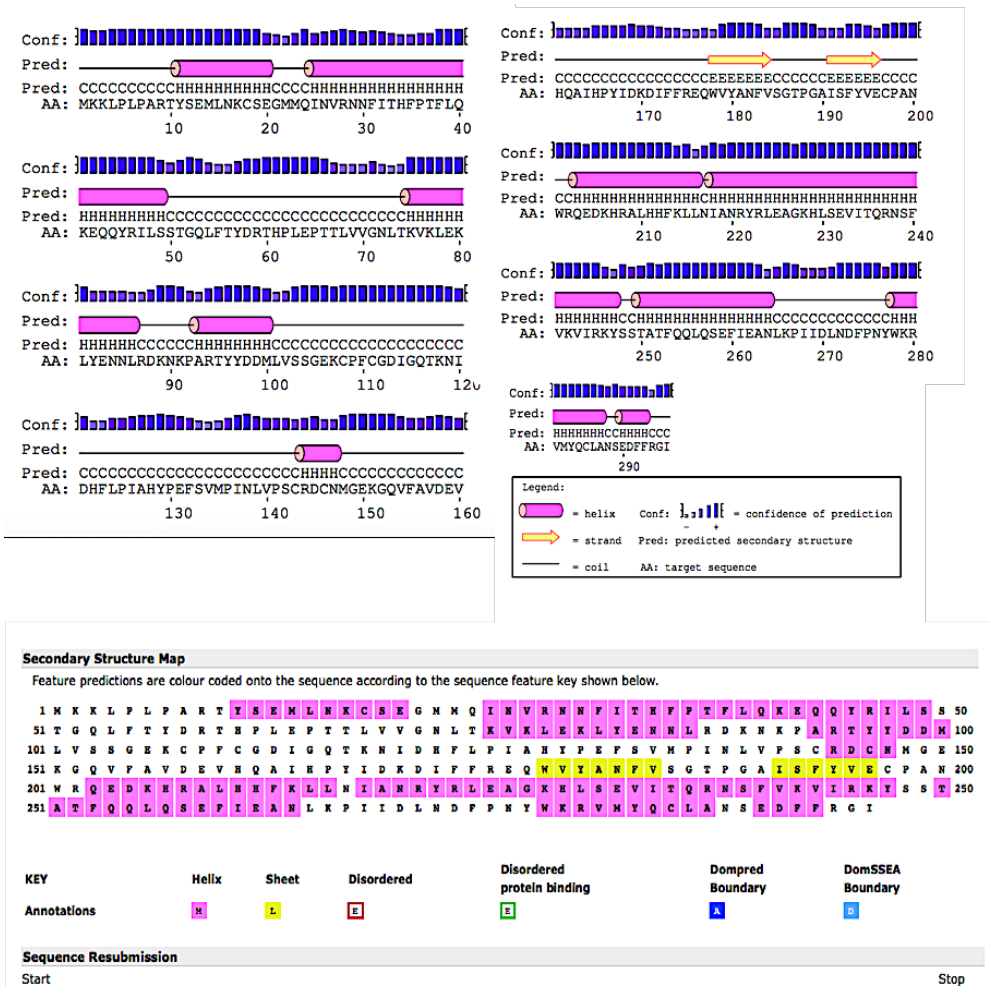


Figure 3.25. Predicted secondary structure.

The predicted secondary structure of the 296 amino acid Ea31 protein sequence via PSIPRED Protein Sequence Analysis server showed the disordered regions at a number of different locations in its structure (residues 50-74, 101-144 and 149-177) that might be the reason for the insolubility problem of the Ea31 protein.

3.3.4 Protein Expression

Conditions had been identified previously in the Sharples laboratory for the expression of the Ea31 protein after growing a 10 ml overnight culture, and the same expression strategy as used for NinH protein was followed to check for a better quantity of the soluble fusion protein. The optimum conditions in 50 ml LB media in a 250 ml flask were found by inoculation with 1 ml overnight culture supplemented by 100 mg/ml ampicillin, followed by induction with 0.5 mM IPTG, 0.2% L-arabinose, at 25 °C and then 18 hours incubation. The samples were checked by SDS-PAGE as described for Rap protein, with the exception that buffer A was 20 mM Tris-HCl pH 7.4, 0.2 M NaCl, and 1 mM EDTA. A large band was found on the gel after induction, between the 66.3 kDa and 97 kDa markers, likely corresponding to the 80 kDa size of the MBP fusion protein. The results revealed a high level of expression of MBP-tagged Ea31 in the supernatant of the cell sample, representing a high level of solubility. Large scale over-expression culture was set up using the same conditions with 3 x 500 ml LB media (section 2.11.2). 5 g of cell pellet was collected 18 hrs after induction and the samples were run on SDS-PAGE to check the expression which confirmed a high level of the soluble expression of MBP-tagged Ea31 protein (Figure 3.26).

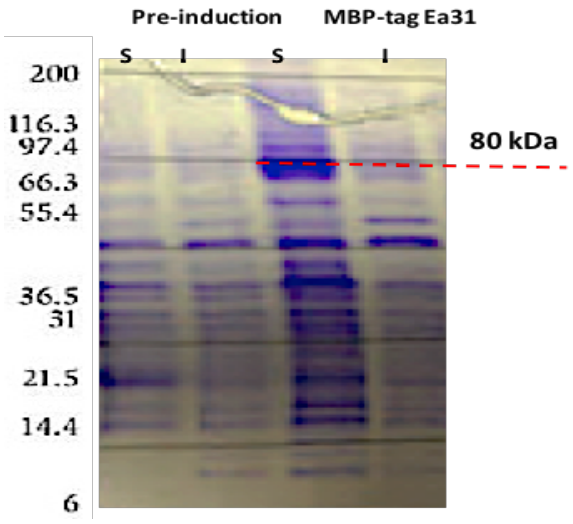


Figure 3.26. Overexpression and solubility for Ea31 protein.

A 12% SDS-PAGE gel shows a high level of 80 kDa soluble expressed MBP-tagged Ea31 compared with the insoluble fraction. The cell paste was resuspended with buffer A (20mM of Tris-HCl pH 7.4, 0.2 M NaCl, and 1 mM EDTA. 15 µg was the amount of protein in the supernatant loaded on the gel. S = soluble, I = Insoluble.

3.3.5 Protein purification

The Expasy protparam server results for the Ea31 protein sequence showed that the exact molecular weight of its 296 amino acids was 34583.6 Daltons or for the MBP-tagged form with 707 residues was 79782.9 Daltons. The predicted isoelectric point (pI) is 8.65 (or 6.22 with MBP-tag), which makes the Ea31 protein more likely to be a positively charged protein by itself (or a negatively charged protein with an MBP-tag) in pH 8 buffer.

The Molar Extinction Coefficient was 37735 $M^{-1}cm^{-1}$ (or 105575 $M^{-1}cm^{-1}$ with MBP-tag) (http://web.expasy.org/protparam/).

The purification process was started with 5 grams of cell pellet from a 1.5 litre culture which was suspended in 50 ml of buffer A (20 mM of Tris-HCl pH 7.4, 0.2 M NaCl, and 1 mM EDTA), and disrupted by sonication in 3 bursts of 20 seconds. The cell free extract of the MBP-Ea31 protein was purified by using an amylose exchange column (MBP column, 5ml). The protein was eluted using 20 ml of buffer B (20 mM of Tris-HCl pH 7.4, 0.2 M NaCl, 1 mM EDTA, and 10 mM maltose). The total amount of the purified fusion protein was 6.5 mg in 7.5 ml volume (0.8 mg/ml). The protein was concentrated to 11.25 mg in 2 ml volume using a Vivaspin unit with 10 kDa MWCO (Figure 3.27). Various crystallization trials were set up in different concentrations for MBP-Ea31 purified protein.

A volume of 250 µl (11 mg/ml) purified protein was used for cleavage of the MBP-tag using 80 µl of Factor Xa protease plud 35 µl $CaCl_2$ (2 mM). The resulting Ea31 native protein was purified by using a Heparin-HP 1 ml column. However, an SDS-PAGE analysis showed that Ea31 purified protein was not fully separated from the MBP-tag by Factor Xa protease (Figure 3.28). Therefore, this project was halted at this stage because of time constraint.

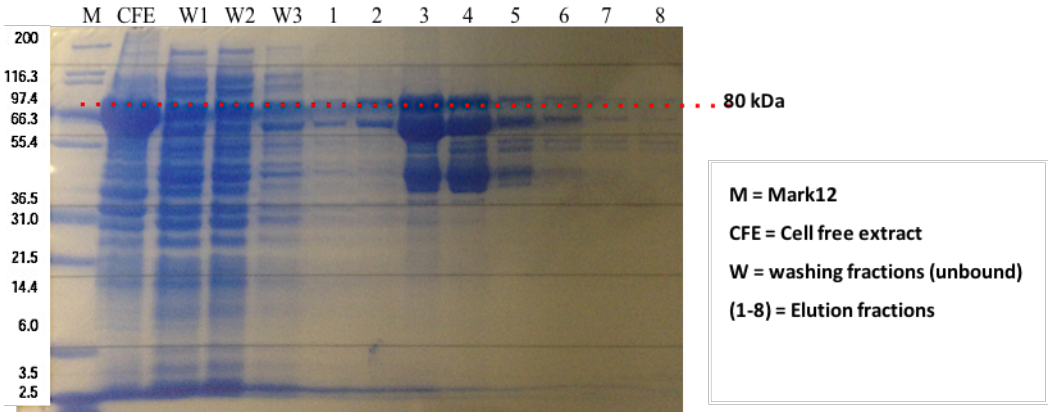


Figure 3.27. MBP purification for MBP-tag Ea31 protein.

SDS gel shows the MBP-Ea31 purified protein after extraction by MBP column, where the MBP-Ea31 purified protein (80 KDa) appears in fractions found in lanes 3-5, which were combined and concentrated to 2 ml volume of 11.25 mg/ml.

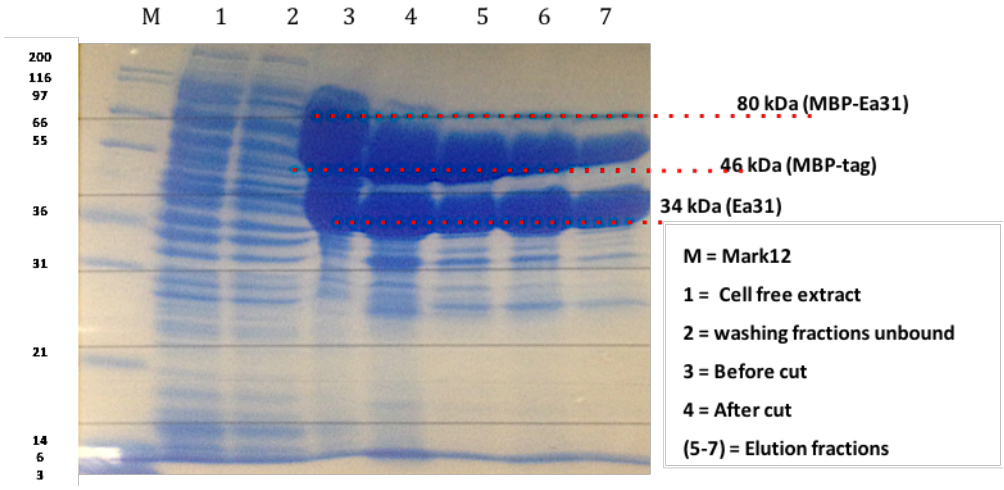


Figure 3.28. Heparin-HP purification for Ea31 protein.

SDS-PAGE gel shows the Ea31 purified protein after extraction by Heparin-HP column, where the Ea31 protein (80 KDa) appears in the fractions found in lanes 4-7. The result showed that the Ea31 purified protein was not fully cleaved by Factor Xa protease and the tagged and cleaved proteins were not separated well by the heparin column.

3.4 Discussion of Rap, NinH and Ea31.

3.4.1 Rap endonuclease

Lambda bacteriophage has had a long history of being a model organism in molecular genetics, and thus has been well-characterized. However, there remain several genes in its genome that still are not entirely understood. In addition to *ea10* described in Chapter 4, we report three additional such genes, *rap*, *ninH,* and *ea31*, that are also found in the tail fibers of lambda phage. While the function of the *rap* gene product has been determined, those of *ninH* and *ea31* remain vague (Rajagopala et al., 2011).

To understand the protein products of these three genes in greater detail, their structures could be determined, for example, by X-ray crystallography. By determining their structures, this would help us in visualizing the molecules, which may lead to clues about their function or interacting partners. This chapter reports the initial stage of the crystallization process, where all three genes have been cloned, expressed, and purified to a reasonable degree, and crystal trials have also been set up.

3.4.1.1 Protein expression level of P3 and P4 truncated Rap protein using pET21

There was a very low level of expression of both P3 and P4 truncated Rap proteins under all conditions and this was not suitable to carry out purification even when cultures were grown in significant volumes of media. It was also unclear why the pET21b vector led to low-level expression of the proteins, since its expression system works in a similar way to the pET22b vector, which resulted in high-level expression of full length Rap protein previously. It may be possible that the alteration in the length of protein by truncating the N-terminal part may have affected its ability to be well expressed due to it becoming a toxic protein which kills the host cell or a dangerous product and thus the host cell eliminates it quickly (fast degraded protein), or perhaps the pET21a vector had gained a mutation because its sequence was not checked via sequencing analysis. Moreover, the low level expression or the quick degradation of product can also happen when using a cell line unable to express and cope with this kind of protein in comparison with BL21 AI cell line, which has been designed for production of proteins that might be toxic to the other competent BL21 cell strains (Laitinen et al., 2005). Therefore, the *rap* P3 and P4

genes were sub-cloned into another vector, pJONEX4, for comparison of expression (Section 3.1.2.5.1.2).

3.4.1.2 Crystallization of Rap proteins

There are a number of possible reasons why the Rap endonuclease protein may not have crystallized:

A) The low level of purity due to fast degradation of the target protein, which was obvious in the results of the SDS-PAGE and mass-spectrometry analyses, may have occurred due to the presence of protease contaminant with the sample. The presence of small fragments of cleaved target protein in the mixture can cause a serious problem by poisoning a crystal lattice during crystal growth.

B) Crystal trials of the Rap protein were performed at various concentrations ranging from 8-26 mg/ml. It is possible that the concentration at which Rap protein crystallizes falls outside the range tested. Proteins have been known to crystallize below and above this range.

C) The ratio of the protein to the precipitant solution was carried out using a 1:1 ratio. Although this is the default ratio for which crystal trials are set up, it could be that this may in fact not be the optimum ratio for driving crystallization of the Rap protein. Therefore, using different ratios of protein to precipitant could be tested.

D) Disordered or flexible regions in the protein can also affect the crystallization process. Looking at the predicted secondary structure of Rap and also the Phyre2 predicted structures, Rap endonuclease appears to have a large loop region in the middle, as well as a large possible disordered region in its N-terminus. These regions can interfere with crystal packing, thereby making the protein difficult to crystallize. Experiments were carried out with N-terminal truncated versions of the Rap protein, which also did not result in successful crystallization. However, these do not take into account the large possible disordered region in the middle of the protein.

3.4.1.3 Possible solutions for Rap protein crystallization.

The Rap endonuclease is a DNA binding protein. Therefore, the presence of a suitable piece of DNA with the protein could be helpful in crystallization. Choosing a specific DNA fragment can be challenging due to the difficulty of determining the appropriate form of DNA (single-stranded, double-double-stranded with or without over-hanging ends, branches junction). However, the Rap protein is more likely to bind to the duplex DNA substrate such as is found in Holiday junction DNA because of its predicted role as a DNA resolvase (Sharples et al., 2004). This cleavage property of Rap should be inhibited before crystallization to allow binding to DNA without cutting it. This could be achieved by mutating the active site although sometimes that can risk the protein state because of a change in the solubility and stability. Alternatively, the resolvase enzyme could be inactivated by eliminating the key cofactors through addition of either competing inactive species such as calcium ions or EDTA to remove the metal ions required. However, this approach risks removing the metal ions required for the protein to bind to DNA or other proteins.

Based on previous study, the Rap protein activity to bind and resolve the DNA junctions is highly increased in the presence of magnesium or manganese ions (Sharples et al., 2004). Therefore, adding a calculated amount of these metal ions to the protein solution might help in crystallization trials.

3.4.1.4 Possible solution to solve Rap endonuclease structure

3.4.1.4.1 Relationship between Rap endonuclease and other endonucleases

There is no known homologous for Rap endonuclease protein. However, sequence homology for a small section of Rap with T4 endonuclease VII and Gmet endonuclease has been proposed (Prof Gary Sharples, personal communication) even though the sequence alignment of Rap with either of these two proteins in NCBI Blastp shows no homology for the rest of the protein. This small section corresponds to the active site of Rap required for DNA binding and cleavage of Holliday junctions (Prof Gary Sharples, unpublished data), which aligns with the active sites of phage T4 endonuclease VII and *Geobacter metallireducens* HNH nicking endonuclease (Gmet) (Figure 3.29).

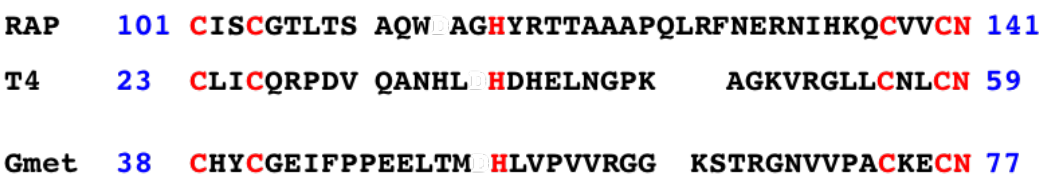


Figure 3.29. Sequence alignment for Rap protein.

Conserved residues are found between the lambda Rap, phage T4 endonuclease VII (T4) and *Geobacter metallireducens* HNH nicking endonuclease (Gmet). The conserved residues include 4 Cysteines, one Histidine and one Asparagine (red).

3.4.1.4.2 Superposition for the predicted structure of Rap Protein

The structural superposition between the phage T4 endo VII (pdb: 2QNC) and Gmet (pdb: 4H9D) endonucleases showed a noticeable alignment in the conserved part, which binds to a DNA junction, with RMSD score of 5.49/100, while the rest of structures were not aligned (Figure 3.30 A, B). This result suggests that the Rap structure might also have a similar structure in this conserved region to T4 endo VII and Gmet, but that the rest of the Rap protein would display structural differences. This presumption was supported further by superposition of the Rap Phyre2 predicted structure, which was based on the gve2 hnh2 endonuclease from the deep-sea thermophilic bacteriophage Geobacillus virus E2 (pdb: 5H0O), with the T4 endo VII and Gmet structures which revealed the compatibility in the conserved region (1.84/40 RMSD) (Figure 3.30 C). However, it is just a small part of the Rap protein (40 residues out of 203 residues). Therefore, it is unlikely that the T4 endo VII and Gmet structures would make a good model for a molecular replacement to solve the Rap structure in case of crystal production. The best strategy for structure solution is by experimental phasing using Seleno-Methionine incorporation, which is the most reliable way, or heavy atom ion soaks such as mercury, iodine, and tantalum bromide solutions.

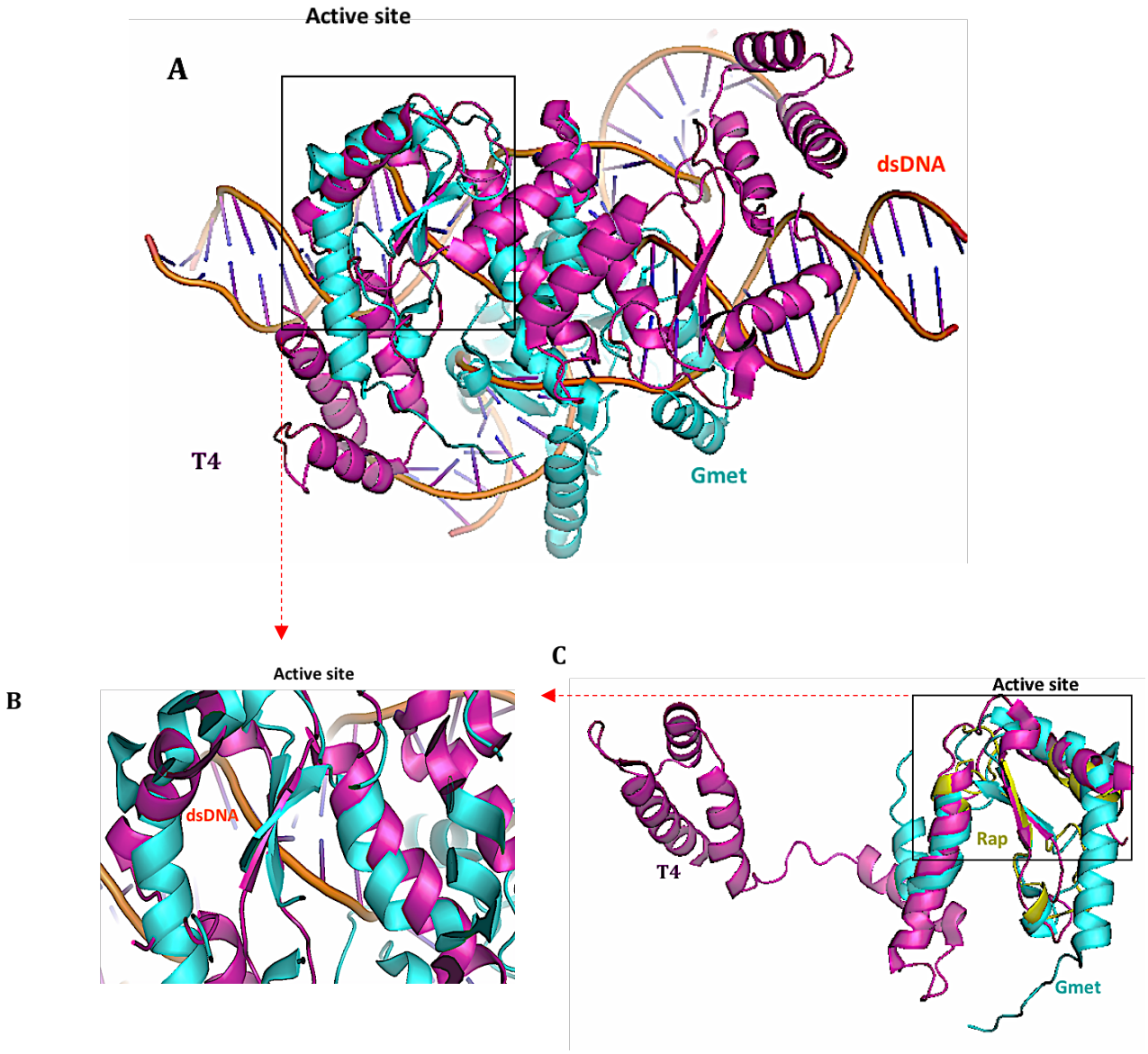


Figure 3.30. Structure superposition of Rap protein.

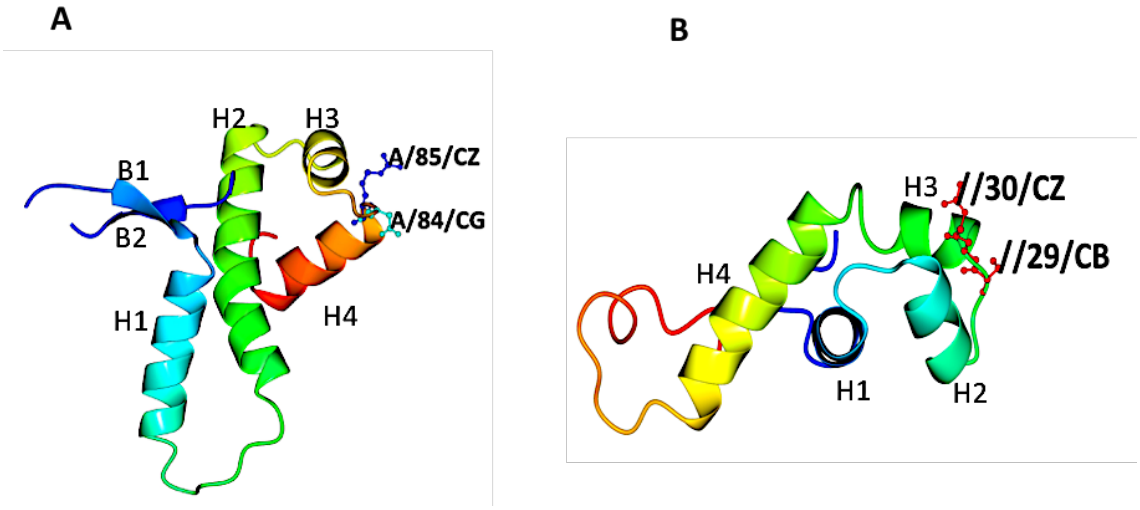
A) Superposition for the full T4 endo VII (magenta) and Gmet (cyan) structures (5.49/100 RMSD). The figure includes two subunits of T4 protein and its bound observed the DNA Holliday junction (orange) with three subunits of Gmet. **B)** The conserved region of the superposed structure (T4, Gmet). The figure also shows the aligned region, which refers to the active sites of these proteins (Helix turn Helix and antiparallel beta strands). **C)** Superposition of the Phyre2 predicted structure of Rap protein (yellow) with one subunit of T4 and Gmet structures.

3.4.2 NinH

All the attempts to express a soluble product of a wild-type version of NinH protein were unsuccessful. However, NinH protein was cloned in pET14b with an N-Terminal 6x histidine tag with additional thrombin site to improve the expression and solubility, and also to facilitate purifying the protein with a possibility of His-tag removal in future**.**

Although the requirements for better production of NinH protein needed either a His-tagged or non-tagged product (cleaved His-tag product), and different strategies were used for setting up crystallization trials, no crystal has formed. As discussed for Rap protein (section 3.4.1), there are many sources that could affect the protein crystallization. For example, the presence of disordered or flexible regions can affect crystal packing and the predicted secondary structure of NinH (Figure 3.20). has perhaps loops or unstructured regions at the N-terminal and the C-terminal ends. Therefore, the attempt to shorten these regions may resolve the crystallization problem. Unsuitable ranges of protein concentration or ratios between the protein and the precipitant solution could also be a cause for not crystallizing the protein.

The predicted NinH structural model via the Phyre2 server is formed from four distinct helices at the N-terminus end (residues 12-30) (Figure 3.31 A, B). The Phyre2 server gave a 34% sequence identity of 67 residues of NinH with a DNA/RNA-binding FIS-like family member with a 92% confidence value. Sequence analysis revealed 42% identity of NinH sequence (residue 6-35) with the *E. coli* Fis DNA binding region (Figure 3.30C). This region forms 3 helices, which is described as comprising a helix plus a helix-turn-helix motif at the C-terminal part of Fis structure (residue 60-90) (Stella et al., 2010). While, this 3-helices bundle is present at the N-terminal region of NinH sequence.



C

Score	Expect	Method	Identities	Positives	Gaps
26.6 bits(57)	6e-06	Compositional matrix adjust.	13/31(42%)	21/31(67%)	1/31(3%)

```
NinH    6   KTIPDMLVE-TYGNQTEVARRLKCSRGTVRK 35

            + + DM+++ T GNQT   A   +   +RGT+RK

Fis    60   QPLLDMVMQYTLGNQTRAALMMGINRGTLRK  90
```

Figure 3.31. Structure analysis for Lambda NinH.

the image shows a helix-turn-helix motif (H3 and H4) from Fis structure **(A)** that shares a structural homology with Lambda NinH protein (H2 and H3) **(B**), and the same conserved residues (Asp84 and Arg85), which are in contact with DNA, but in different location in NinH sequence (Ser29 and Arg30). **C)** Sequence analysis refers to 42% identity between NinH and Fis proteins and the location of the conserved residues in red.

Since NinH is a DNA bending protein and shares structural features with Fis protein (Esposito and Gerard, 2003), it may have some similar functions to Fis protein. For instance, being involved in site-specific recombination or gene regulation, possibly of those genes in within the *ninR* region (Gittens, W. *et al*, 2015). These properties might suggest that adding an appropriate metal ion or piece of DNA substrate (single or double stranded) may improve the state of the protein for crystallization.

3.4.3 Ea31

Ea31 was found to be very difficult to express as a soluble protein. Many different techniques and methods were applied, following a variety of strategies, and these included, producing His-tagged and non-tagged versions, using different expression system and vectors (pET14b, pT7-7), a diverse set of conditions of temperature, incubation and concentration of inducer (Prof Gary Sharples, Unpublished data). Analysis of the predicted secondary structure of the Ea31 protein revealed a number of potentially disordered areas within the structure which might be the possible causes for the insolubility through aggregation of the protein (Churion and Bondos, 2012).

One of the solutions for solubility issues is to use a chaperone to improve folding and hence the solubility of unfolded proteins. The MBP protein is considered as a solubility enhancer that acts efficiently to increase the expression and promote the solubility of overexpressed proteins of interest. The MBP protein has a chaperone behavior through its activity as a translocator for maltose and maltodextrins (Fox et al., 2001). In addition, it is a natural binding protein, which contains a large hydrophobic site which might play a very important role in interacting with a target protein and preventing the aggregation. MBP is also able to form different conformations to bind to various proteins.

The MBP system was found to work well in producing a fusion product with Ea31 (MBP-Ea31) that was soluble, and therefore could be easily purified on an amylose column. A reasonable yield of purified MBP-Ea31 was obtained, at around 11 mg of purified protein from 5 g of cell paste. This was enough to set up at least some crystal trial experiments. However, the large size of the MBP tag and the flexibility of the linker to the target protein can affect negatively on crystallization. Crystal formation has not been observed for MBP-Ea31 samples. An attempt was made to cleave the MBP-tag from the MBP-Ea31 protein. However, it was found that the Ea31 protein could not be reasonably cleaved after the addition of Factor Xa protease and passage down a heparin column. It is unclear about poor activity of the cleavage enzyme, but possible reasons include using an unappropriated ratio of the protease enzyme to protein to cofactor $CaCl_2$, inappropriate contents of the buffer solution, wrong incubation time and temperature.

4 Chapter 4. Ea10 protein results

Summary

No previous crystallographic studies have been made of the Ea10 protein structure. This chapter describes the protein purification, crystallization and the structure determination. The Ea10 protein structure was solved to 2.4 Å, and it was found to form a dimeric structure. As yet, no homologous structures have been found for Ea10 to help predict its main function. Unpublished data from the laboratory of Dr. Gary Sharples at the University of Durham had suggested that it is a DNA binding protein. In contrast, structural comparisons for Ea10 via the Dali Server suggest a notable similarity with part of the Q-beta replicase core complex, indicating an association with RNA instead. Experimental results are described about the affinity of the protein for DNA.

4.1 Predicted structure for Ea10 protein

Prior to determining a structure for Ea10, the sequence of Ea10 protein was run in intensive mode on the Phyre2 server to obtain a set of predicted structures. The chosen predicted structure was selected to model Ea10 protein based on the program's maximum confidence value, the percentage sequence identity and sequence alignment coverage. The model showed three anti-parallel beta strands in the middle and three helices including a long helix (H3) (Figure 4.1). However, the structure was modelled *ab initio,* which is generally highly unreliable, and the protein sequence recorded a very low score of confidence (21%) using the hemagglutinin structure (pdb: c5c0sA) as a template. This predicted structure allowed an interesting comparison between the true structure that was obtained later via X-ray crystallography with this putative structure.

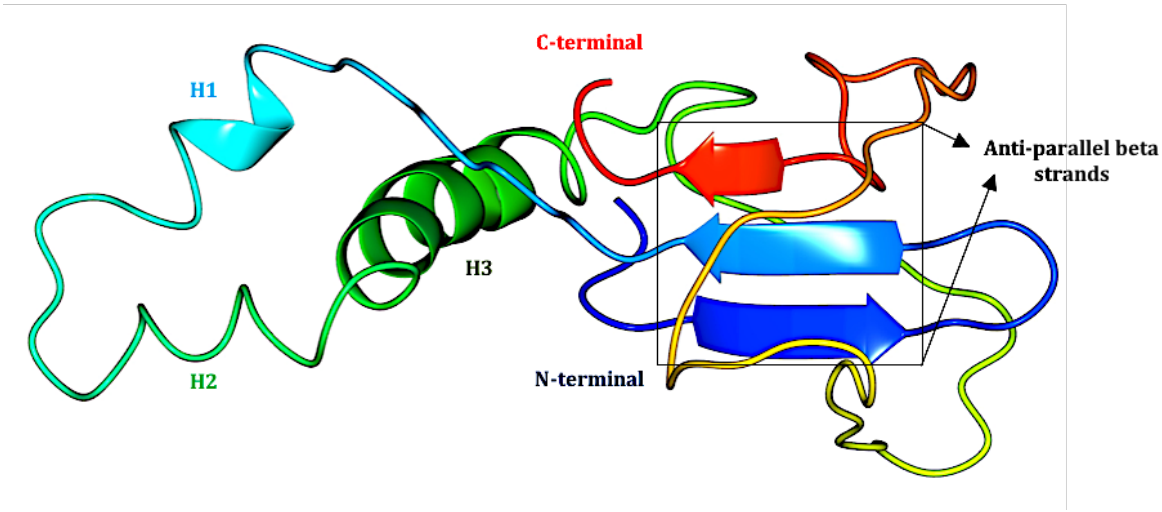


Figure 4.1. Predicted structure for Ea10 protein.

The predicted structure of Ea10 was generated by *ab initio* modelling via the Phyre2 server using the intensive mode option. The model shows three anti parallel beta strands, and three alpha helices (H1-H3).

4.2 Studying the secondary structure of Ea10

The Ea10 protein sequence was also examined using the PSIPRED Protein Sequence (http://bioinf.cs.ucl.ac.uk/psipred/) to predict the secondary structure of the Ea10 protein and possible disordered regions. The result showed two beta strands in the first 25 residues of the N-terminus end, and also two beta strands in the final 25 residues of

the C-terminus. Three of these strands were shown as anti-parallel beta strands in the Phyre2 predicted structure. In the middle of the structure (27^{th}-70^{th}) residues, the secondary structure also showed two small α helices followed by a long α helix, which were quite compatible with the Phyre2 predicted structure. There was no significant sign of any disordered protein structure (Figure 4.2).

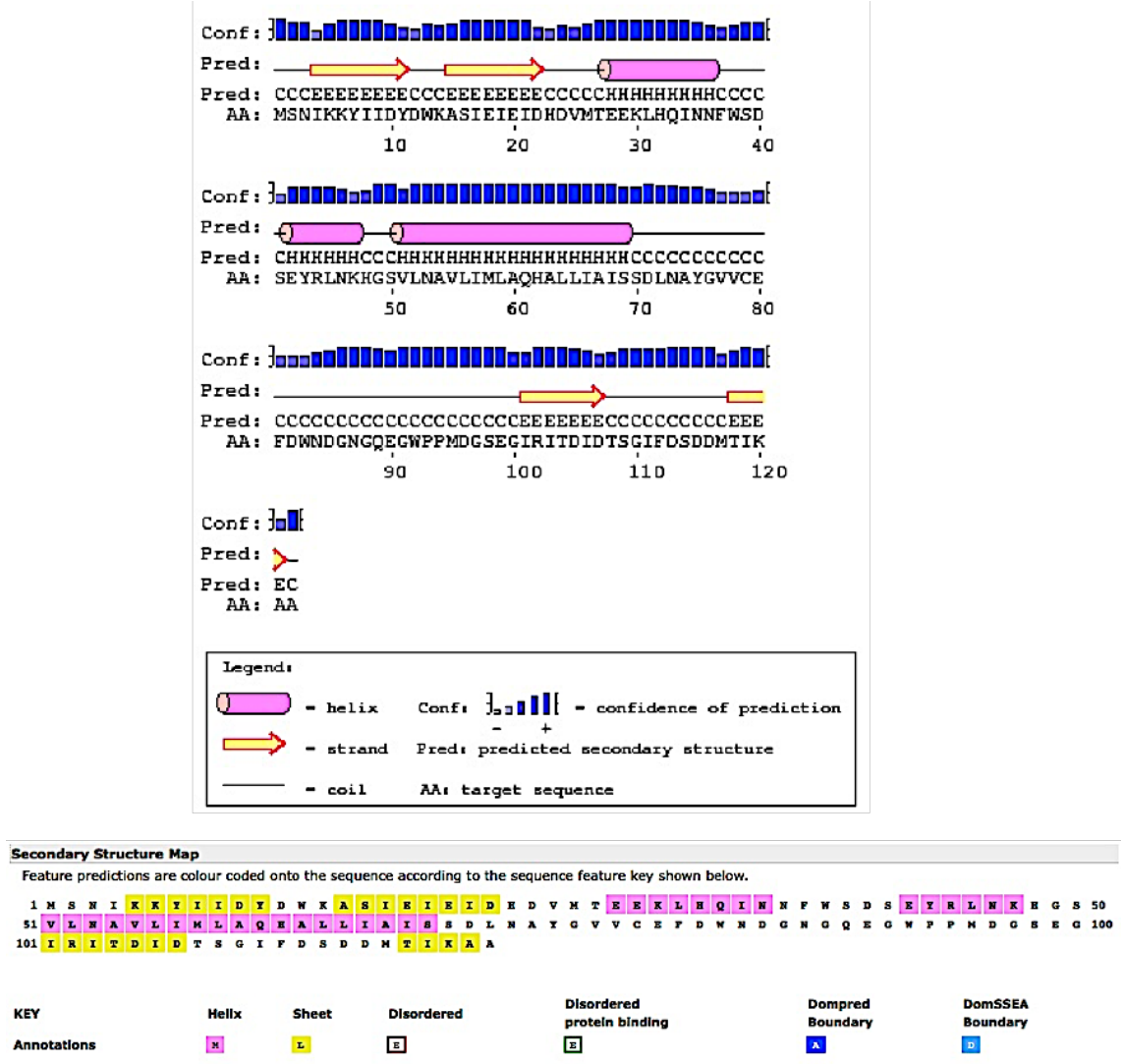


Figure 4.2. Predicted secondary structure of Ea10 protein.

The predicted secondary structure of Ea10 protein sequence via the PSIPRED Protein Sequence Analysis server showed two beta strands in each end with three distinct α helices in the middle of the structure (27^{th}-70^{th} residue). No disordered regions were detected.

4.3 Ea10 cloning background

The *ea10* gene was cloned into a pET16b vector (Figure 4.3) containing an ampicillin resistance marker and *Lac* operon system by collaborators at Durham university (G. Sharples).

NcoI and BamHI were used as restriction sites. The pET16b vector has a very developed system for cloning and expression of target genes controlled by the bacteriophage T7 system of transcription and translation and is under *Lac* control.

The recombinant plasmid was transformed into the *E. coli* host cells (BL21 AI) using a heat shock method. This protein expression cell line contains a chromosomal copy of the T7 RNA polymerase gene in the *ara*B locus of the *ara*BAD operon, thus the T7RNAP expression is regulated by the arabinose-inducible *ara*BAD promoter. This property makes the BL21 AI cell very useful for expression of genes that might be toxic to other BL21 strains (Laitinen et al., 2005). This cell also has a *LacUV5* promoter and *LacI* gene because of a lysogenized bacteriophage DE3 system in the genomic DNA of this cell. The expression was induced by the addition of IPTG (Isopropyl-ß-D-thiogalacopyranoside) to a growing culture for inducing the LacUV5 promoter to direct transcription of the T7 RNA polymerase, which in turn transcribes the target gene in the plasmid (Studier and Moffatt, 1986).

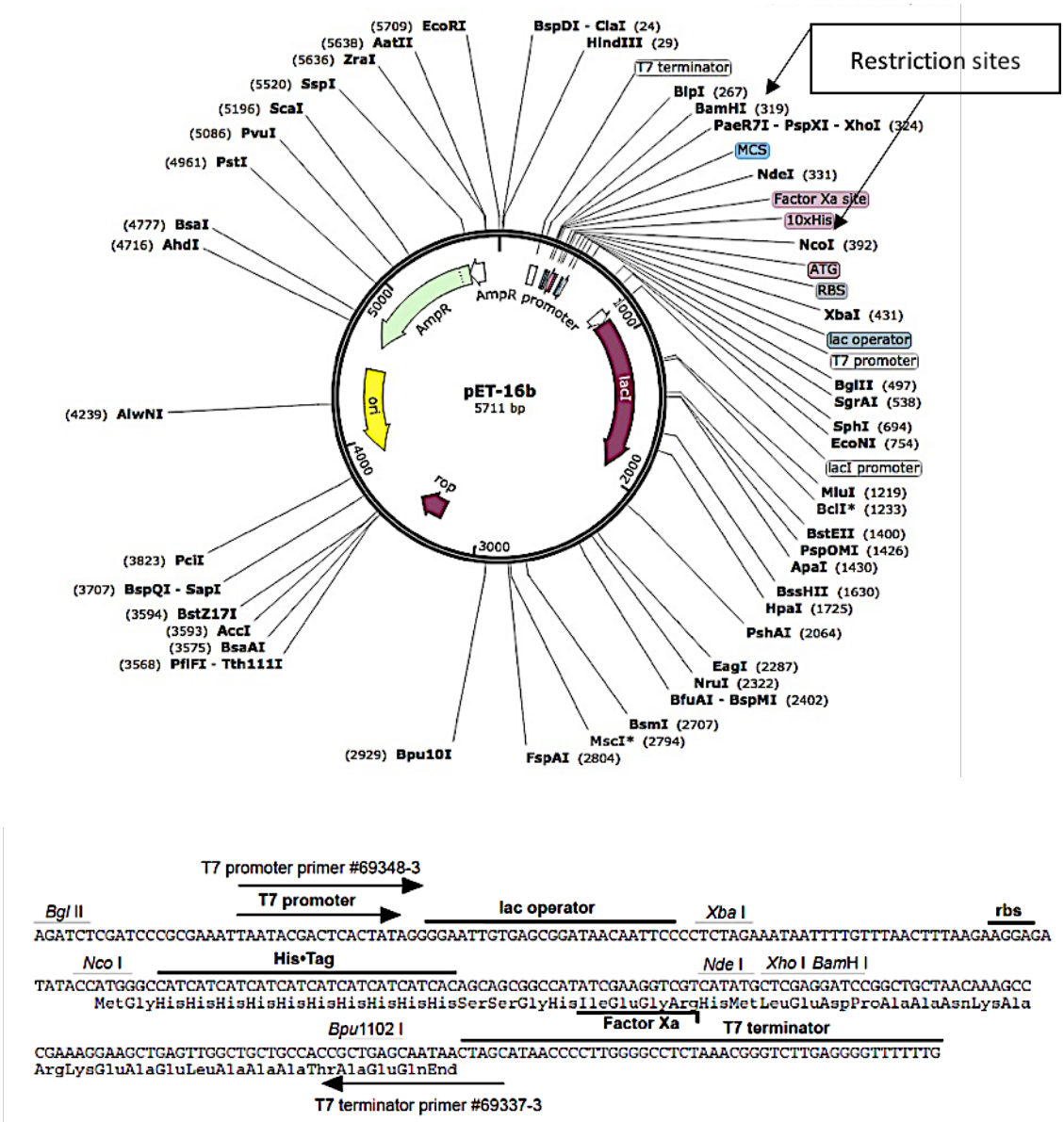


Figure 4.3. pET16b plasmid map.

Ea10 lambda gene was cloned into pET-16b vector using NcoI and BamHI restriction sites to remove the possible His-tag. This vector contains an ampicillin resistance marker and *Lac* Operon and was transformed by collaborators at Durham University for overexpression in the *E. coli* cell line BL21-AI. Image form Snap gene.

4.4 Colony screening

The BL21 AI cells containing the plasmid construct were plated out on LB agar plates supplemented with 50 µg/mL ampicillin, to obtain single colonies. Colonies were screened by PCR using the original primers for specific gene amplification, and then the plasmids were extracted from the colonies using the Qiagen Plasmid Miniprep Kit. 10 colonies were selected to perform sequencing to ensure that no mutation occurred

within the *ea10* gene sequence. Sequencing of the construct was carried out by the GATC sequencing company (Section 2.10). Clear sequencing reads were obtained for four colonies. Sequence alignment with the standard λ phage *ea10* sequence in NCBI GenBank came up with 100% identity using Blastn and Blastp.

4.5 Protein Expression

4.5.1 Small Scale Expression and Solubility

A confirmed *ea10* gene containing colony was selected for carrying out small-scale expression studies. Induction of the Ea10 protein from its pET16b vector was performed by the addition of IPTG (Section 2.11.1). The media was supplemented with 0.2% L-Arabinose to induce the expression of T7 RNA polymerase. Varying concentrations of IPTG (0.1, 0.5, 1, and 2 mM) and temperatures (16, 25, 30, and 37 ˚C) were tested to compare the levels of Ea10 production. An expression analysis on 12% acrylamide SDS-PAGE was used on 1 mL samples taken before induction, and at 3 hours, 6 hours, and 18 hours after induction. The small-scale expression of Ea10 protein was optimized at 37 ˚C, using 0.5 mM IPTG and 0.2% L-arabinose, and for 6 hours incubation after induction. A large band was found on the gel after induction at a molecular weight around 14 kDa, which was equivalent to the right size for Ea10 (13.8 KDa). This band was not found in a pre-induction sample, revealing a high level of expression of Ea10 protein. The results indicated a high level of solubility due to the presence of the majority of this protein in the supernatant of the cell sample (Figure 4.4).

4.5.2 Large Scale expression

The addition of 0.5 mM IPTG at an OD_{600} of 0.6 and 6 hrs of further growth at 37 ˚C was found to be ideal for over-expression of the Ea10 protein with a high level of solubility. This was scaled-up to 3 L of culture using 500 ml of LB broth in 2 L (baffled flasks). Baffled flasks were found to be better than standard flasks due to creating more aeration in the growth media. The media was supplemented with 0.2% L-Arabinose to induce the expression of T7 RNA polymerase. Samples were collected before induction and 6 hrs after induction, and then analysed on SDS-PAGE (Figure 4.4). All cells were harvested 6 hrs after induction (Section 2.11.2). A total of 6 g of cell pellet was collected from 3 L of

growth media. Equivalent levels of expression of Ea10 protein were obtained on large scale over-expression comparing with small scale expression with a high level of solubility, which looked suitable to do purification (Figure 4.4).

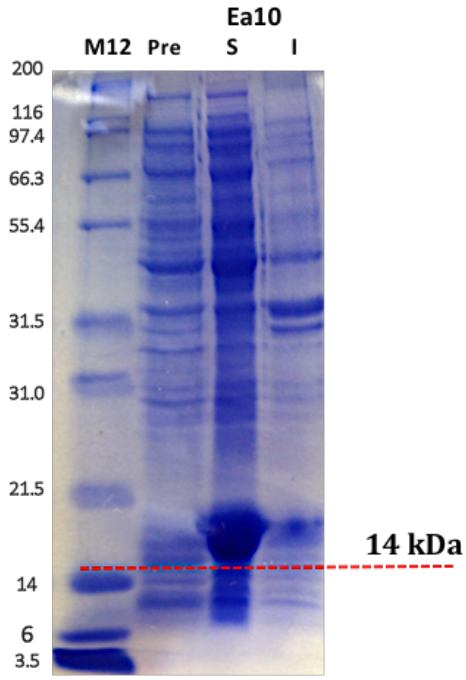


Figure 4.4. Overexpression and solubility for Ea10 protein.

12 % acrylamide SDS-PAGE showed a significant expression level of Ea10 soluble protein, which is around 14 kDa measured by Mark12 ladder, comparing with the sample in the pre-induction lane. 15 µg was the final amount of the protein in the supernatant loaded on SDS gel. The protein was expressed at 37 ˚C, 0.2% L-Arabinose, with a 6 hrs incubation. The cells were induced with 0.5 mM IPTG after an OD_{600nm} of 0.6 was reached. M12= Mark12 Pre= Pre-induction S= soluble, I= Insoluble

4.6 Protein purification

The Expasy ProtParam server results for the Ea10 protein sequence showed that the exact molecular weight of its 122 amino acids was 13780.4 Dalton, and the predicted isoelectric point (pI) is 4.28, which makes the Ea10 protein more likely to be negatively charged in pH8 buffer. The Molar Extinction Coefficient was 27960 $M^{-1} cm^{-1}$. (http://web.expasy.org/protparam/).

The initial purification stage for Ea10 protein was accomplished by re-suspending 4 g of Ea10 cell paste in lysis buffer (Section 2.14). Then it was loaded on a 5ml DEAE Sepharose column (GE healthcare) (Section 2.14.1).

The chromatogram result of this purification step revealed that the protein was eluted in fractions 26-30 with a high level of contaminant proteins based on a 12% acrylamide SDS-PAGE analysis (Figure 4.5).

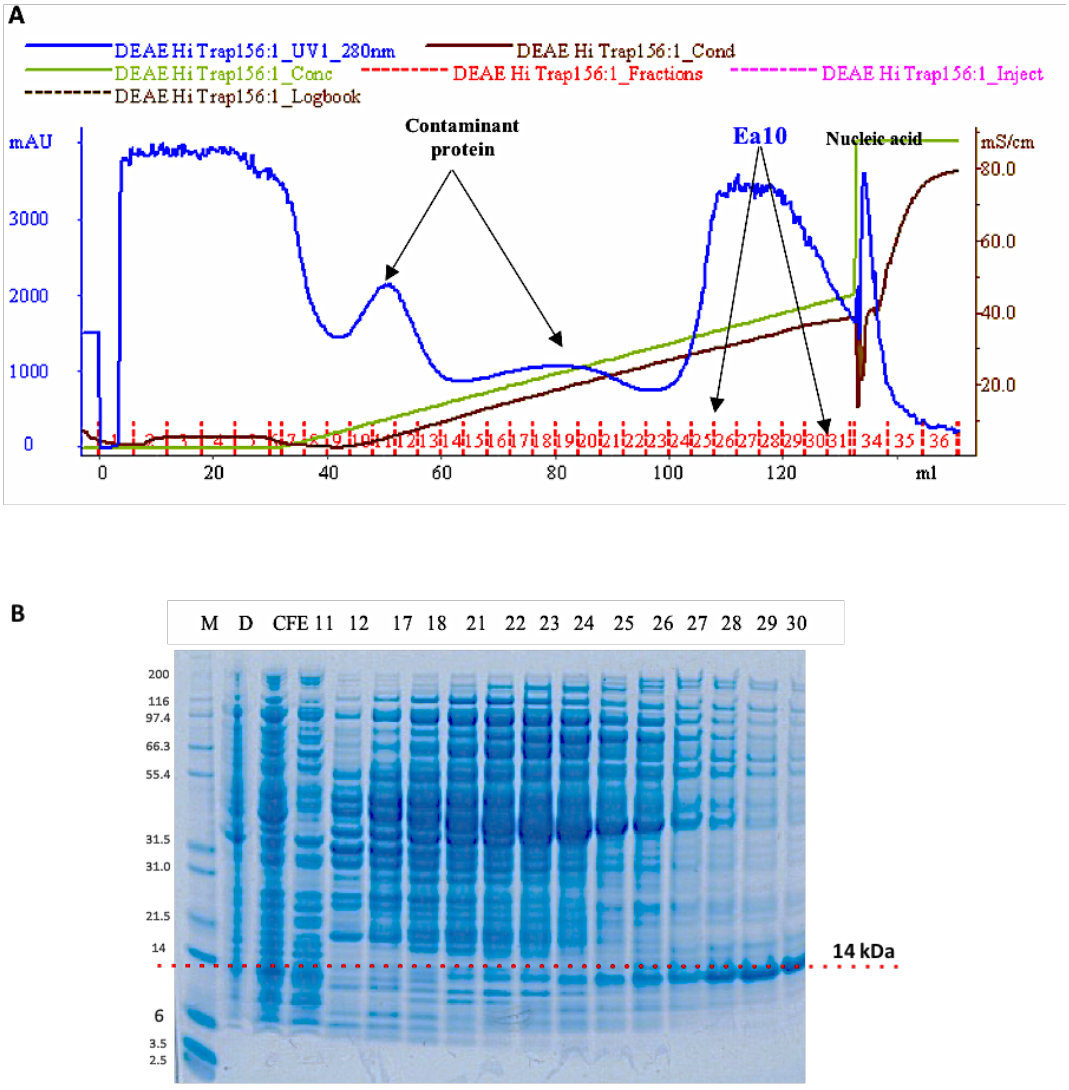


Figure 4.5. DEAE Sepharose purification for Ea10 protein.

A) Chromatogram of the first purification stage result of the Ea10 protein via 5ml DEAE Sepharose column. The blue line represents the UV absorption at 280 nm and the brown line refers to the conductivity. **B)** The results showed that the Ea10 protein eluted in fractions 26-30 as revealed by a 12% acrylamide SDS gel, and these were pooled and used for the next purification step.

A second purification step was carried out using an ion exchange column (Resource Q) after merging the protein fractions (20 ml, 1.8 mg/ml) to increase the purity of the target protein (Section 2.14.2). The protein was eluted using a gradient of 100 ml of buffer containing from 0.1 to 1.0 M NaCl. The chromatogram showed that two peaks were formed; the Ea10 protein was eluted in the first peak (fraction 20-22) based on the results of SDS-PAGE analysis (Figure 4.6).

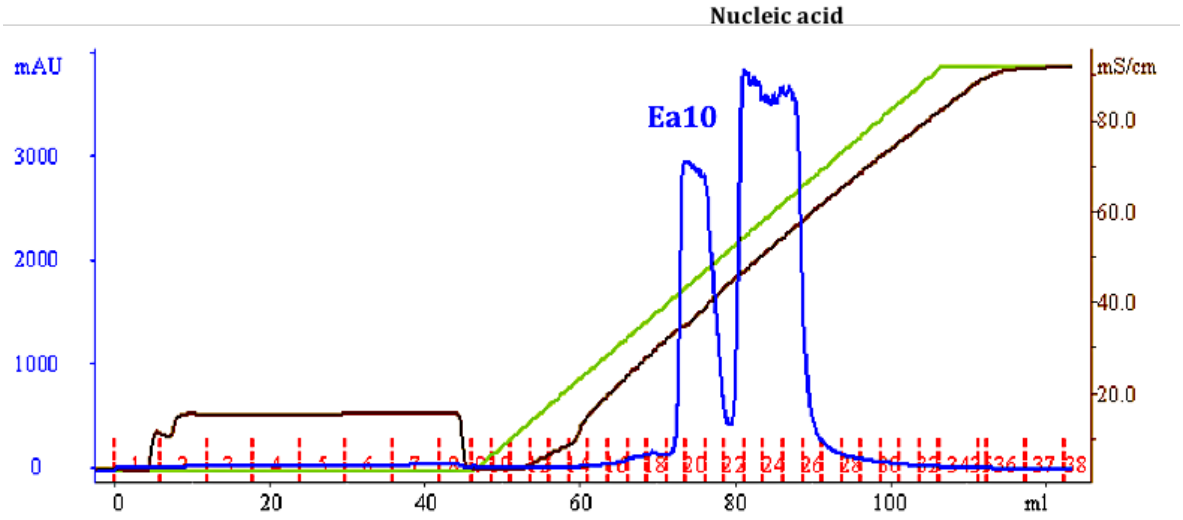


Figure 4.6. Resource Q purification for Ea10 protein.

Chromatogram from Ea10 sample on a 6 ml Resource Q column. The blue line represents the UV absorption at 280 nm, the green line represents the salt ingredient, and the brown line refers to the conductivity. The results showed that the Ea10 protein was eluted in fractions 20-22 as revealed by a SDS-PAGE gel (see figure 4.7), which were combined and concentrated for the next purification step.

These Ea10 protein fractions were combined together and concentrated so that 1-2 ml volumes then were applied on a Superdex^{TM} 200 gel filtration column for better purity and to provide an analysis of the oligomeric state of the protein in solution.

The pooled fractions (1 ml, 23mg/ml) of Ea10 protein were applied to a 1.6x60 cm Superdex-200 gel filtration column (GE healthcare) equilibrated with 50 mM Tris-HCl pH 8.0, 0.5 M NaCl at a flow rate of 1 ml/min. Fractions of 2 ml volume were collected and the protein concentration for each was determined using the Bradford assay. The purification was confirmed using SDS-PAGE with samples after each purification step (Figure 4.7).

The result suggests that the Ea10 protein may be forming a dimeric assembly based on the chromatogram calculations for the highest resulting peak, using the following equation: Kav = (Ve-41)/75, where Ve = the eluted volume, 41 ml = Vo = void volume, 75 ml = V_t = total column volume, and Kav represents the elution volume function of a molecule for the column. The Ve in this case for the highest peak on the chromatogram is 83.7 ml, giving a Kav of around 0.57. From the Kav, the molecular weight of the protein was then estimated by comparison to protein standards plotted on a logarithmic scale. A molecular weight of around 25.1 kDa was calculated for the peak, close to the expected 27 kDa size of a dimeric state of Ea10 protein. The fractions (9-13) with more than 20% of the maximum protein concentration were combined (10 ml, 2.5 mg/ml) and concentrated for use in crystallization trials. The remaining sample was stored at 4 $^{\circ}$C (up to 1 week) or -70°C (for long-term storage) until needed.

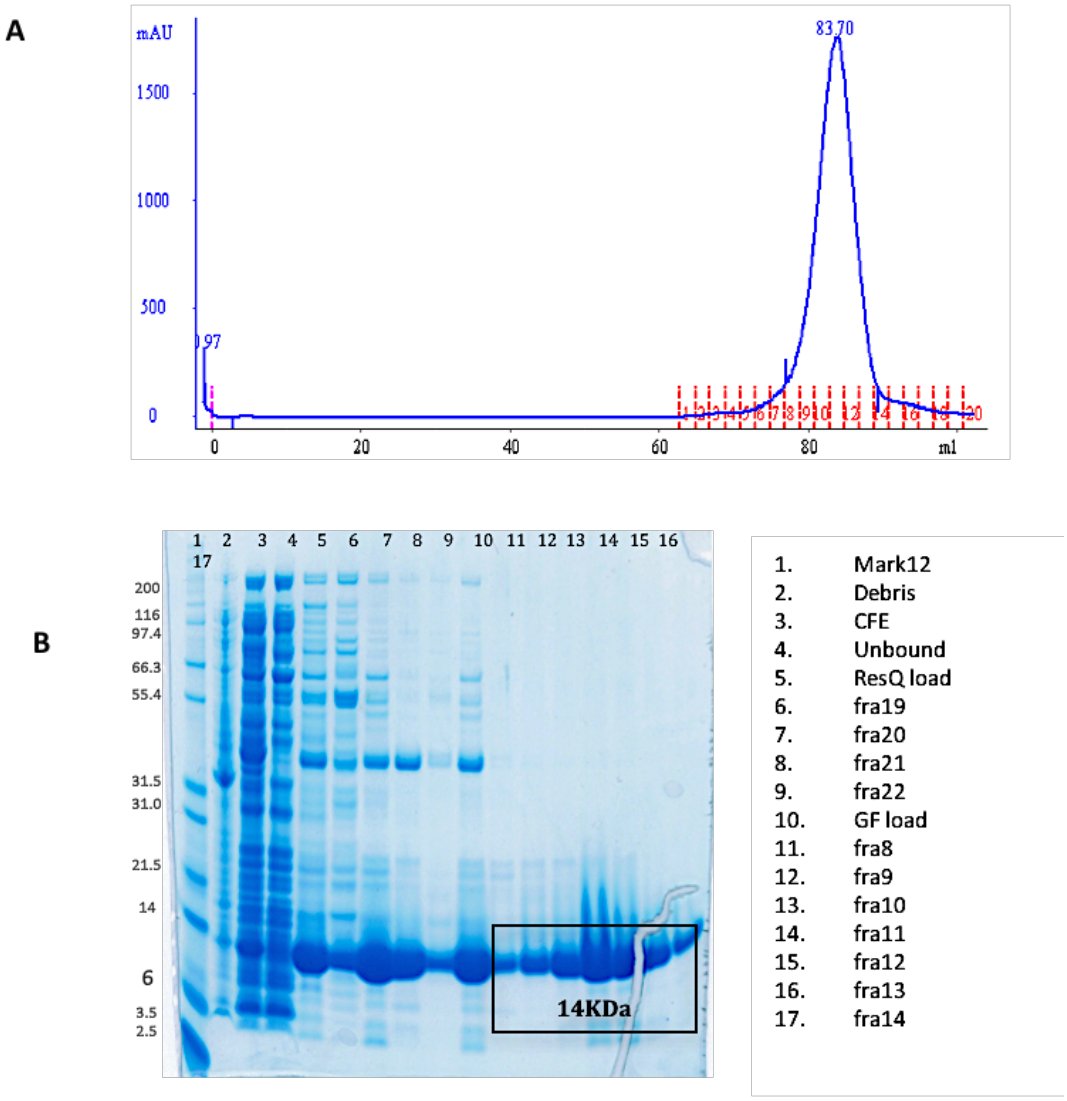


Figure 4.7. Gel-filtration purification of Ea10 protein sample.

A) Chromatogram of Ea10 protein eluted through the final gel filtration Superdex-200ml column. The blue line represents the UV absorption at 280 nm. The fractions 8-14 had a peak volume of 83.7 ml, which suggested formation of a dimer based on the chromatogram calculations. The protein-containing fractions were combined, concentrated and used for crystallization. B) SDS-PAGE gel shows the purified Ea10 protein after extraction by Resource Q column in lanes 8-9, which contain less contaminants. After gel filtration on a Superdex200 column, the majority of the Ea10 protein appears in lanes 11-17 and contains fewer contaminants. 15 µg of protein was loaded in each lane.

An alternative purification process for Ea10 protein was also followed using a 5ml Q-HP column (Section 2.14.3) instead of using two columns (DEAE Sepharose column and Resource Q column) and gave better purity and in less time. The Q-HP column was pre-equilibrated to 50mM Tris-HCl pH8 at 2.5ml/min. A sample of 94mg of protein in 27ml volume of the supernatant was applied. The Ea10 protein eluted in fractions 26-29 with some contaminant protein (Figure 4.8) as revealed by SDS-PAGE analysis (Figure 4.9 B).

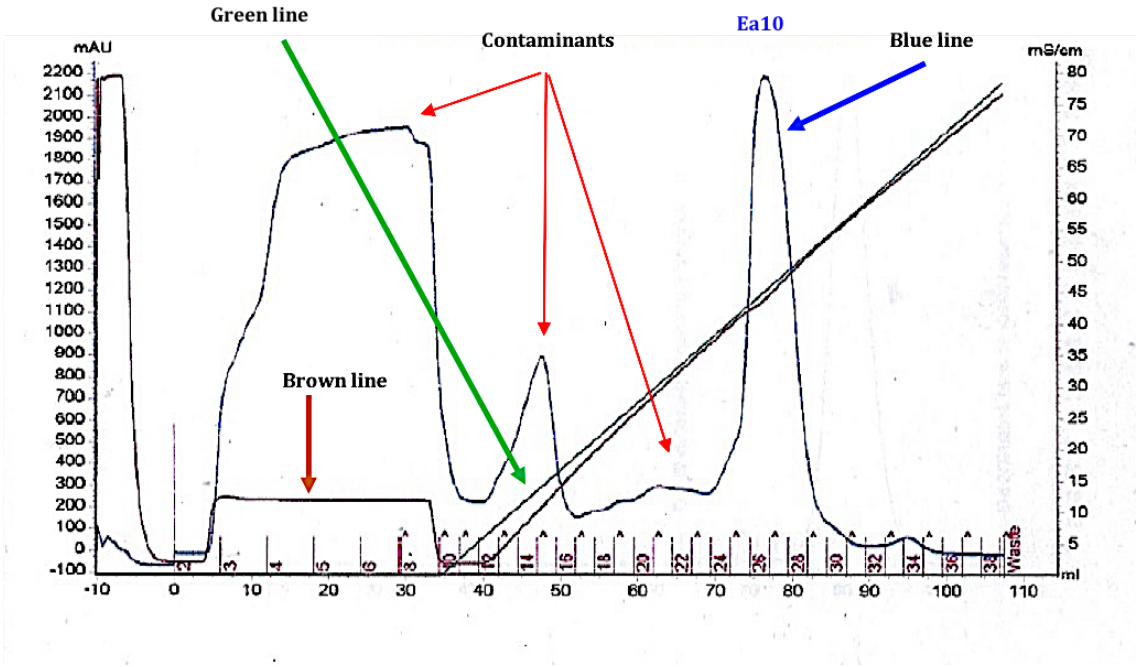


Figure 4.8. Q-HP purification for Ea10 protein.

The chromatogram of the result of the second method used to purify Ea10 protein via a 5ml Hi-Trap Q-HP column. The blue line represents the UV absorption at 280 nm, the green line represents the salt ingredient, and the brown line refers to the conductivity. The SDS-PAGE analysis showed that the Ea10 protein eluted in fractions 26-29, which were pooled and used for the next purification step.

The final volume of protein was 12.5 ml at 4.5mg/ml concentration, and this was concentrated to 2 ml volume to run on a Superdex^{TM} 200 column (Figure 4.9). The SDS-PAGE analysis of gel filtration of the protein sample revealed a dominant band belonging to the full-length molecular weight of the target protein, without any degradation whether the protein was stored at 4 ˚C or at room temperature. Over all, production of purified protein following this protocol was better than the previous one based on the time and the yield of purified protein at the final purification step.

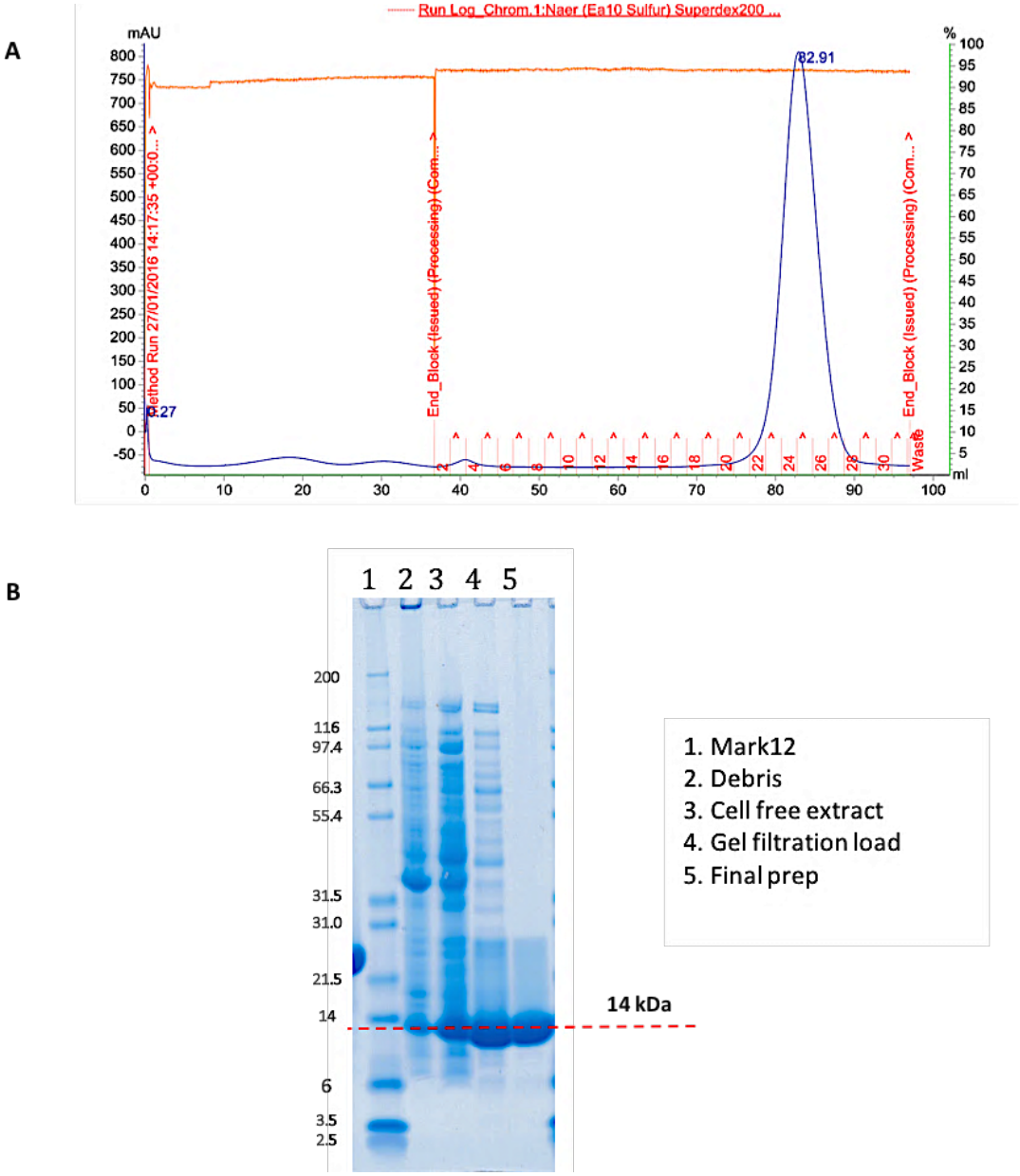


Figure 4.9. Gel-filtration assay for Ea10 protein sample.

A) The chromatogram shows the results of the final gel filtration on a Superdex^{TM} 200 column. Ea10 protein was eluted in fractions 23-26, which was similar to the results of the first method used. These fractions were combined, concentrated and used for crystallization. **B)** The SDS-PAGE gel shows the presence and the purity of Ea10 protein in the final prep from Q-HP and Gel filtration purification columns.

4.7 Mass spectrometry assay for the Ea10 protein

The purified protein sample of Ea10 was examined by mass spectrometry (Dept Chemistry, University of Sheffield). The result illustrated that the protein sample after the gel filtration column gave a major product (13.649 kDa) representing the expected Ea10 protein (13.780 kDa) missing the molecular weight of the first methionine (Figure 4.10). The mass spectrometry results confirmed that the purification strategy did not result in any proteolysis of the full-length target protein.

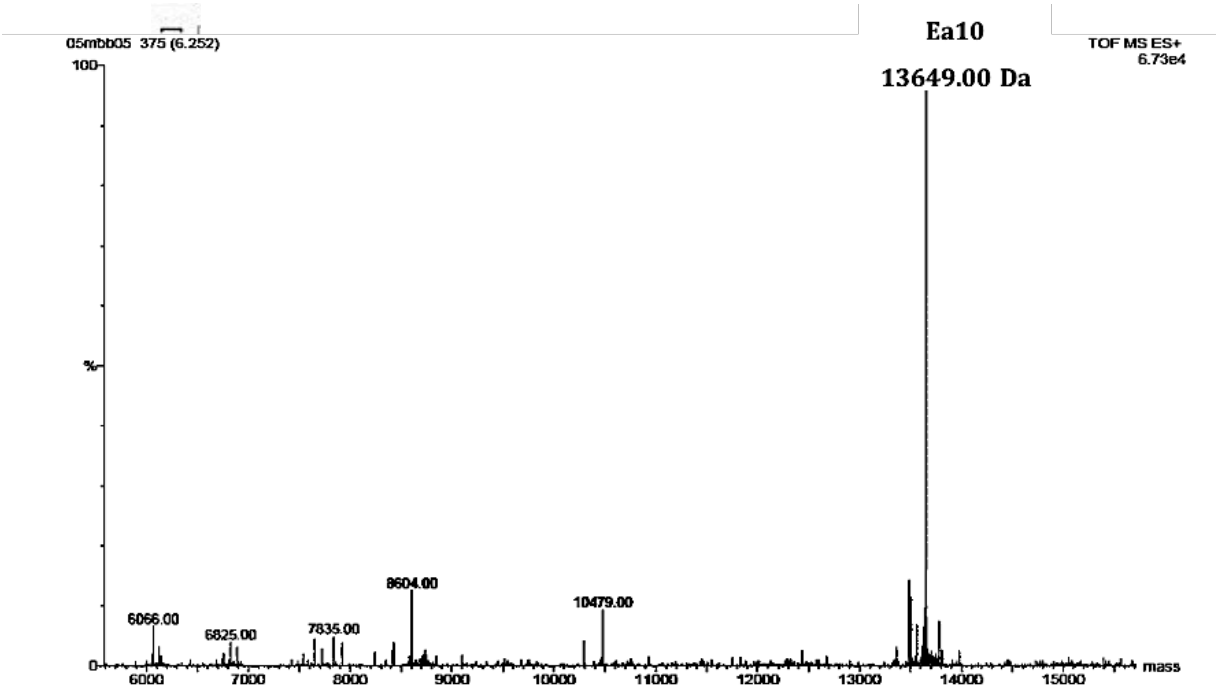


Figure 4.10. Mass-spectrum confirmation of the size of the purified Ea10 sample.

The mass-spectrum analysis confirmed the presence of Ea10 protein based on a dominant peak with a mass corresponding to the full-length molecular weight of the protein, without any degradation.

4.8 NMR (nuclear magnetic resonance) of purified Ea10 protein

The purified Ea10 protein sample was diluted to reduce the salt concentration into a buffer of 10 mM Tris-HCl pH 8.0, 50 mM NaCl, and 10% D_2O (0.5 mM concentration of protein), and then sent for analysis via NMR (MBB department, University of Sheffield) to study the nature of the Ea10 protein. The target protein was folded based on the NMR results as revealed single sharp peaks in the methyl region (before 0 ppm), and also by the peaks in the amide region (from 6 to 10 ppm), which were dispersed and sharp (Figure 4.11). The peaks of folded protein cover a great range of chemical shifts, which relies on the environment of protons and their positions in the tertiary structure of folded protein (Page et al., 2005).

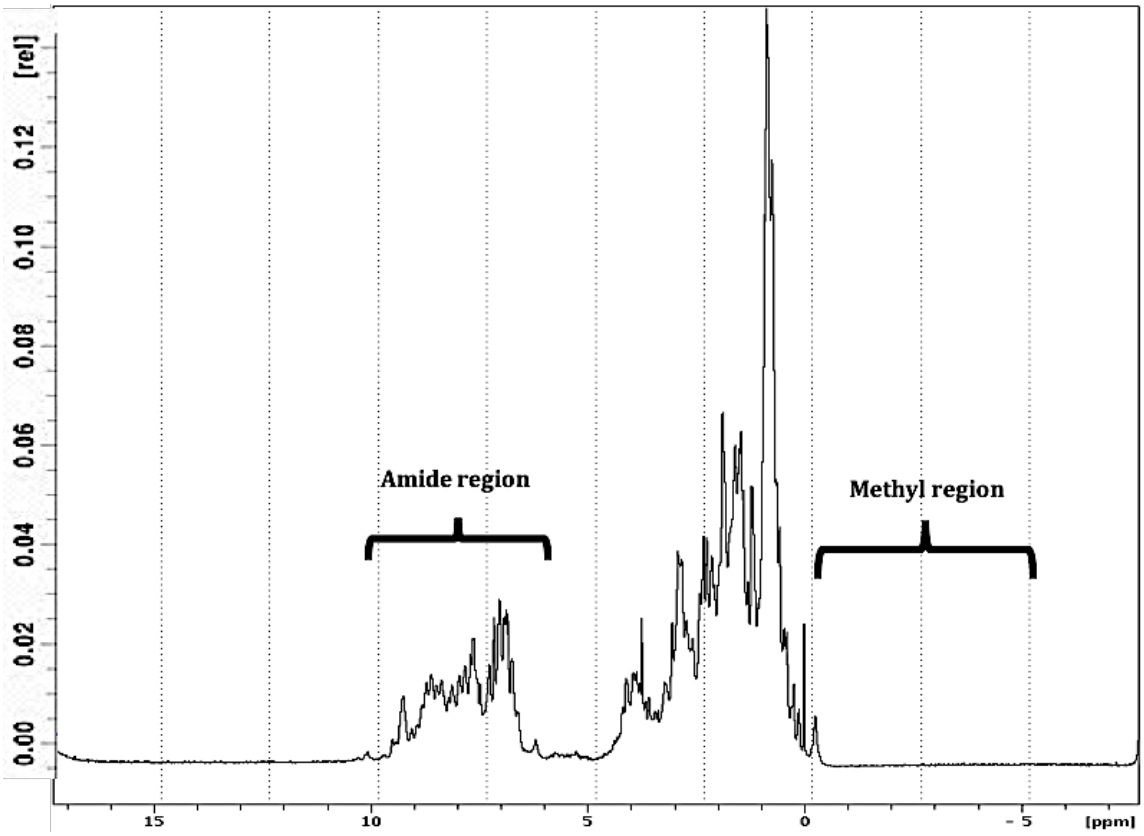


Figure 4.11. The NMR for Ea10 protein.

The NMR spectrum for the Ea10 protein and confirms that the protein was well folded, due to the presence of single sharp peaks in the methyl region (before 0 ppm), and also the dispersed and sharp peaks in the amide region (from 6 to 10 ppm). A volume of 1 ml with 0.5 mM concentration of protein was used to set up the NMR assay.

4.9 Crystallization of a native Ea10 protein

The purified Ea10 protein was concentrated to different concentrations (6, 10, 16 and 20 mg/ml) using a 1 mL 5 kDa Vivaspin column. The buffer was also exchanged into 5mM Tris-HCl, pH8 and 50 mM NaCl using a Zepa column.

The different concentrations of Ea10 were then used to set up crystal trials using various screening kits (PACT, MPD, JCSG, Ammonium Sulfate, Proplex and Classic). Sitting drop 96-well plates containing the different crystallization conditions from each screen were set up by the Hydra robot (Section 2.17.1). The best crystal forms were observed from 16 mg/ml protein concentration after 4 days at 7 ˚C in several conditions (Figure 4.12). The exact growth conditions for these crystals were optimized and the crystals harvested for collecting diffraction datasets at the Diamond Synchrotron, UK.

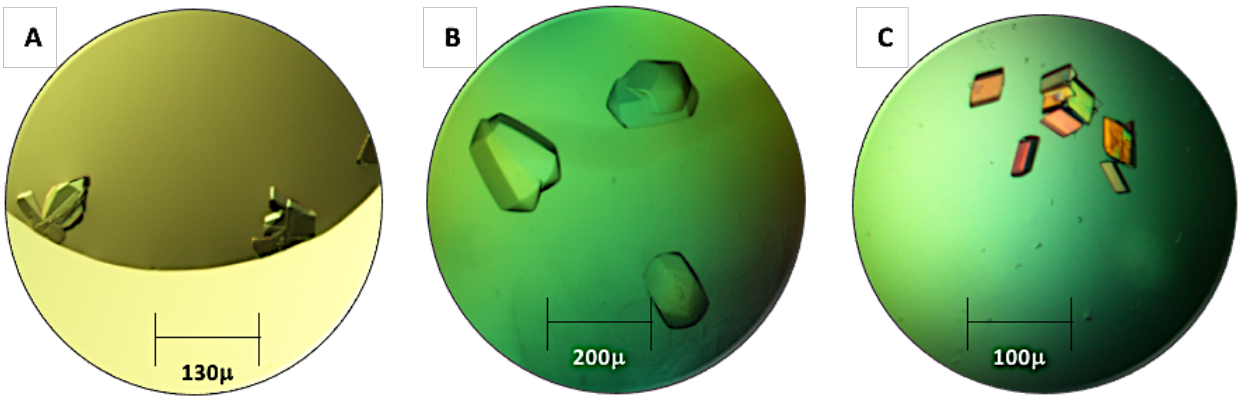


Figure 4.12. Crystallization screens of the native Ea10 crystals.

Many crystals of the native Ea10 crystals were grown in various screens: **A)** Many triangle shaped crystals were grown in screen condition B8 of the JCSG screen: 0.2 M Magnesium Chloride, 0.1 M Tris-HCl pH7, 10% PEG 8000. **B)** Giant diamond crystals were formed in the screen condition F6 of the PACT Screen: 0.2 Na formate, 0.1 M Bis-Tris propane pH 6.5, 20% PEG 3350. **C)** Many square shape crystals were grown in the screen condition E11 of the PACT Screen: 0.2 Na citrate, 0.1 M Bis-Tris propane pH 8.5, 20% PEG 3350.

4.10 collection strategy for Ea10 crystals

Many crystals of Ea10 protein were cryoprotected using 20-25% ethylene glycol to examine their diffraction properties via X-rays at the Diamond Synchrotron, UK. The highest resolution diffraction dataset was collected and observed to 2.4 Å resolution and belonged to a crystal grown from 0.2 M Na formate, 0.1 M Bis-Tris propane pH 6.5, 20% PEG 3350 (Figure 4.12 B). The strategy of collecting data was to collect 3600 images with 0.10° oscillation angle with 0.05 s exposure time, and 20, 40, 60, or 100% transmission.

4.11 Structure determination for the processed native dataset

The X-ray diffraction dataset was analyzed using the statistics from the AIMLESS program as a part of the XIA2 system for processing data in its xia2 3dii mode, which identified the space group as P 4_12_12 with unit cell a= 50.2 Å, b= 50.2 Å, c= 236.5 Å, α= 90°, β= 90°, γ= 90°. The Matthews coefficient calculations gave an estimate of two for the number of molecules in the AU according to a 14 kDa estimated molecular weight (2.66 Matthews coeff, 53.79 % solvent, 0.95 for P (2.35) and P (tot) (Figure 4.13). All the data statistics are described in Table 4.1.

Although the collected dataset analysis gave a promising sign for the likelihood of solving the Ea10 structure, this data was not quite enough to obtain the structure because there is no known structure of an Ea10 homologue that can give a model for use in molecular replacement. Therefore, it was not possible directly to get the missing phase information.

For estimated molecular weight 14000.

Nmol/asym	Matthews Coeff	%solvent	P(2.35)	P(tot)
1	5.32	76.89	0.00	0.01
2	2.66	53.79	0.95	0.95
3	1.77	30.68	0.04	0.04
4	1.33	7.57	0.00	0.00

Figure 4.13. Matthews's coefficients calculations for native Ea10 protein.

The Matthews's calculations suggest two molecules in the AU according to a 14 kDa estimated molecular weight.

Table 4.1. The data collection statistics for Ea10 native protein crystal.

The processed merging statistics for Ea10 native protein crystal. Analyzed by AIMLESS program via Xia 3dii.

SAD	Overall	Low	High
High resolution limit	2.38	10.64	2.38
Low resolution limit	50.19	50.19	2.44
Completeness	99.9	99.6	99.6
Multiplicity	13.4	9.2	11.8
I/sigma	14.0	55.3	1.4
Rmerge	0.163	0.030	2.055
Rmeas(I)	0.176	0.033	2.252
Rmeas(I+/-)	0.175	0.033	2.242
Rpim(I)	0.047	0.010	0.650
Rpim(I+/-)	0.063	0.012	0.887
CC half	0.999	0.999	0.607
Wilson B factor	34.465		
Anomalous completeness	99.8	100.0	99.5
Anomalous multiplicity	7.6	7.6	6.3
Anomalous correlation	0.043	0.320	0.016
Anomalous slope	1.028	0.000	0.000
dF/F	0.096		
dI/s(dI)	0.822		
Total observations	174985	1944	10762
Total unique	13054	211	911
Space group	P $4_1 2_1 2$		
Unit cell dimensions	a	50.2	
	b	50.2	
	c	236.5	
	α	90	
	β	90	
	γ	90	

$$R_{merge} = \frac{\sum_h \sum_{i=1}^{N} |I_{(h)i} - \bar{I}_{(h)}|}{\sum_h \sum_{i=1}^{N} I_{(h)i}}$$

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{j=1}^{n} |I_{hkl,j} - \langle I_{hkl} \rangle|}{\sum_{hkl} \sum_j I_{hkl,j}}$$

4.12 Overexpression and Purification for the Seleno-Methionine Ea10 protein

A selenomethioine (Sel-Met) incorporated Ea10 protein was produced in order to provide a way to obtain the first estimates of the phase angle information by using the anomalous signal from the selenium. The sequence of Ea10 contains four methionine residues, which can be replaced with seleno-methionine, and in addition to the protein phase information, their locations can also be used to help in the building of the Ea10 structure.

The protein was overexpressed following the same strategy as native Ea10 protein but using M9 Minimal media supplemented with seleno-methionine prior to protein induction. Additionally, a longer incubation time (18 hrs.) for the post induction cells was applied to obtain an acceptable quantity of expression (Section 2.13).

The same purification steps as for the native Ea10 protein were applied to purify the Sel-Met product by using a Q-HP column. The Ea10 protein was eluted in fractions 20-24 with contaminant proteins (Figure 4.14). The presence of Ea10 in these fractions was confirmed via SDS-PAGE analysis (Figure 4.16).

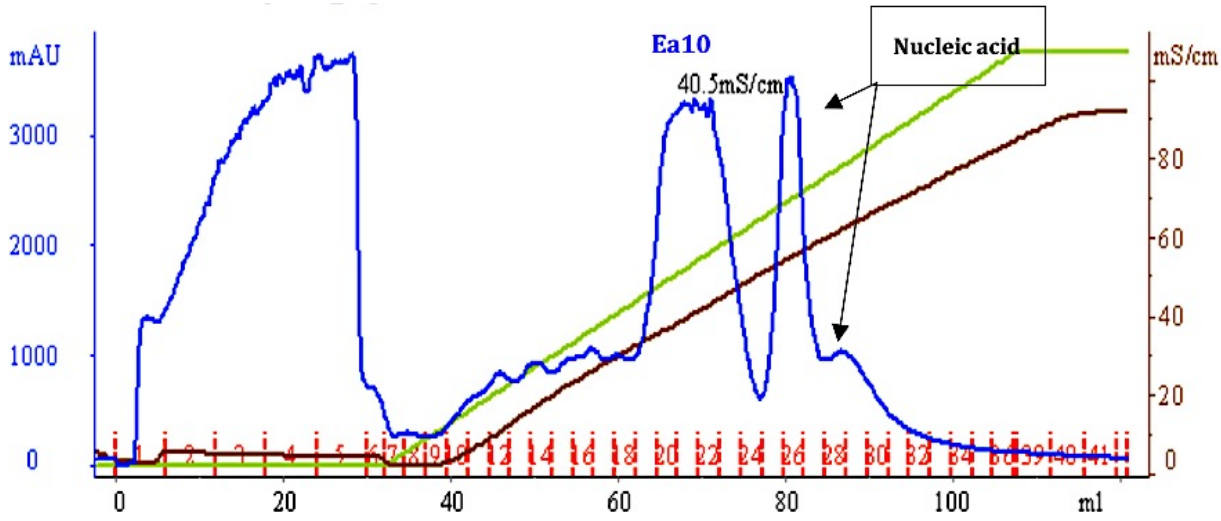


Figure 4.14. Q-HP purification for Sel-Met Ea10 protein.

The chromatogram shows the result of Ea10 protein elution from 5ml Hi-Trap Q-HP column. The blue line represents the UV absorption at 280 nm, the green line represents the salt ingredient, and the brown line refers to the conductivity. The results showed that the Ea10 protein was in fractions 20-24, which were pooled to 2 ml (46 mg) for Gel filtration use. Two peaks were shown in the chromatogram after Ea10 peak elution that are related to nucleic acid elution.

The 12.5 ml volume from the five fractions contained 56 mg of protein which was concentrated to 2 ml volume (46 mg) to run on a Superdex^{TM} 200 column following the same strategy as before. The result of gel filtration showed that the Ea10 protein was eluted in 84 ml volume, which corresponds to a dimeric state of the protein (Figure 4.15). The results showed a high level of purity on SDS-PAGE revealed by the dominant bands of the right size for the Ea10 protein (Figure 4.16), without any degradation or contaminant proteins.

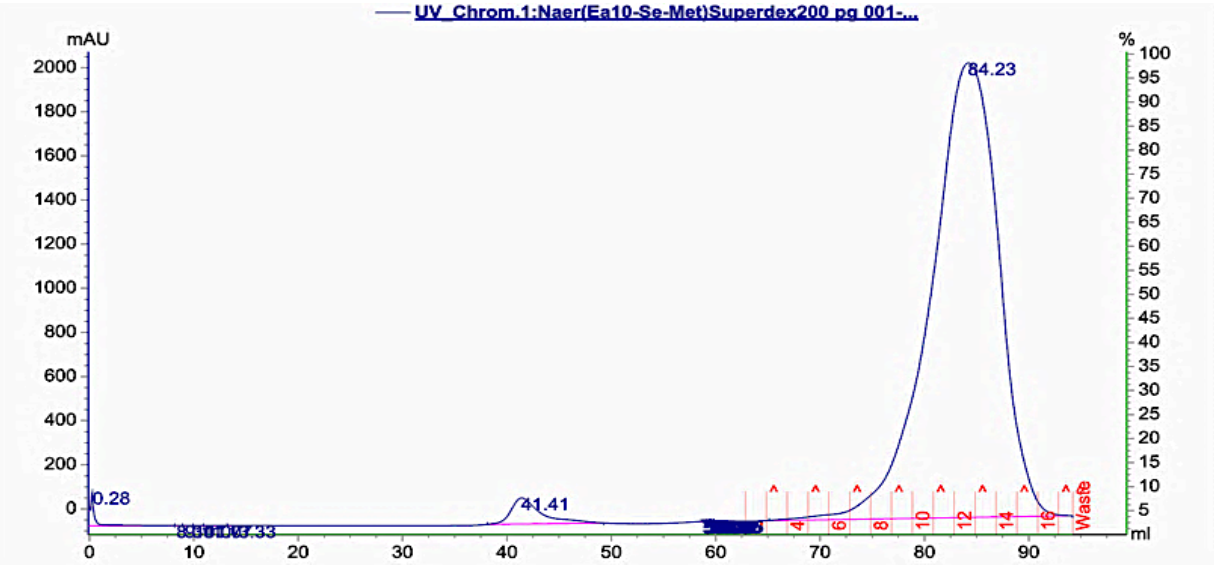


Figure 4.15. Gel-filtration assay for Sel-Met Ea10 protein sample.

The chromatogram shows the results of the gel filtration stage on a Superdex^{TM} 200ml column, which was run with 2 ml volume containing 46 mg protein. Fractions of 2 ml volume were collected throughout. The results showed that Ea10 protein was eluted in fractions 10-15. A sample of 24 mg of protein in 6 ml volume was concentrated to a volume of 0.4 ml (45 mg/ml) for crystallization trials.

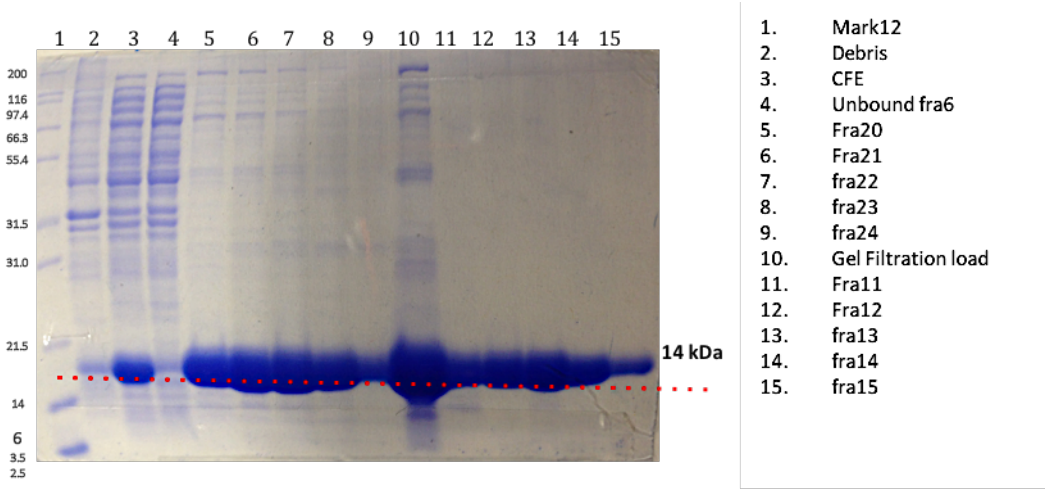


Figure 4.16. 12% acrylamide SDS-PAGE gel analysis of the Sel-Met Ea10 protein purification.

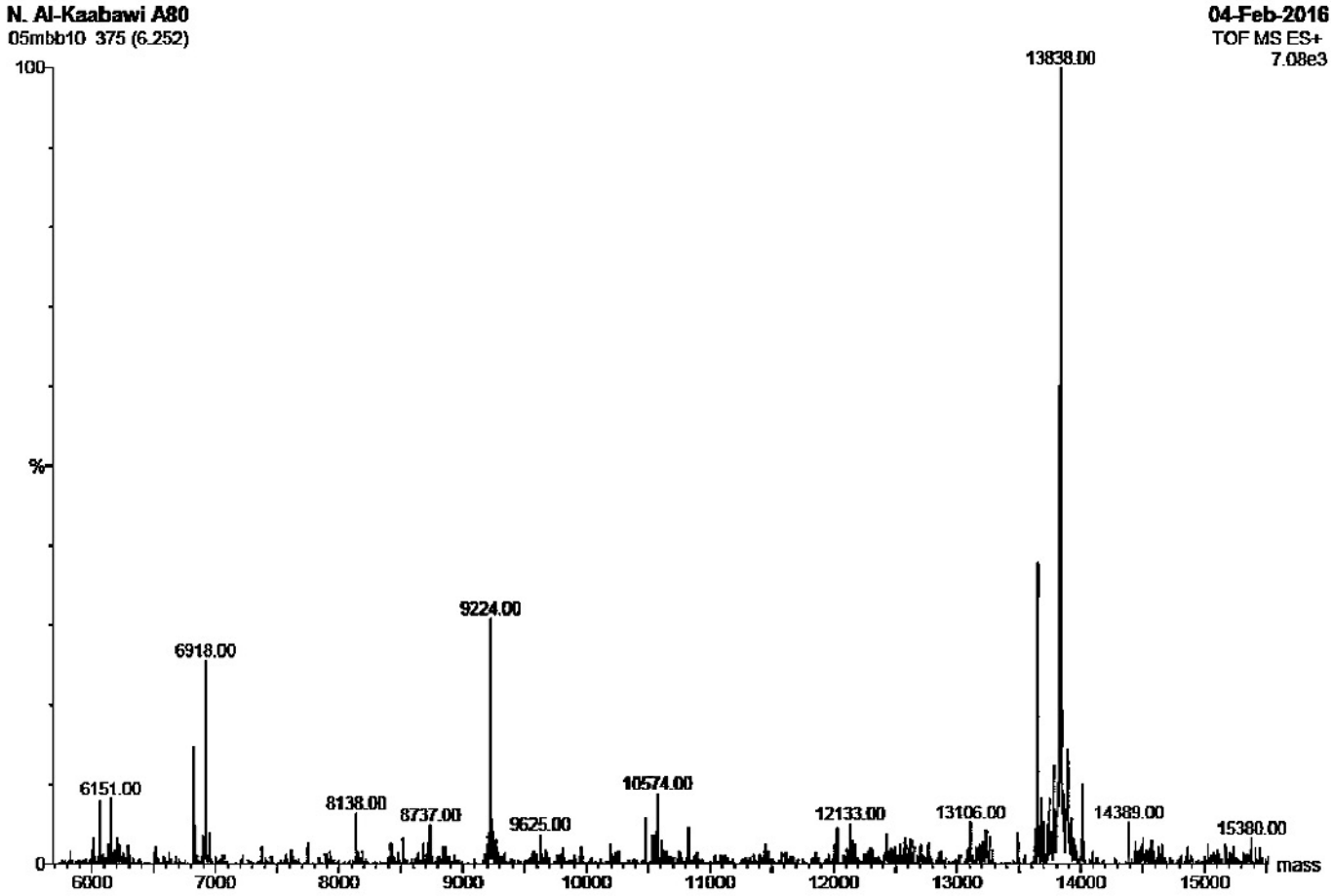
The SDS-PAGE gel shows the results of the purification stages of the Sel-Met Ea10 protein via Q-HP and Superdex-200ml columns, and also shows the level of purity of protein after each stage. 15 μg of protein was loaded in each lane.

The purified Ea10 protein was diluted to the concentration of 16 mg/ml and the buffer was also exchanged into 10 mM Tris-HCl, pH 8 and 50 mM NaCl using a Zepa column for setting up crystallization trials.

The presence of selenium was confirmed by analyzing the molecular weight for the Ea10 purified protein using mass-spectrometry, and the result revealed that all the methionine (4 amino acids) in the protein molecule were saturated with selenium due to obtaining a dominant band of 13838 Da, which corresponds to the endogenous molecular weight of Ea10 protein including the weight of four selenium atoms (316 Da) rather than four Sulphur atoms (128 Da) i.e. an extra 188 Da (Figure 6.17).

Figure 4.17. Mass-spectrum confirmation for the Ea10 Sel-Met protein.

The mass-spectrum analysis confirmed the presence of selenium in the Ea10 protein molecules based on a dominant band belonging to the full-length molecular weight of the target protein, with few contaminants.



4.13 Crystallization of Sel-Met Ea10 protein

The purified Sel-Met protein was then used to set up crystal trials using various screening kits, which were the same screens as mentioned in the crystallization section of native Ea10 protein. Each screen was set up by the Hydra robot (Section 2.17). Crystals were grown from a protein concentration of 12 mg/ml after 10 days at 7 ˚C in several conditions (Figure 4.18). These crystals were optimized and harvested for collecting diffraction datasets at the Diamond Synchrotron, UK.

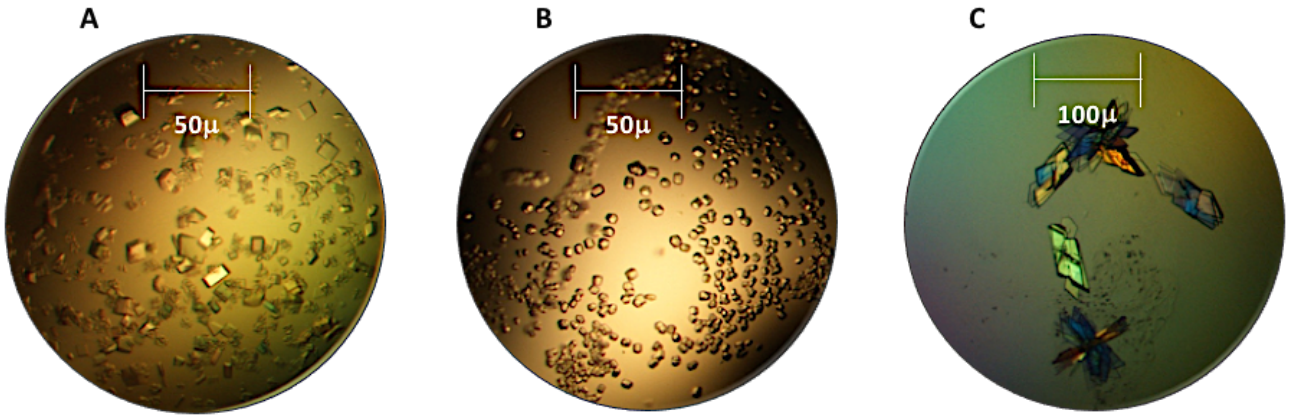


Figure 4.18. Crystallization screens of Sel-Met Ea10 crystals.

Many crystals of Sel-Met Ea10 crystals were grown in various screens: **A)** Variety of crystals forms were grown in the screen condition G11 of the Ammonium sulphate Screen: 1.6 M Ammonium sulphate, 0.1 M HEPES Na salt pH 7.5, 2% PEG 1000. **B)** Small square crystals formed in the screen condition G2 of the PACT Screen: 0.2 M Potassium Thiocyanate, 0.1 M Bis-Tris propane pH 7.5, 20% PEG 6000. **C)** Many square, clustered crystals were grown in the screen condition E11 of the PACT Screen: 0.2 M Na citrate, 0.1 M Bis-Tris propane pH 8.5, 20% PEG3350.

4.14 Data processing of Sel-Met Ea10 crystals

The same methodology for looping and cryo-protecting Sel-Met crystals was used as for the native crystals. The data were collected from singular mounted crystals on the I03 beamline at the Diamond Light Source. All the initial processes to test the crystals were similar to those described in section (Section 2.17.6).

The selenium incorporation was confirmed after performing a fluorescence scan at the selenium K edge to determine the proper X-ray energy for collecting Single Wavelength Anomalous Dispersion (SAD) experiment data. Based on the output of the fluorescence spectrum, the anomalous data were collected by selecting two energies, which were 12658 eV for the peak energy (maximized f ”), and 12655 eV of the inflection point (minimized f ’), from a plot of the transformed spectral data using the CHOOCH program (Figure 4.19) (Evans and Pettifer, 2001).

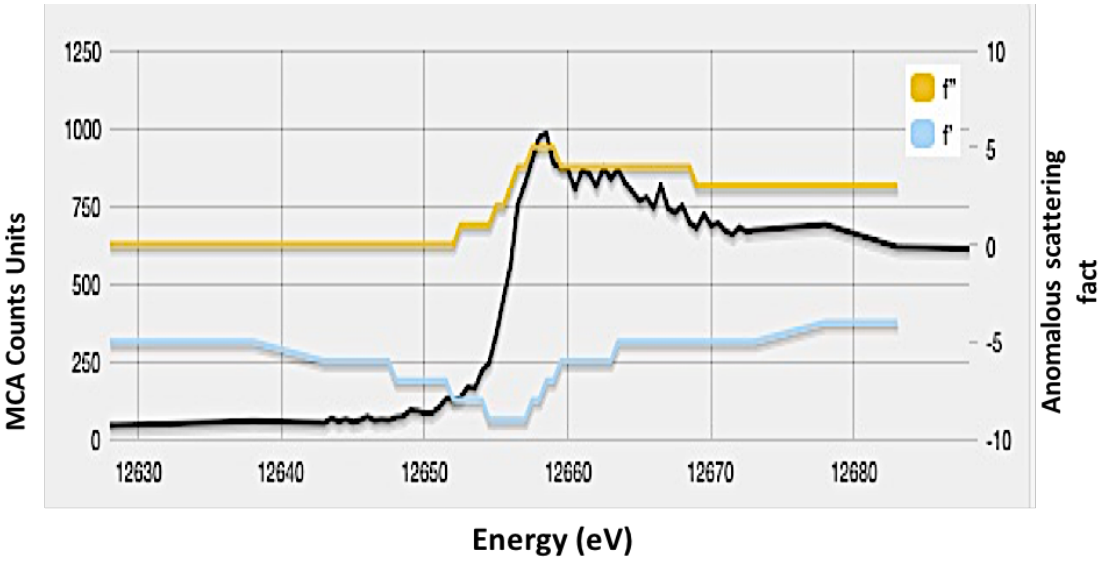


Figure 4.19. Fluorescence and CHOOCH plot against energy.

The incorporation of selenium into Ea10 protein crystals was confirmed by a fluorescence spectrum (black line) at the selenium K-edge, which gave a maximal fluorescence value at 12658 eV. The diagram also shows the chosen energies, from a plot of the fluorescence transformed using the program CHOOCH, for a MAD experiment. The selected energy peaks were 12658 eV for the maximum f “ and 12655 eV for the minimum inflection f ’.

For each data collection, 1800 images at each wavelength were collected with 0.05 sec exposure time and 0.1˚ oscillation per image (180˚ total) at various values of transmission (20-100%).

The datasets were auto-processed using the 3dii-run option via the xia2 system. A good resolution and diffraction data were obtained from optimised crystals in condition E11-JCSG: 0.6 M calcium acetate, 0.08 M sodium cacodylate pH 6.5, 14.5% PEG8000.

The highest resolution dataset was to 3.1 Å and belonged to space group $P2_1\,2_1\,2$ with cell dimensions a=72.8 Å, b= 65.6 Å, c= 68.3 Å, $\alpha = \beta = \gamma = 90°$ (see Table 4.2). The Matthews' coefficient calculations suggested two molecules in the AU according to a 14 kDa estimated molecular weight with a 2.92 Matthews coeff, 57.84 % solvent, 0.72 for P (3.1) and P (tot) (0.58) (Figure 4.20). All the data statistics are described in table 4.2.

For estimated molecular weight 14000.

Nmol/asym	Matthews Coeff	%solvent	P(3.13)	P(tot)
1	5.83	78.92	0.01	0.01
2	2.92	57.84	0.72	0.58
3	1.94	36.77	0.27	0.41
4	1.46	15.69	0.01	0.00

Figure 4.20. Matthews's coefficients calculations for Se-Met Ea10 crystals.

The Matthews's coefficient calculations suggest two molecules in the AU according to a 14 kDa estimated molecular weight.

Table 4.2. The data collection for Ea10 Sel-Met protein crystal.

The processed merging statistics for Ea10 Sel-Met protein crystal. The data were analyzed by the AIMLESS program via Xia 3dii.

SAD	Overall	Low	High
High resolution limit	3.1	14.0	3.1
Low resolution limit	65.5	65.5	3.21
Completeness	99.2	97.9	99.0
Multiplicity	6.7	4.7	7.3
I/sigma	2.8	8.7	0.8
Rmerge	0.570	0.128	2.132
Rmeas(I)	0.576	0.116	2.253
Rmeas(I+/-)	0.619	0.145	2.294
Rpim(I)	0.297	0.063	1.35
Rpim(I+/-)	0.237	0.066	0.839
CC half	0.97	0.99	0.57
Wilson B factor	38.6		
Anomalous completeness	98.4	97.3	99.0
Anomalous multiplicity	3.6	3.2	3.8
Anomalous correlation	0.26	0.77	0.03
Anomalous slope	1.089	0.000	0.000
Total observations	41092	410	3150
Total unique	6105	88	431
Space group	$P\ 2_1 2_1 2$		
Unit cell dimensions	a	72.8	
	b	65.51	
	c	68.48	
	α	90	
	β	90	
	γ	90	

4.15 Structure determination of Seleno-Methionine Ea10 protein

The programs of the SHELX package were used to attempt a Single Wavelength Anomalous Dispersion (SAD) experiment using the peak wavelength dataset and also a MAD (Multiple wavelength anomalous dispersion) experiment by adding the inflection wavelength dataset.

The anomalous normal probability reported a midslope of 1.089 using the peak dataset and contained a Rpim value in the inner shell of 0.066.

Analysis of the data using SHELXC showed that anomalous signal was present to 2.9 Å, and SHELXD was run using both enantiomorphs in a $P2_12_12$ space group. The best scores for 10 Se sites were found with a CCall of 27.73, CCweak of 12.84, and CFOM of 40.58%.

Moreover, these data were not enough to obtain a map, which gave a clear initial image of the Ea10 structure. It was not obvious how to interpret features in the map, for example the section is shown in Figure 4.21 could be an alpha helix or a beta sheet.

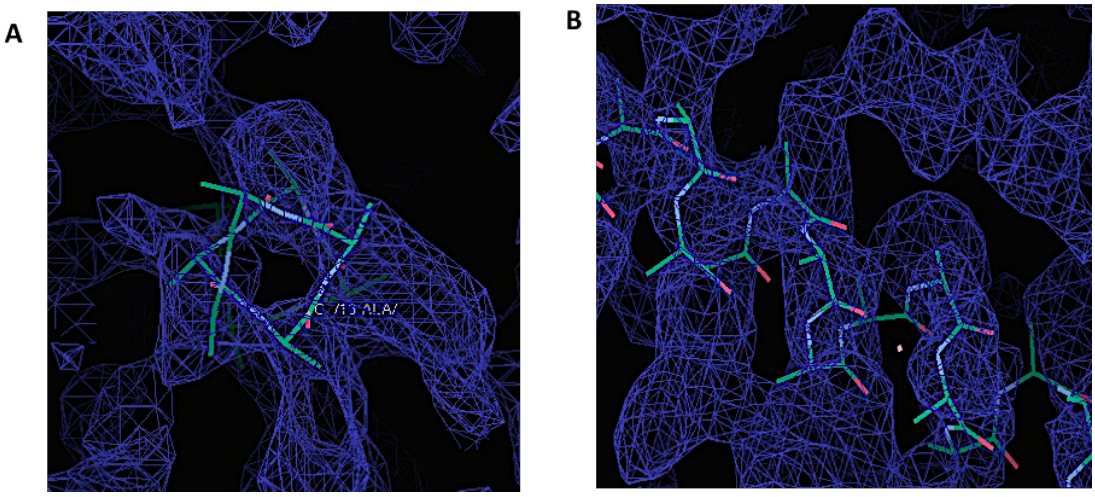


Figure 4.21. Electron density map generated from the Sel-Met Ea10 crystal.

A section of the electron density map (blue mesh) fitted with an alpha helix in different views. Although, the helix appears to fit in A), the perpendicular view in B) shows poor coverage of the model by the map. Overall, the fit of a model to the map was poor and the map could have been interpreted in a number of different ways.

4.16 Heavy Atom Derivatization.

The idea for using heavy atoms in a crystal (Section 2.18) is as a scattering source to obtain phasing information at low resolution. Their use helps in simplifying the solution of a heavy atom substructure which can then be used for obtaining the first estimate of the phasing angle for the protein map to build the Ea10 protein structure. Two approaches were used for heavy atom derivatization and hence determining the starting phase. A heavy atom soak with iodine (53 electrons), tantalum (73 electrons) or mercury (80 electrons) compounds was used for soaking protein crystals. The large number of electrons in the atoms of these elements relative to carbon, oxygen, nitrogen, and Sulphur in proteins has a measurable impact on the values of the scattering factors and diffracted x-ray intensities. These atoms also scatter x-ray anomalously at certain wavelengths and the effect of this on the reflection intensity can also be measured and used. Ea10 protein was proposed to have possibly a high affinity to bind to these heavy atoms because it contains four tyrosine residues (noted for potential iodine binding) and one accessible cysteine (noted for potential mercury binding) in its sequence.

The accessibility of the cysteine within the protein was confirmed by a dithionitrobenzoic acid (DTNB) reaction test (Section 2.16). A 1.5 mM solution of the Ea10 protein was used in this test. The results showed that the DTNB reagent was bound to the Ea10 protein potentially and it confirmed the accessibility of the cysteine present in the protein.

The crystals were left overnight to allow complete soaking with the reagent, and then mounted to examine their X-ray diffraction. Although, many derivative crystals were tested at the Diamond Light Source in Harwell, not one diffracted well enough to solve the structure. Therefore, an alternative method was applied by setting down co-crystallization screens with ethyl mercury phosphate (EMP) using a final 2:1 molar ratio of EMP: protein for crystallization. The mix was incubated for 2 hours at room temperature, and then the same crystallization strategy was applied as for the native crystals (section 2.17). Many crystals in different sizes and shapes were grown in many conditions using a protein concentration of 30 mg/ml after 10 days at 7 ˚C incubation (Figure 4.22).

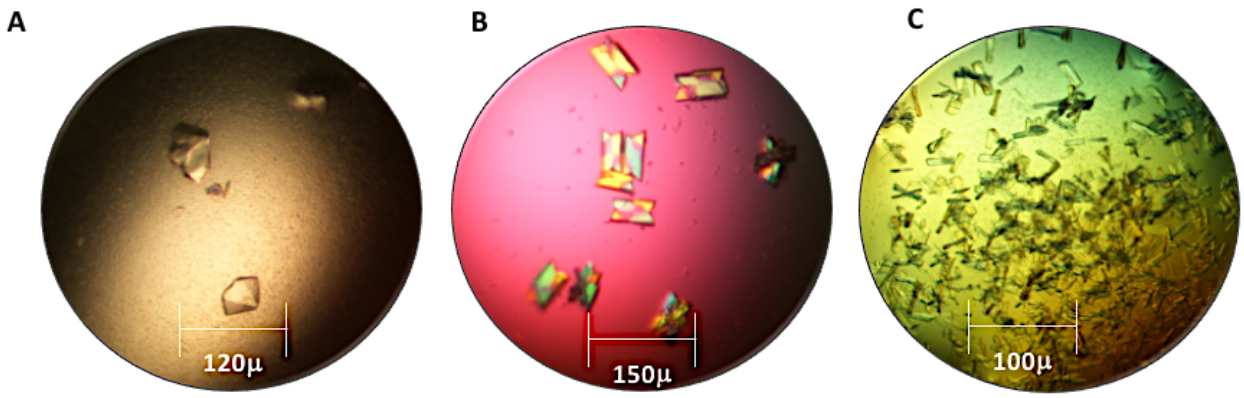


Figure 4.22. Co-crystallized Ea10 crystals with EMP.

Various crystals of Ea10 were grown co-crystallized with EMP: **A)** Big diamond-shaped Ea10-hg crystals were grown in the screen condition F4 of PACT Screen: 2 M thiocyanate, 0.1 M Bis-Tris propane pH 6.5, 20 % PEG3350. **B)** Big square crystals forms of Ea10-hg were grown in the screen condition G2 of MPD Screen: 0.05 M magnesium Chloride, 0.1 M Tris-HCl pH 8.5, 12% MPD. **C)** Various rod-shaped crystals of Ea10-hg were formed in the screen condition D2 of JCSG Screen: 2 M magnesium chloride, 0.1 M HEPES pH 7.5, 30% PEG400.

4.17 Data processing of Ea10-Hg crystal

The optimized crystals were from condition F1-PACT: 0.2 M Na fluoride, 0.1 M Bis-Tris propane pH 6.5, 20% PEG 3350. A crystal centering was carried out during initial tests on each crystal at a wavelength of 0.9795 Å, in which 4 images of a crystal were used at 45˚ apart on the rotation axis with 0.5˚ Oscillation, 0.5 sec exposure time, and 20 % transmission for each image. The initial indexing and collection strategies were suggested by EDNA or Mosflm programs and used to guide a starting point for the rotation range before collecting 180˚ or 360˚ of data in 1800 or 3600 images.

A fluorescence scan of EMP co-crystallized Ea10 crystal was taken over energy range of 12.240-12.4 keV, which confirmed the presence of mercury. A peak dataset was collected at a wavelength of 1.0073 Å for 3600 images with 0.05sec exposure time and 0.1˚ oscillation per image (360˚ rotation range) at 40% transmission. The best crystal diffraction was obtained to 2.7 Å resolution.

The space group was determined as $P3_1$ 21 with cell dimensions a = 85.3 Å, b = 85.3 Å, c = 159.5 Å, α = 90°, β = 90° and γ = 120° (Figure 4.23). The overall anomalous midslope of the data was 1.14, but with a high CC half score (0.45) in the high-resolution bin. The overall score for Rmerge and the Rpim per batch were 0.14 and 0.03, respectively.

The processing statistics are described in Table 4.3. The estimated number of molecules from a Matthew's coefficients calculation (48.6 % solvent, 0.54 for P (2.35) and P (tot)) was five molecules in the AU according to a 14 kDa estimated molecular weight (Figure 4.24).

Table 4.3. SAD dataset for the Ea10 crystal co-crystalized with EMP.

The auto-processed statistics for the Ea10 crystal co-crystalized with EMP. The data were analyzed by the AIMLESS program via Xia 3dii.

SAD	Overall	Low	High
High resolution limit	2.68	7.27	2.68
Low resolution limit	73.88	73.91	2.73
Completeness	100	100.0	100.0
Multiplicity	19.5	16.8	19.0
I/sigma	15.8	43.6	1.4
Rmerge(I)	0.139	0.080	2.138
Rmerge(I+/-)	0.131	0.070	2.088
Rmeas(I)	0.143	0.082	2.196
Rmeas(I+/-)	0.137	0.073	2.201
Rpim(I)	0.033	0.020	0.501
Rpim(I+/-)	0.042	0.023	0.697
CC half	0.998	0.998	0.448
Wilson B factor	0		
Anomalous completeness	100	100.0	5.1
Anomalous multiplicity	10.4	10.0	10.0
Anomalous correlation	0.594	0.643	-0.035
Anomalous slope	1.144	0.000	0.000
dF/F	0.078		
dI/s(dI)	1.205		
Total observations	380334	18237	18097
Space group	$P3_1$ 21		
Unit cell dimensions	a	85.3	
	b	85.3	
	c	159.5	
	α	90	
	β	90	
	γ	120	

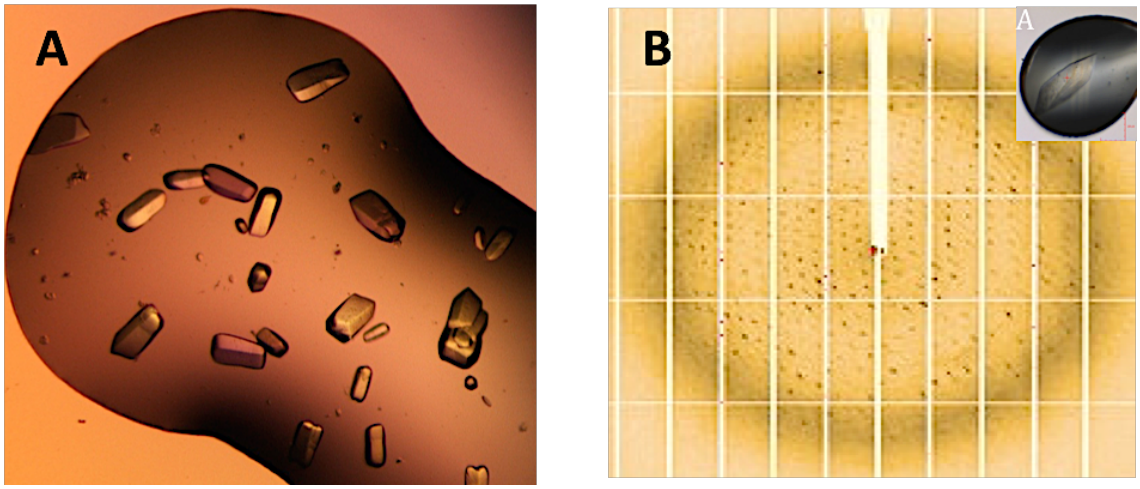


Figure 4.23. Ea10-Hg crystals and diffraction.

Purified Ea10 was co-crystallized with ethyl mercury phosphate for heavy atom derivatization. **A)** Crystals were observed in the condition containing 0.2M Na fluoride, 0.1M Bis-Tris propane pH6.5, and 20% PEG3350. **B)** Diffraction image from data collected on Ea1-Hg crystal (F1), which gave data up to a resolution of 2.7 Å.

For estimated molecular weight 14000.

Nmol/asym	Matthews Coeff	%solvent	P(2.68)	P(tot)
1	11.96	89.73	0.00	0.00
2	5.98	79.45	0.00	0.00
3	3.99	69.18	0.03	0.02
4	2.99	58.90	0.21	0.20
5	2.39	48.63	0.54	0.53
6	1.99	38.35	0.21	0.24
7	1.71	28.08	0.00	0.01
8	1.50	17.80	0.00	0.00
9	1.33	7.53	0.00	0.00

Figure 4.24. Matthews's coefficients calculations for Ea10 crystals co-crystallized with EMP.

The Matthews's coefficients calculations suggest five molecules in the AU according to a 14 kDa estimated molecular weight.

4.18 Structure determination of Ea10-Hg dataset

Attempts to solve the structure of the Ea10 crystal by EMP were carried out via SHELX in SAD mode. The anomalous signal present in the dataset was determined in SHELXC, based on where the d"/sig fell below 0.8. SHELXC showed that anomalous signal was present to 2.68 Å. SHELXD was run using both enantiomorphs in the $P3_121$ space group with the recommended resolution cut-off of 2.68 Å, with 1000 tries and searching for 10 Hg heavy atoms. The best scores were obtained for 1 Hg per monomer with an average of CCall and CCweak of 16.18 and 7.1 respectively. Finally, SHELXE was run to calculate phases and generate an initial electron density map using the best solution of SHELXD. Cycles of density modification and chain tracing were performed using RefMac5 and coot programs. Both original and inverted hands for the heavy atom coordinates in order were used to break the phase uncertainty. The interpretation of the resulting output electron density map showed that the heavy atom determination was of satisfactory quality. However, it was not a fully connected density map with incomplete observation of sidechains (Figure 4.25).

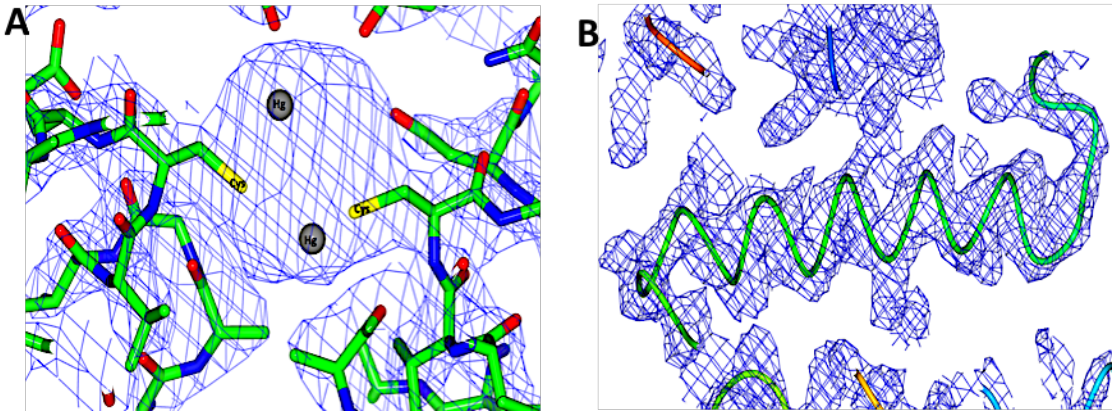


Figure 4.25. Electron density maps created by SHELXE for Ea10-Hg crystal.

A) . Large density peak (20 rmsd) for Hg heavy atoms, which were observed in the electron density maps. The bound Hg atoms appeared to occupy two adjacent locations, probably corresponding to alternative conformations for the cysteine sidechain to which they are bound. **B)** The map following a density modification and chain building of the original hand of the heavy atom substructure model, which shows a good solvent boundary comparing with the inverted hand model, using RefMac5 and coot programs.

4.19 Construction of Ea10-Hg model using Buccaneer

The BUCCANEER program was run to build an initial model of Ea10 using the map generated in SHELXE with initial phase estimates as well as the protein sequence. The BUCCANEER program created a model after 5 cycles of building and refinement (Cowtan, 2006), which consisted of 822 residues built in 21 fragments (Figure 4.26), where the biggest fragment contained 100 residues. The residues were initially assigned as alanine amino acids. They were built as a poly-alanine to obtain the overall chain trace and improve the initial estimated phases, and then the program tried to identify amino acid sidechains aided by the knowledge of the sequence. An improved model was constructed after joining the fragments, and 466 residues were assigned an amino acid type and allocated to four chains in the AU. The R factor and R free after model refinement were 0.36 and 0.38, respectively.

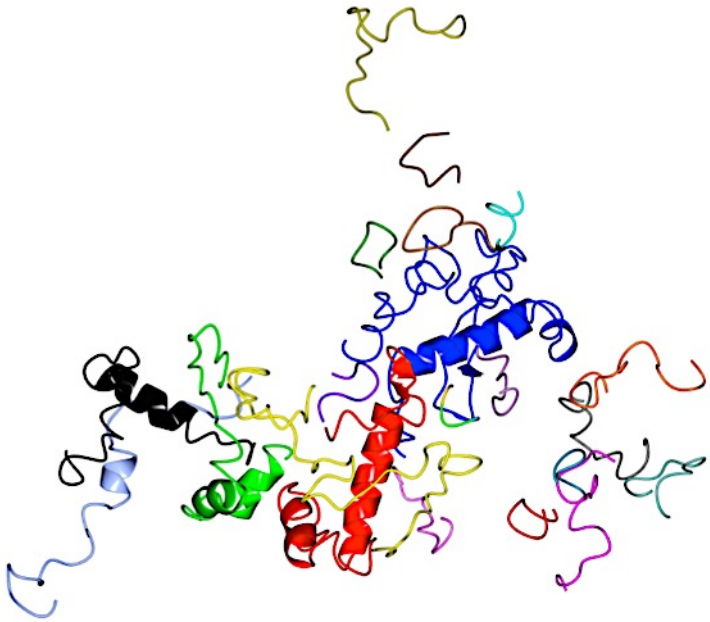


Figure 4.26. The generated model of Ea10 structure by Buccaneer.

The initial structure model of Ea10-Hg protein was built and refined by Buccaneer. The final model after 5 cycles of refinement had 21 fragments (822 residues) shown here in different ribbon colours. It also reported 0.36 R factor and 0.38 R free.

The model from Buccaneer was used to rebuild the protein structure manually in COOT and refine it via REFMAC5 (Emsley et al. 2010: Murshudov et al. 1997) from the CCP4 suite. After a first round of building and refinement, the results showed four subunits (A, B, C, and D) with R factor (0.34) and R free (0.39). The model construction continued until there were no further features to interpret in the electron density map. The residues and their sidechains were fitted to the electron density map, and missing residues were added, where there were obvious features in the electron density. A model containing four molecules in the asymmetric unit resulted during the final stage of building and refinement. The R factor and R free of the final refinement were 0.34 and 0.38 respectively. A model containing four molecules in the asymmetric unit resulted during the process of building is shown in figure 4.27.

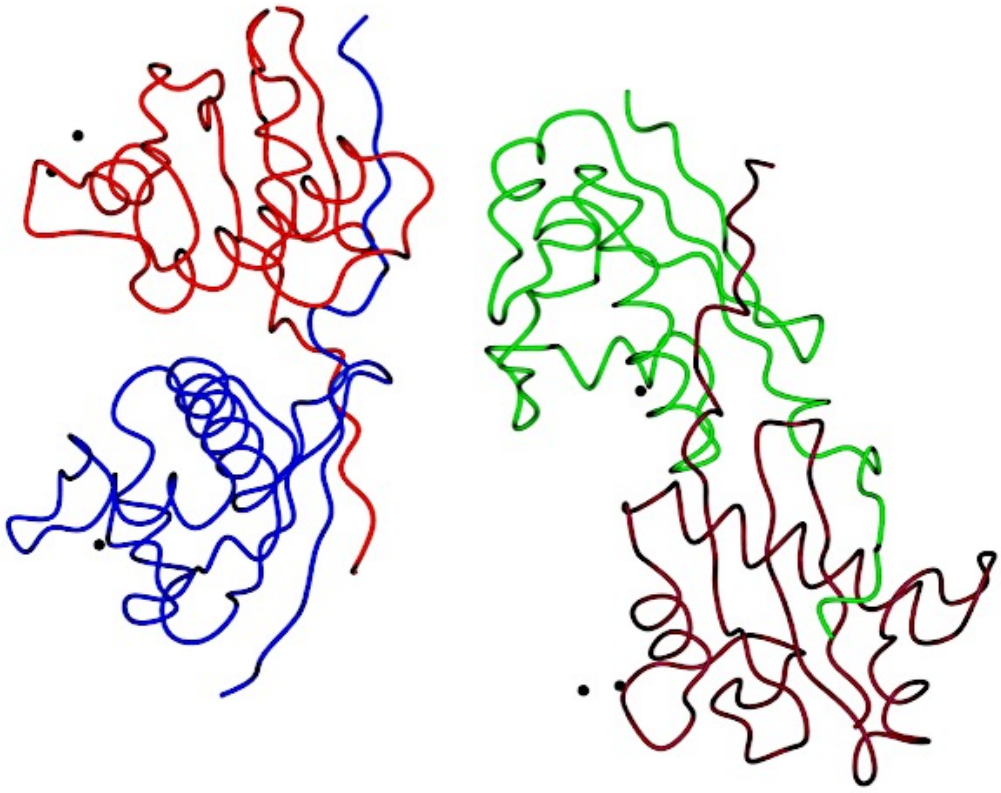


Figure 4.27. The final refined structure of the Ea10-Hg crystal.

The model of the Ea10-Hg crystal structure was built and refined with COOT and REFMAC5 programs from CCP4 suite. The model was made of four molecules shown in different colours. R factor (0.34) and R free (0.38), respectively.

4.20 Molecular replacement for the native dataset of Ea10 crystal.

The final structure of Ea10 from the mercury derivative data was used as a model for molecular replacement with the higher resolution native dataset of 2.4 Å, and the structure was refined versus the native dataset of the original crystal in space group $P4_12_12$ (Table 4.1). The molecular replacement procedure was accomplished by using PHASER-MR from the CCP4 suite to search for two subunits in the asymmetric unit based on the Matthews coefficients calculations (Figure 4.13). Alternative space groups related to P 4_12_12 were tested (McCoy, 2006).

The molecular replacement results suggested two potential solutions that gave 28.8 of TFZ scores per molecules with an 896 overall LLG scores. These values indicated a correct solution. The resulting electron density map was slightly improved and revealed new density for certain residue sidechains in comparison with the density for the map with the Ea10 model. The model was then put through 14 cycles of restrained refinement via REFMAC5 after placing all the residues in two subunits manually in COOT (Table 4.4).

Table 4.4. The statistics of the native Ea10 structure refinement after molecular replacement.

The table shows the improvement in the R values in the final refinement cycle for the Ea10 structure after molecular replacement in comparison with the initial values.

Value	Initial	Final
R factor	0.3343	0.2042
R free	0.4193	0.2426
Rms Bond length	0.0192	0.0203
Rms Bond Angle	2.2246	2.0483
Rms Chir Volume	0.1558	0.1407

4.21 Structure validation of Ea10 crystal By MolProbity.

The final refined model of Ea10 contained 100 % of residues of two chains (244 residues) and was analyzed in MolProbity (Chen et al., 2010), to analyze the structure in terms of properties such as Ramachandran outliers, bond lengths and sidechain rotamers. Two statistics are used by MolProbity analysis to calculate the model quality by comparing the query model with other submitted structures in the PDB across a resolution range of 0.25 Å. The Clash score indicates the number of the modelled atoms within clashing distance of one another after incorporation of hydrogen atoms into the model. The final refinement and validation statistics are shown in Table 4.5. The results showed 0.00 % poor rotamers with 98.57% Favored rotamers. The overall clash score was 3.48 which qualified for the 99th percentile * (N=792, 1.68Å ± 0.25Å), 1.35 overall MolProbity score qualified for the 100th percentile * (N=937, 2.35Å ± 0.25Å). No residues were Ramachandran outliers, 229 residues (69.64%) were in preferred regions, while 8 residues were in the allowed region (33.36%) (Figure 4.28). The last refinement of the final structure showed a noticeable improvement for the electron density map and reported an R factor of 0.2 and R free of 0.25 (Figure 4.29).

Table 4.5. Validation statistics of Ea10 structure.

MolProbity analysis output (Chen *et al.,* 2010) shows the statistics of the validation score of Ea10 structure after the final refinement, which reported 0.2 R factor and 0.25 R free.

All-Atom Contacts	Clashscore, all atoms:	3.48		99th percentile* (N=335, 2.35Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	0	0.00%	Goal: <0.3%
	Favored rotamers	207	98.57%	Goal: >98%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	230	96.64%	Goal: >98%
	MolProbity score^	1.35		100th percentile* (N=9377, 2.35Å ± 0.25Å)
	Cβ deviations >0.25Å	1	0.44%	Goal: 0
	Bad bonds:	0 / 1958	0.00%	Goal: 0%
	Bad angles:	2 / 2658	0.08%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 4	0.00%	Expected: ≤1 per chain, or ≤5%

In the two column results, the left column gives the raw count, right column gives the percentage.

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst. For clashscore the comparative set of structures was selected in 2004, for MolProbity score in 2006.

^ MolProbity score combines the clashscore, rotamer, and Ramachandran evaluations into a single score, normalized to be on the same scale as X-ray resolution.

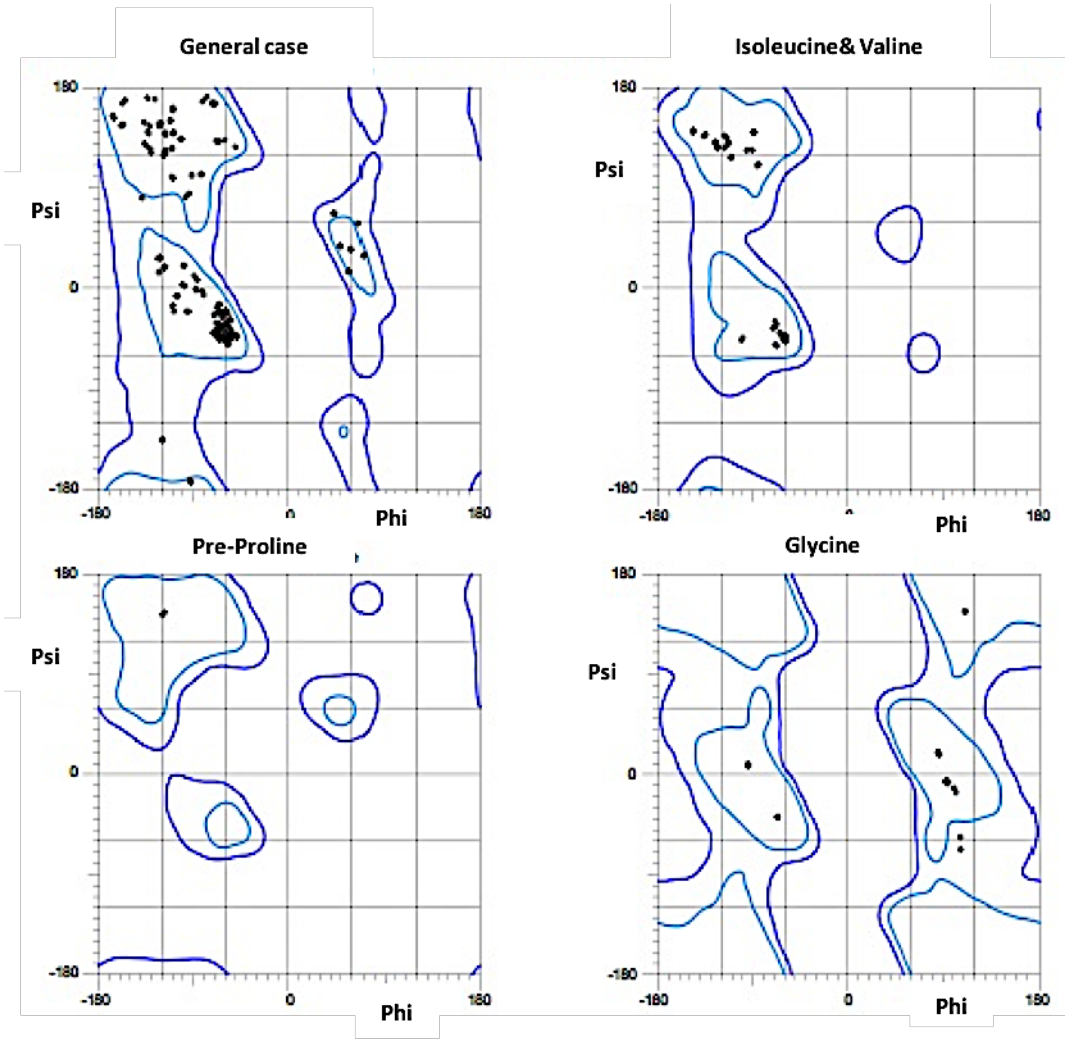


Figure 4.28. Ramachandran analysis of the final structure of Ea10.

A Ramachandran plot map for the final refined model of Ea10 structure produced using the MolProbity server. http://molprobity.biochem.duke.edu/. The general case of each amino acid in the structure is represented as a one plot. Other plots were produced for isoleucine and valine, pre-proline residues and glycine. MolProbity Ramachandran analysis torsion viewpoints shows the Phi angles plotted versus Psi angles for each residue, the islands of preferred Ramachandran angles in light blue, and the allowed Ramachandran angles in dark blue. Outside of this region are unacceptable Ramachandran angles. For Ea10 crystal, 229 residues (69.64%) were in preferred regions, while 8 residues were in the allowed region (33.36%). MolProbity analysis output (Chen *et al.,* 2010).

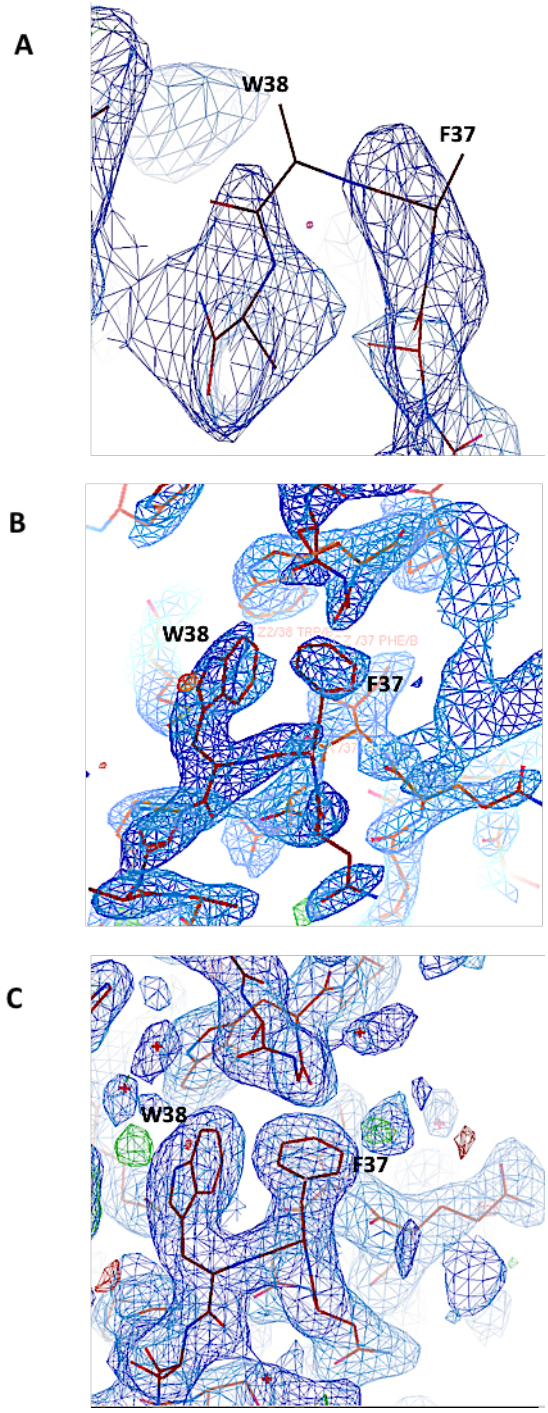


Figure 4.29. Progress of electron density maps of Ea10 structure.

The progression of electron density maps during the refinement process. The images show the map at F37 and W38 residues (initially as an alanine) in chain B. **A**) The electron density as it looked after BUCCANEER (0.43 R factor, 0.40 R free). **B**) The map after the final refinement for the Ea10-Hg crystal dataset that reported 0.34 R factor and 0.38 R free. C) The final electron density map after the MolProbity analysis for the Ea10 dataset, with final refinement reported R-factors of 0.2 R factor and 0.25 R free.

4.22 Structure analysis of Ea10 protein

The structure of Ea10 was determined and it was seen that one monomer (122 residues) starts with two anti-parallel β strands (S1 and S2) followed by four distinct α-helices (H1-H4), which are connected by short loops, and characterized by a long α-helix (H3, 51st-71st residue), then followed by a third β strand (S3). In addition, the C-terminal polypeptide extends away interestingly to a long distance of 35 Å from base to tip (S107-A122) and 41 Å from top loop to tip (S39-A122) to form the fourth strand (S4) (Figure 4.30). This structure seemed to share some similar characteristics with the predicted model from the Phyre2 server (Section 4.1) which contained two anti-parallel β strands followed by three possible α-helices and ended with a β-strand at the C-terminal end.

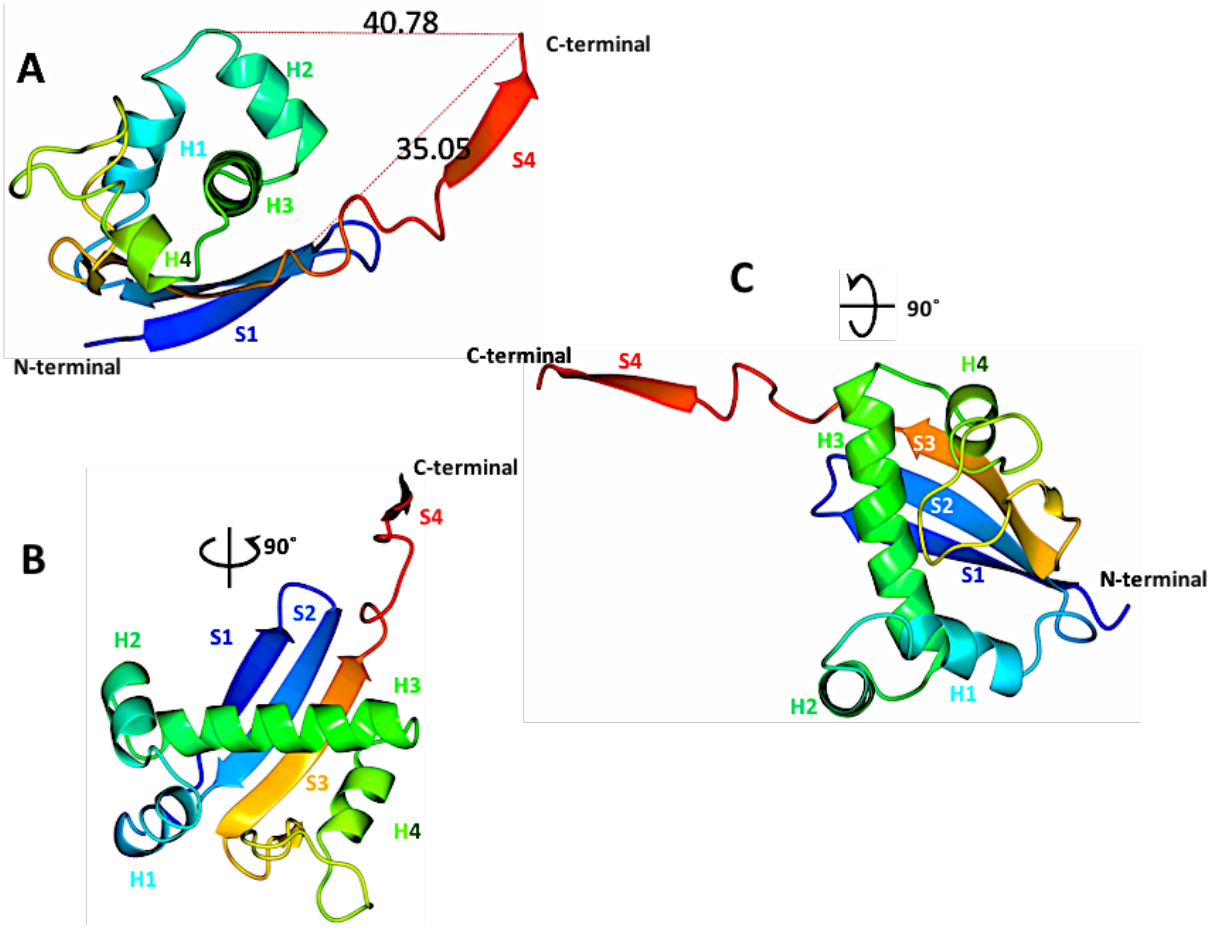


Figure 4.30. Structural form of the Ea10 protein.

The figure shows a ribbon form in rainbow colours for a monomer structure of the Ea10protein. This subunit contains four α-helices (H) characterized by a long α-helix (H3, 51st-71st residue), and three anti-parallel β-strands (S), and a fourth β-strand formed at the C-terminal end, which extends further from the body of the structure to a distance of 35 Å from base to tip (S107- A122). It is shown in different orientations: the front view A; the side view B; and the top down view C.

4.23 Overall structure of Ea10

The structure of Ea10 was built up of two monomers in a dimeric form. It is found that the extended C-terminal polypeptide forms the fourth strand (S4) with the three anti-parallel β-strands bundle of the neighboring monomer (Figure 4.31).

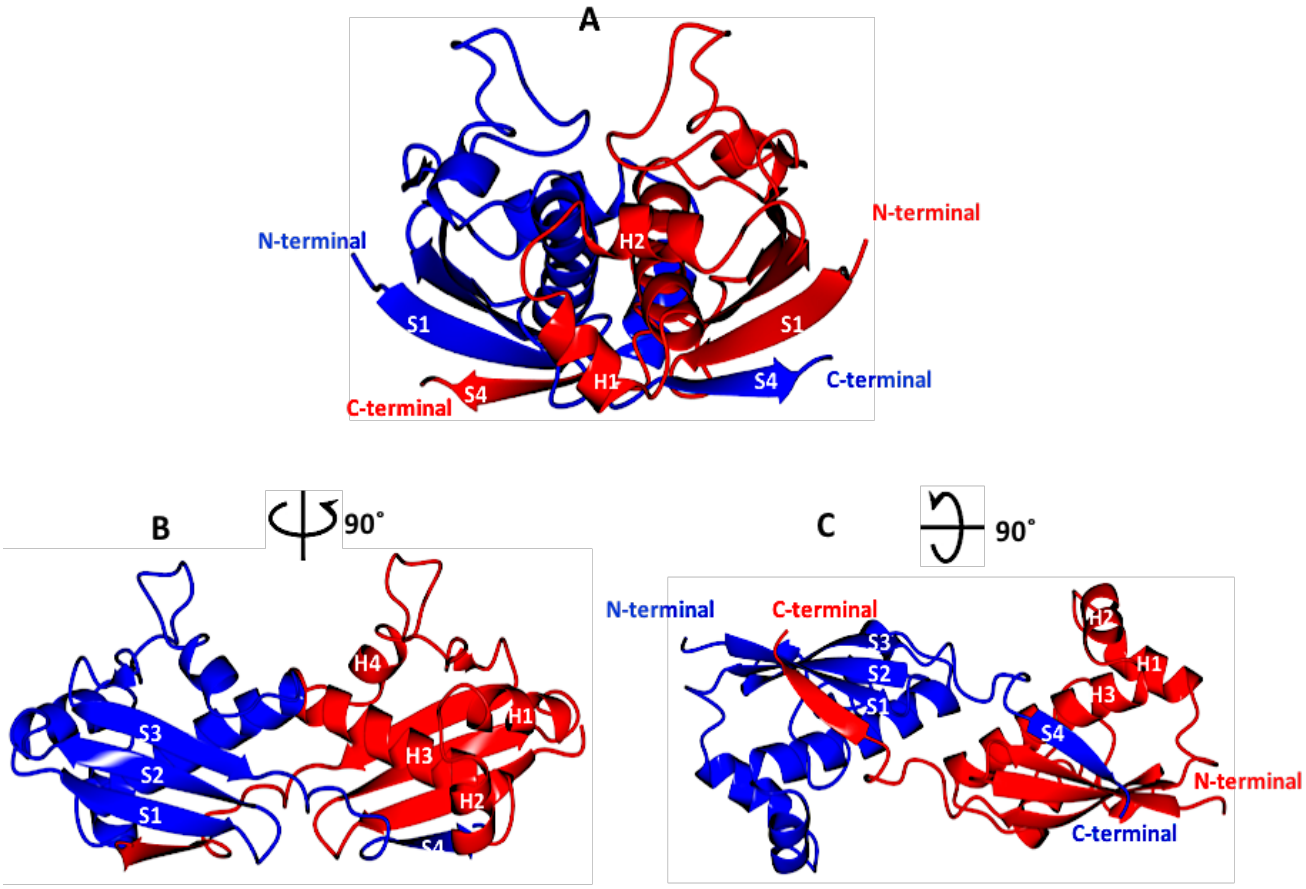


Figure 4.31. Dimeric form of Ea10 structure.

Structure of the Ea10 protein at 2.3 Å. Ea10 forms a dimer, each subunit contains four helices characterized by a long α-helix (H3, 20 residues), and three antiparallel beta-strands. The C-terminal region of the polypeptide from each monomer creates the fourth strand (B4) in a β-sheet with strands from the other monomer. It is shown in different orientations and color (blue and red) to see the interaction between the subunits: the front view A; the side view B; and the top down view C.

4.24 Function determination of Ea10 protein.

A structural comparison of the Ea10 model from the X-ray diffraction data with all other entries in the Protein Data Bank (PDB) using the Dali server was performed. The Dali server suggested a notable similarity with part of the Qbeta-replicase core complex (PDB: 4Q7J). The Qβ-replicase is an RNA-dependent RNA polymerase from phage Qβ and its core complex formation involves the hijack of *E. coli* host EF-Tu and EF-Ts

ribosome associated proteins (Figure 4.32). The overall alignment from the Dali server has an RMSD of 4.3 Å over 111 residues and a Z-score of 4.2. Examination of the structure showed the alignment of Ea10 to a part of the *E. coli* EF-Ts, which is a 281 residue protein and is the guanine exchange and elongation factor that promotes GDP released by EF-Tu. This region specifically has an RMSD of 1.19 Å and it showed a potential alignment for the three anti-parallel β-strands of Ea10 with the EF-Ts structure (I4-Y11 with N59-E66, K14-D22 with Y69-N76, I101-D107 with N130-E138 of Ea10 with EF-Ts respectively), followed by a partially aligned part of the large helix (H3) (L52-Q61 of Ea10 with F91-G102 of EF-Ts) (Figure 4.33). This result might suggest a role for Ea10 protein in RNA metabolism rather than as a DNA binding protein.

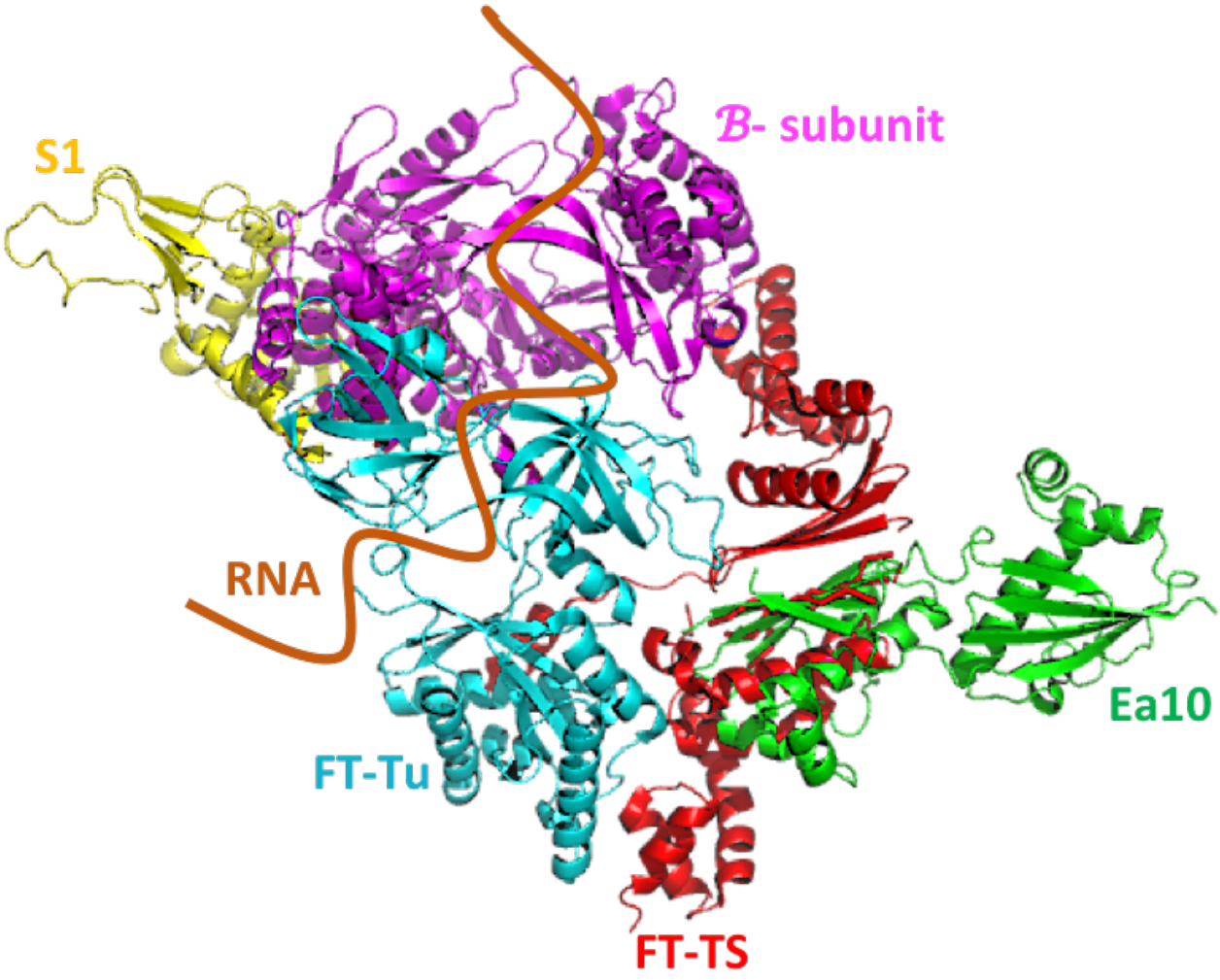


Figure 4.32. Structure suggestion of Ea10 by Dali server.

Structural alignment for the Ea10 Dimeric structure (green) with Qβ-replicase complex (different colors): EF-Ts (red), EF-Tu (cyan), ß- subunit (purple) and N-terminal domain of S1-independent RNA polymerase (yellow) protein structures suggested by the Dali server. The overall RMSD was 4.3 Å over 111 residues.

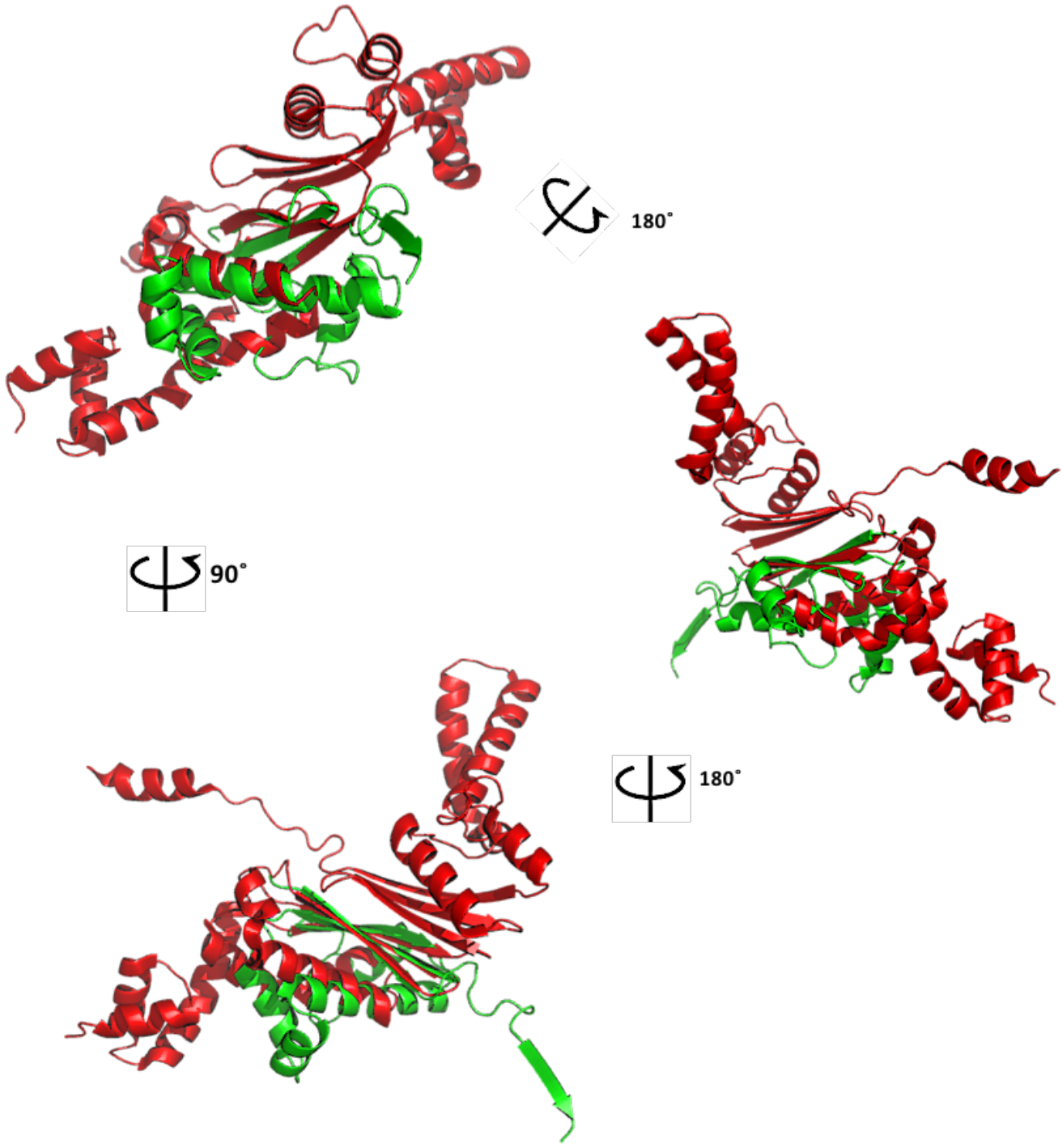


Figure 4.33. Superimposition of Ea10 with EF-Ts protein.

Ea10 monomer structure (green) aligned to a part of the 281-residue EF-Ts protein (red). The aligned region has an RMSD of 1.19 Å and it showed a potential alignment for the three anti-parallel β-strands of Ea10 with the EF-Ts structure and a partial alignment to a part of the large helix H3.

4.25 Investigation of the DNA binding function of Ea10 protein

The Ea10 protein was considered as a DNA binding protein based on unpublished fluorescent anisotropy data from Dr. Gary Sharples (University of Durham), and its gene is under the control of the same promoter as other Red pathway genes (Section 1.9). Attempts were thus carried out to confirm the ability of the protein to bind to DNA that may help in the discovery of the Ea10 protein function.

4.25.1 Electrophoretic Mobility Shift Assay (EMSA)

This assay was carried out in order to prove the ability of the Ea10 protein for DNA binding activity by using concentrations of 2.2 to 4.4 mM of single strand DNA and double strand DNA in gel. The oligonucleotides were used in different sizes and sequences (Table 2.5, section 2.19). Initially, short length DNA substrates were designed with hopes of use in future crystallization experiments, which can be difficult because of the flexibility of longer length DNA substrates. Dna1 (5' – CACCAA - 3'), was used as a single strand probe, and Dna2 (5' – GCTAGC - 3') was used to prepare a double strand probe. A concentration of 1.1 mM of the Ea10 protein was used. Binding reactions were incubated for 60 minutes at room temperature and then resolved through 12% v/v native polyacrylamide gels (Section 2.19.2). No band shift was detected in the native gel when either DNA substrates were added to the Ea10 protein (Figure 4.34).

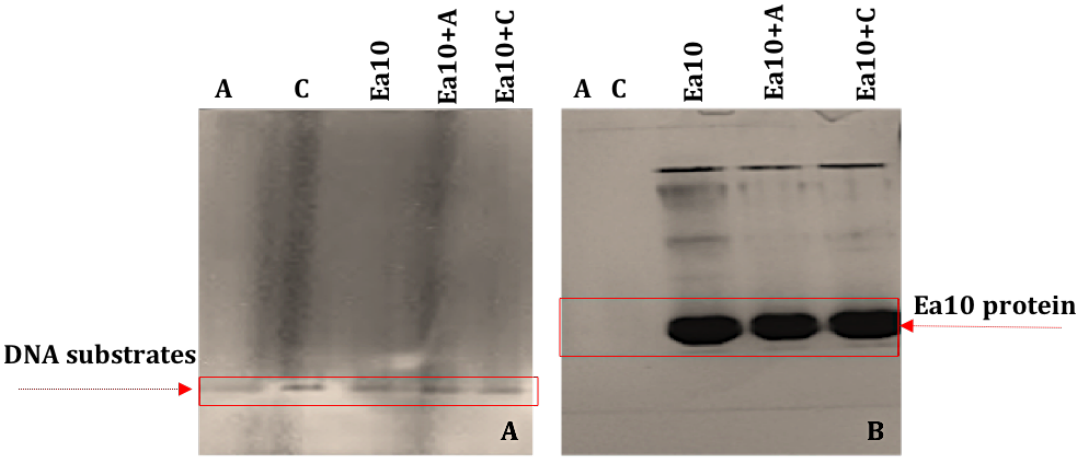


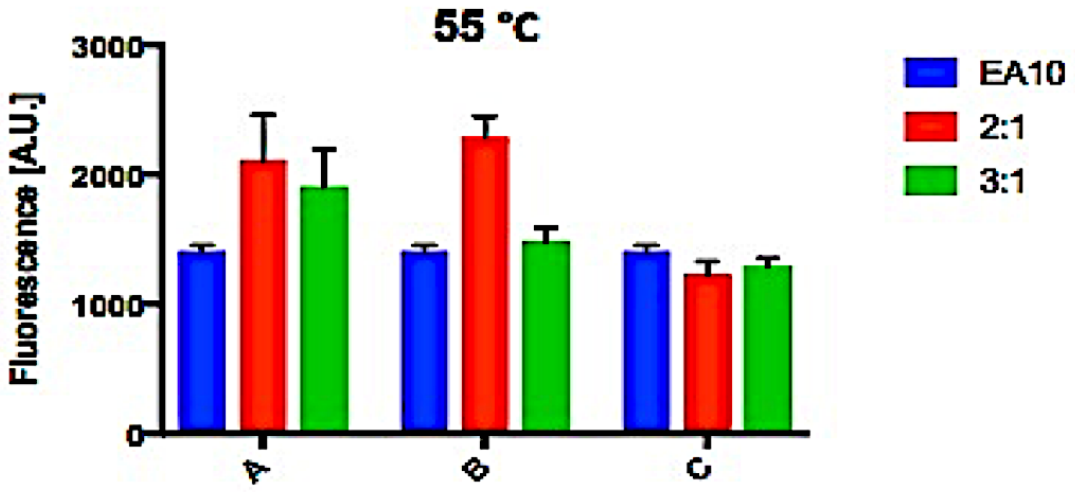
Figure 4.34. Electrophoretic Mobility Shift Assay for the Ea10 protein.

The figures show a 12% polyacrylamide gel, which was treated once with gel red dye to see the running place of DNA bands (A), and once with Bromothymol blue to see the protein bands (B). The result showed no DNA shifting in any of the Ea10-DNA complex bands in comparison with the DNA substrates bands (A and C) under UV. light that indicates no binding to Ea10. **Substrates:** A = 6 base of single strand DNA, C = 17 base of double strand DNA

4.25.2 Thermo-shift assay

The same DNA substrates as used in the EMSA (Section 2.19.1) were carried out to check the affinity of Ea10 protein for DNA substrates by measuring the shift of the denaturation temperature for the protein after adding the DNA substrates. In order to obtain an optimal signal to noise for the fluorescence, different titrations of the DNA substrates (1:1, 2:1 and 3:1 of ligand to protein) and buffers were used to determine the appropriate condition for the protein. The results revealed that the 2:1 ratio in Tris-HCl buffer pH8 was the best condition based on the statistical analysis using triplicate sample data of the highest fluorescence score (55 ˚C) in two-way ANOVA multiple comparisons analysis, which gave a significant score of P value (<0.0001) (Figure 4.35 A). The results also showed that no shift in the temperature was detected in this experiment when either DNA substrate was added to the Ea10 protein, supporting the results obtained from the EMSA that Ea10 does not bind these specific DNA substrates (Figure 4.35 B).

A



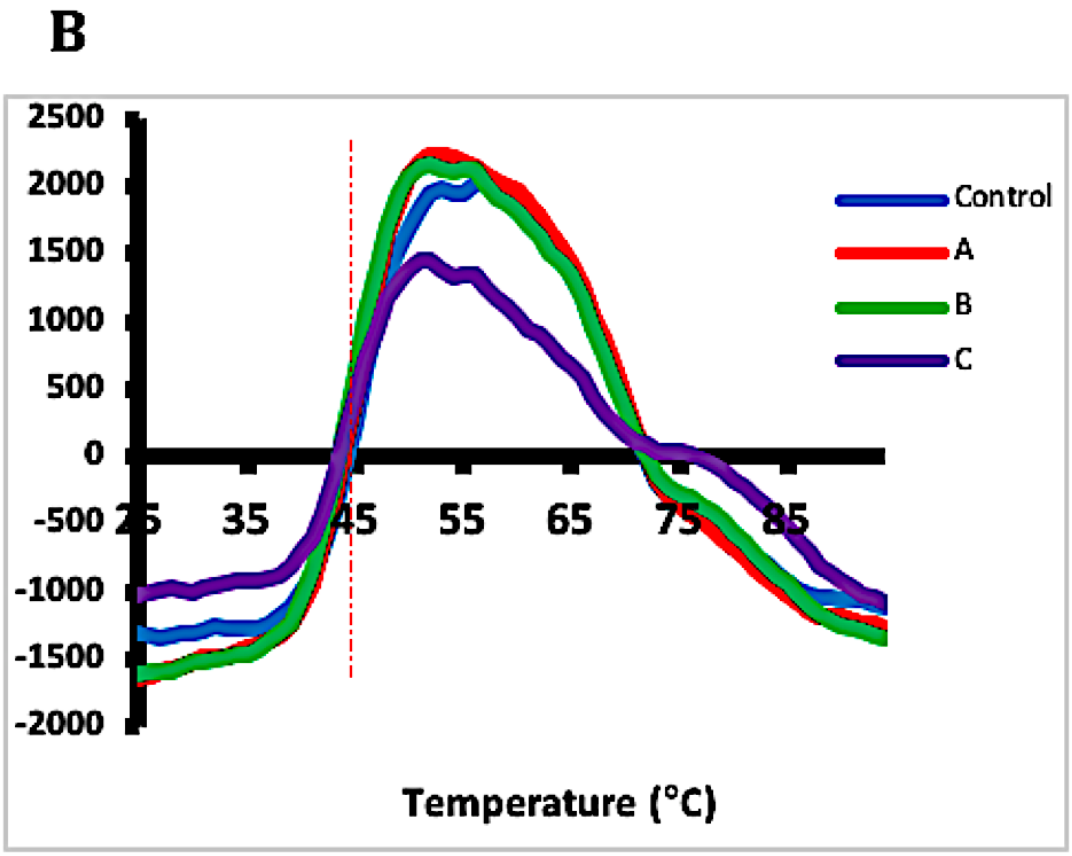
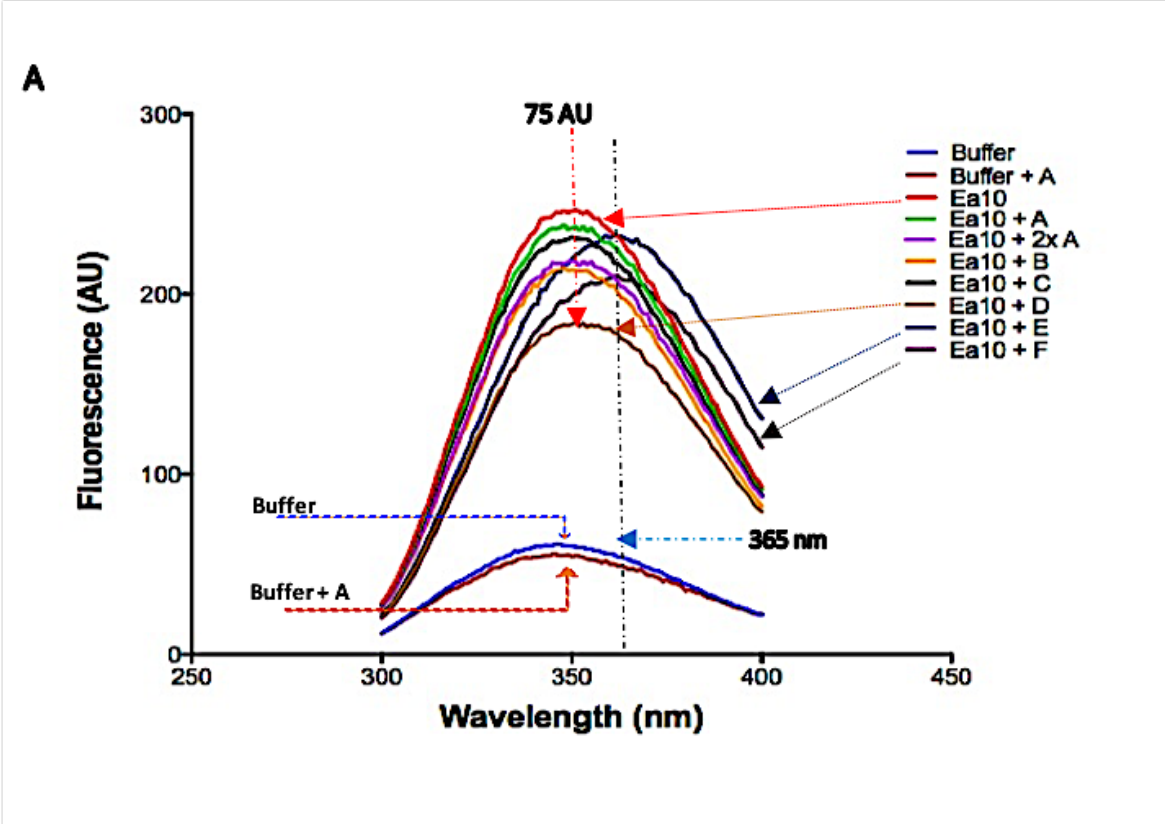


Figure 4.35. Thermal Shift assay for Ea10 protein.

The graphs of the thermal shift assay showed **A)** Statistical analysis for determining the best DNA substrate to Ea10 protein ratio based on the strength of the fluorescence signal as judged by a significant P value (P< 0.0001) using a two-way ANOVA multiple comparison, which suggested a 2:1 ratio at 55 ˚C**. B)** Measurements at a DNA-protein 2:1 ratio that showed no shift in temperature at the midpoint temperature (45 ˚C, red dashed line) of the protein Ea10, indicating no binding to Ea10 for either of the DNA substrates tested. **Substrates: A** = 6 base of single strand DNA. **B** = 6 base of double strand DNA. **C** = 17 base of single strand DNA, Control= only protein

4.25.3 Tryptophan Fluorescence assay

The capability of Ea10 protein to bind to DNA was also tested using the same DNA substrates as used in the thermo-shift assay. The tryptophan fluorescence assay relies on the excitation and the emission of tryptophan in the native protein by measuring the quenching level of the fluorescence peaks (Section 2.19.3). Substrates composed of 60 bases of single and double strand DNA were also used in this assay (Table 2.5). The results illustrated a possible interaction between the Ea10 protein and all of the DNA substrates based on the varying quenches of the fluorescence peaks compared with the peak of Ea10 (Figure 4.36 A). It was also found that the 60 bp single strand DNA (D) scored a noticeable quenching, of approx. 75 AU. A wavelength shift at 365 nm was reported for the peaks of the 60 bp DNA substrates (E= single strand DNA, and F= double strand DNA). The results were confirmed by a statistical analysis (two-way ANOVA multiple comparisons) using the triplicate sample data of the fluorescence peaks at 350 nm wavelength, and the results came up with a significant score of P value (<0.0001) for all of the DNA substrates supported with the small error bars (Figure 4.36 B).



B

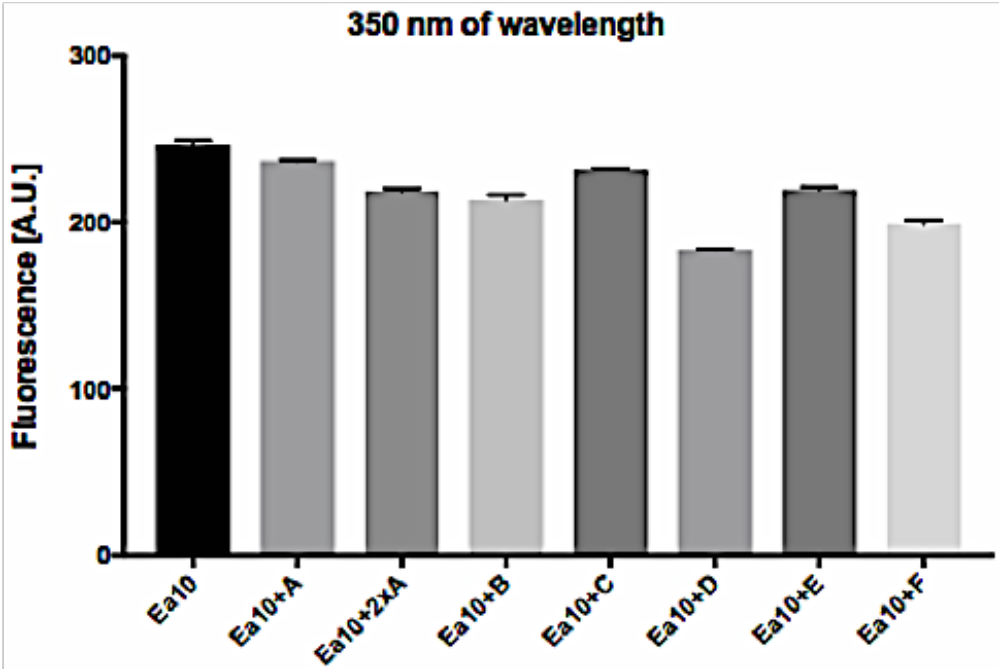


Figure 4.36. Tryptophan fluorescence assay for Ea10 protein.

A) A decrease in the fluorescence signal is observed for substrate D, which might indicate a possible weak DNA interaction with Ea10 for substrate D, which scored a 75 A.U of quenching at 350 nm (red dashed line). However, this was not replicated by substrate E, which is also 60 base single strand DNA. A wavelength shift at 365 nm (blue dashed line) was found in the peaks of 60bp DNA substrates (E and F). **B)** A bar plot showing the fluorescence of each substrate plus Ea10 at the peak at 350 nm. The two-way ANOVA multiple comparison revealed a significant difference between each substrate with Ea10 compared to Ea10 alone ($P < 0.0001$). **Substrates**: **A** = 6 base of single strand DNA, **B**= 6 base of double strand DNA. **C** = 17 base of double strand DNA. **D** = 60 base of forward single strand DNA, **E** = 60 base of backward single strand DNA, **F** = 60 base of double strand DNA.

4.25.4 Electrostatic potential analysis for the Ea10 structure

The surface charge of the Ea10 protein was analysed in the CCP4mg program. A qualitative model for the electrostatic surface potential of the Ea10 dimer showed the protein is predominantly negatively charged on its surface, indicating that it is unlikely to bind DNA (Figure 4.37). However, it is possible that a negatively charged area could bind to a positively charged ion, which could in turn help interaction with DNA.

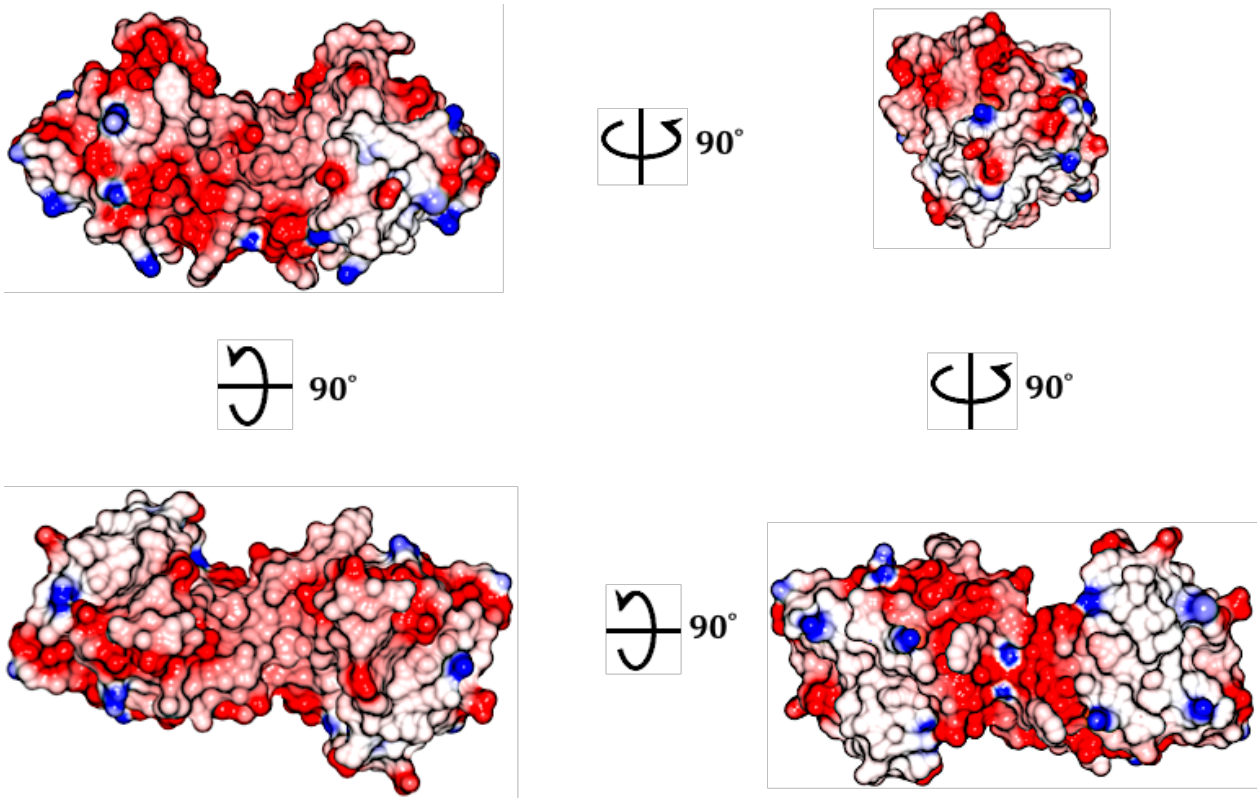


Figure 4.37. Electrostatic surface of the Ea10 structure.

Electrostatic surface potential of the Ea10 dimer in different views shows that this protein structure is predominantly negatively charged on its surface (red areas), indicating that it is unlikely to bind DNA.

4.26 Discussion

4.26.1 Possible DNA binding function of Ea10

Ea10 as a DNA binding protein has previously been hinted at by unpublished anisotropy data and its location in an operon containing Red pathway proteins. Based on this, initial experiments investigated the DNA binding function of Ea10 further. There are not many conserved members of the Ea10 family but it does have a domain assignment (DUF2528; PF10800) via http://pfam.xfam.org/family/PF10800#tabview=tab4. The conserved 'FW' pairing appears to be on the surface and could interact with bases in ssDNA as seen for *E. coli* SSB, which has many interactions via aromatic residues (Figure 4.38).

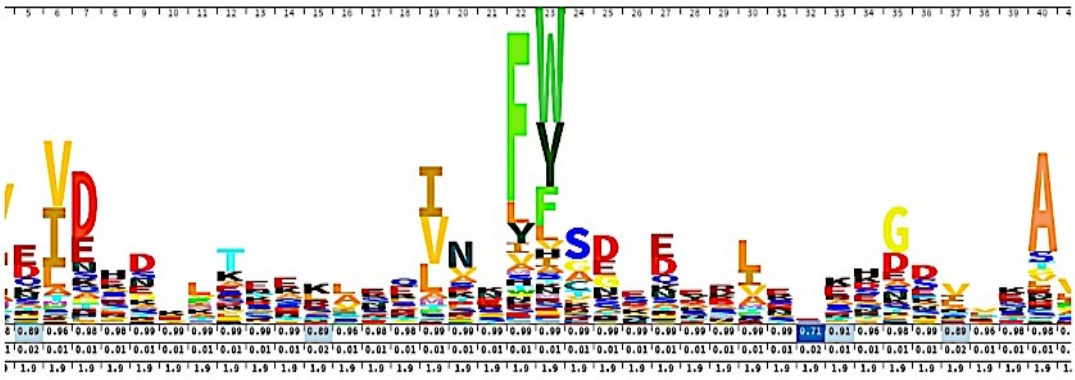


Figure 4.38. The conserved HMM logo for the Ea10 structure.

The conservation of residues F22 and W23 on the surface of Ea10 structure that might possibly interact with ssDNA as seen with *E. coli* SSB.

The first attempt to understand the ability of the protein to interact with DNA was the EMSA assay. No shift in the position of the DNA was observed after adding Ea10, indicating that Ea10 does not bind to the DNA substrates used. Similarly, no major shift in the melting temperatures were observed in the thermo-shift assay for any of the substrates, again indicating a lack of binding of Ea10 to the DNA. The tryptophan fluorescence assay did show a possible interaction between a 60 base ssDNA and Ea10, where noticeable fluorescence quenching of the peaks was observed between that and the control (Ea10). However, this fluorescence interference alone does not prove the

binding activity for Ea10 protein. It could occur simply due to substrate size or concentration in the sample solution. Therefore, using different titrations of substrates might help in confirming the binding activity for Ea10. From these experiment, there is not yet any clear evidence for Ea10 as a DNA binding protein.

Additionally, electrostatic surface calculations of the solved Ea10 structure revealed that the protein has a predominantly negatively charged surface, meaning that it may be unlikely that Ea10 has the ability to bind to a negatively charged substrate (*i.e.* DNA). One consideration, however, that was not taken into account in these experiments is that Ea10 may have metal ion binding sites that are necessary for its function. Perhaps, positively charged metal ions are able to bind to specific sites in Ea10, which would alter its surface charge, thereby allowing it to bind to DNA in this manner. Many proteins work in such a way, requiring metal cations as a cofactor to help it carry out its function. For Rap endonuclease (another protein in the Red Pathway system), it was found that its activity as a resolvase for Holliday junctions increased in the presence of magnesium ions (Sharples et al., 2004). Therefore, the magnesium ion may also be useful to improve Ea10 DNA binding.

A BLAST search also showed an alignment of a small section of the Ea10 protein sequence with a specific sequence of an apo-aequorin protein from Aequorea Victoria (pdb: 1SL8) (Deng et al., 2004), where conserved residues were specifically used to form a pocket to bind to a calcium ion (Figure 4.39 A). Aequorin is a protein that is used for bioluminescence in jellyfish, which is activated by calcium binding. Although bioluminescence is not a property that lambda phage has, it is possible that this specific site could bind a calcium ion for another function like binding DNA. However, studying the sites of the conserved residues using the determined structure of Ea10 reveal that they do not form a pocket like in Aequorin, but instead, are found in the beta-sheet region, making it an unlikely place to bind a calcium ion unless some sort of conformational change can take place. Therefore, the alignment may just have occurred by chance (Figure 4.39 B and C). Nonetheless, further work could repeat the assays already carried out, but using a range of different metal cations (*e.g.* magnesium, calcium, zinc, manganese, iron ions) at different concentrations to see if DNA binding could be improved.

For these experiments, only a limited number of DNA substrates were used. Therefore, using varying substrates of DNA in different sizes and structure like junction-type substrates could help to find the specific substrate for its function. Conditions of the assay might also be varied, for example, different pH and temperature. In reality, an infinite number of different assay conditions and DNA substrates could be tested, so they should be chosen strategically for further experiments.

A

Alignment statistics for match

Score	Expect	Method	Identities	Positives	Gaps
26.9 bits(58)	5e-05	Compositional matrix adjust.	12/23(52%)	14/23(60%)	0/23(0%)

```
Query    96   DGSEGIRITDIDTSGIFDSDDMT  118
              D  E  R+ DID SG  D D+MT
Subject  146  DCEETFRVCDIDESGQLDVDEMT  168
```

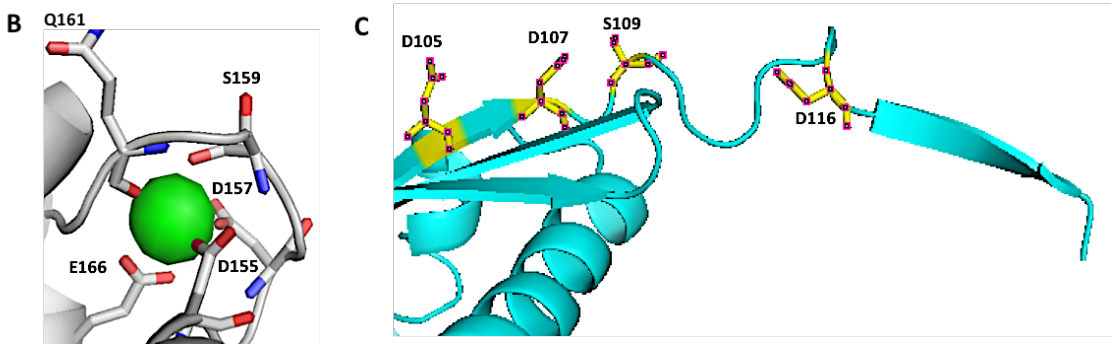


Figure 4.39. Comparison for conserved residues between Aequorin and Ea10.

A) Sequence alignment between Aequorin and Ea10 with an expectation value of 5e-05. Possible conserved residues in Ea10 are shown with the calcium binding region of Aequorin.

B) Residues in Aequorin that form the calcium ion binding pocket (Deng et al. 2004).

C) Corresponding residues in Ea10 to the residues found in (B). The residues are found along the 3rd to 4th beta strands of Ea10, and do not appear to form a possible calcium ion binding pocket.

4.26.2 Possible RNA binding function of Ea10

Since the DNA-binding function of Ea10 is not evident, other potential alternative functions of Ea10 should be investigated. For example, the Ea10 protein may possibly have affinity to interact with RNA instead of DNA. The Dali server suggested a notable identity for the Ea10 structure with Qb-replicase core complex, which is an RNA-dependent RNA polymerase and involves the hijack of host EF-Tu and EF-Ts ribosome associated proteins. However, the Ea10 is aligned to part of EF-Ts which is not in contact with EF-Tu or indeed with the replication protein in the Q-beta replicase core. Additionally, as discussed previously with DNA, RNA is also a negatively charged molecule, which would pose the same problems with binding to Ea10 as a predominantly negatively charged protein. Thus, if Ea10 is involved with RNA metabolism, it would suggest a role not directly involved in binding the RNA.

Almost equally good Z-scores for matches were achieved by the Dali server for a soluble bacterial homologue of one domain of stomatin (a eukaryotic membrane protein) and for the "shoulder" domain of a subunit of rat liver vault ribonucleoprotein (a 39-identical residues in the subunit chamber of unclear function). Therefore, although a match was found to part of an RNA-binding protein, this may not have much significance due to the seemingly random matches of Ea10 structural motifs to a range of other proteins with differing functions. These matches in Ea10 may just represent common motifs found across many proteins. Swapping strands in the Ea10 structure to form a dimer is still curious, and it is not a unique feature but interesting. EF-Ts also dimerizes but in a completely different manner.

4.26.3 Possibility of protein-protein interaction.

Ea10 protein may possibly have an ability to interact with other proteins during the phage life cycle either from the *E. coli* host or from lambda itself. A protein interaction network for phage lambda proteins with *E. coli* host proteins was developed based upon a yeast two hybrid screen (Blasche, 2013) and the results for Ea10 can be found at https://www.ebi.ac.uk/intact/ with search entry C6ZCV8. A number of *E. coli* proteins were found to interact with Ea10 protein including the ribosomal proteins RpmA and RpsG, the ribosome modulation factor (Rmf), primosome replication protein (PriC), macrodomain Ter protein (Matp), NAD-dependent protein deacylase (Npd) and

oxidative damage regulatory proteins (SoxS). These findings remain to be validated by other approaches. Therefore, various experiments could be employed, for example, pulldown assays with *E. coli* protein lysates and tagged Ea10 or production of a complex between Ea10 protein with purified samples of the suggested *E. coli* proteins. Alternatively, the transcriptome of *E. coli* might be investigated if recombinant Ea10 is overexpressed and expression of the host genes examined for those which are upregulated or down regulated.

5 Chapter 5. Introduction for DnaD protein of Bacillus subtilis

5.1 General background of Bacillus subtilis

B. subtilis is considered scientifically as one of the best classified gram-positive bacteria. It is an aerobic, endospore-forming, rod-shaped bacterium, and commonly present in soil, water sources and in plants. It was named at its first discovery as *Vibrio subtilis* by Christian Ehrenberg in 1835, and then the name was changed *to Bacillus subtilis* in 1872 by Ferdinand Cohn (Claus & Berkeley 1986). *B. subtilis* is among the most common *Bacillus* species., which have been exploited for protein production studies and for development as a host for the production of heterologous proteins owing to their tremendous fermentation characteristics, high product yields (20 to 25 gram per litre) and the entire loss of toxic by-products (van Dijl and Hecker, 2013). In particular, *B. subtilis* and its close relatives are an essential source of industrial enzymes (such as amylases and proteases), and a great commercial interest in these bacteria has emerged from their capability to produce these enzymes at gram per litre concentrations (Kunst and Al., 1997). *B. subtilis* was demonstrated to be an appropriate model organism for Systems Biological analyses on gene regulation throughout different conditions (Buescher et al. 2012; Nicolas et al. 2012). Furthermore, *B. subtilis* is also used in the production of Natto, which is a traditional Japanese dish of fermented soya beans (Kunst and Al., 1997).

5.2 Biological interest of *B. subtilis*

Bacillus spp. possess distinct properties that make them good nominees as biological control agents. They are well known for their antibiotic production with antagonistic action against some bacterial and fungal pathogens (Loeffler et al. 1986; Krebs et al. 1998). The majority of the antibiotics obtained from *Bacillus* spp. have been dipeptides or cyclic peptides with low molecular weight (Loeffler *et al.,* 1986; Nakano & Zuber 1990). In addition to antibiotic production, *Bacillus* species are among the most common bacteria found to colonize plants endophytically (Lilley et al. 1996; Mahaffee & Kloepper 1997). Possibly their endophytic ability plays a role in the biocontrol of vascular plant pathogens (Williams, 1980).

5.3 Structural characteristics

Bacillus species have been identified and characterized, and the sequence of the 16S *rRNA* gene was analysed from several *Bacillus* species (Ash *et al.,* 1991). In the study of Ash and colleagues, five highly different lines within the genus were recognised based on nucleotide sequence. *B. amyloliquefaciens*, *B. subtilis* and *B. pumilus* are among the recognised lines, which all belong to group 1 with another 27 species (Ash et al., 1991).

The *B. subtilis* genome contains 4,100 protein-coding genes with a size of 4,214,810 bp. Fifty-three percent of the protein-coding genes are characterized, whereas a quarter of the genome resembles gene families that have been significantly enlarged by gene duplication (Kunst and Al., 1997).

5.4 DNA replication system

DNA replication is a central process in the life cycle of all organisms. It is divided into three main stages known as initiation, elongation, and termination. Chromosomal replication relies on the initial assembly of replication forks at specified origins called the replication origin (oriC) and on the reassembly of the ongoing replication forks in case of their arrest (Zhang et al., 2005). Afterwards, the other replication apparatus proteins are recruited for DNA replication such as the replisome, which is necessary to prime and expand DNA synthesis using two polymerases (Figure 5.1). This replication initiation can be performed by various sets of initiation proteins, describing diverse strategies of helicase recruitment and loading (Grabowski & Kelman 2003; Méndez & Stillman 2003).

Loading the replicative helicase onto the DNA is the main role of the initiation of DNA replication. The initiation process, priming, recruits cooperatively a number of primosomal proteins that are responsible for employing the replicative helicase (Zhang et al., 2005).

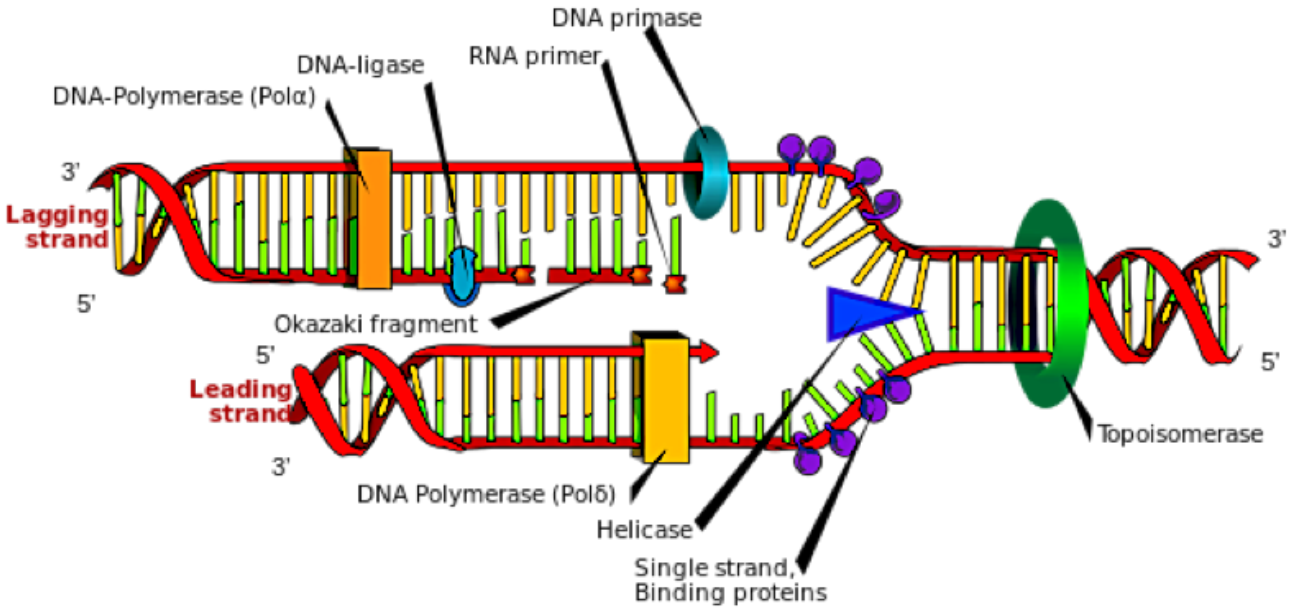


Figure 5.1. Bacterial replisome.

The figure shows the DNA replication of *B. Subtilis*. In the initiation stage, two polymerases are recruited, one moving clockwise (leading strand, yellow) and another moving anti-clockwise (Lagging strand, Orange) throughout the chromosome. The replisome contains a large group of proteins working cooperatively to transcribe the DNA. Three main functions are essential to duplicate the chromosomal DNA: 1) Unwinding the existing DNA strands by Topoisomerase followed by Helicase to break the hydrogen bonds between base pairs to produce single-stranded DNA. 2) Synthesizing complementary single-stranded DNA by polymerases: DNA polymerase (pol III) reads one of the two ssDNA (leading strand) to create dsDNA. 3) Replicating the lagging strand. Since the DNA polymerase (pol III) is able only to build DNA in a 5'-3' orientation and the Lagging strand is 3'-5' fashion, DNA primers are generated made from short ds-DNA ~2kbp from the 5' end of the replicated DNA that facilitates DNA polymerase activity to continue. Since this pathway creates long sections of ssDNA, it needs to be protected via recruiting ssDNA binding proteins. These pieces are identified as Okazaki fragments, which are joined by DNA ligases (G. Duggin 2006; Marians 2008).

5.5 Mechanism of replication pathway

Extensive genetic and biochemical studies in the Gram-negative bacterium, *E. coli,* have solved two mechanisms for activating and reactivating DNA replication. The first takes place at the oriC of the circular chromosome and confirms the precise timing of replication within the cell cycle (reviewed by (Marsin et al., 2001). This highly controlled initiation process is activated by the DnaA protein, which specifically recognizes oriC. The second reactivation process is initiated by the PriA protein, which boosts replication restart by binding to specific DNA structures (Marians, 1999). In *B. subtilis*, as in *E. coli,* the initiation and re-initiation of chromosomal DNA replication are initiated either at the single chromosomal origin oriC or at the arrested replication forks, and require the presence of all the DnaB, DnaD and DnaI proteins. These proteins were found only produced in the Gram-positive bacteria, which have low G+C bases content (Bruand et al., 2005).

Bruand and co-workers suggested a model of the different pathways existing in *B. subtilis* for loading the DnaC helicase and DnaG primase onto DNA (Bruand et al., 2001). Two major pathways are present, one originates at oriC and requires DnaA (pathway 1) and another operates at ssiA (single-strand initiation site located on the lagging strand template of bacterial plasmids, -150 bp down-stream from the replication origin) and at forked structures (D-loops or stalled forks), and requires the primosomal protein PriA (pathway 2) (Bruand et al., 2001). Both pathways (DnaA and PriA) require a set of three intermediate proteins (DnaB, DnaD and DnaI) to recruit and load the DnaC replicative helicase at the chromosomal origin oriC or at arrested replication forks (Marians 1999; Marsin et al. 2001). The replication initiation proteins associate stably with oriC in a defined order, in which DnaA binds first, followed by DnaD and then DnaB, and finally the DnaI mediates loading of helicase (Smits et al., 2010). Similarly, in the PriA pathway, PriA, DnaD and DnaB were suggested to assemble the nucleoprotein sequentially (Marsin et al., 2001). Nevertheless, a third pathway called the 'DnaD pathway' was suggested by (Bruand et al., 2001) that comprises the same DnaB, DnaD and DnaI as the other two pathways, and is similar to the *E. coli* PriC pathway (Sandler, 2000). Using yeast two-hybrid analysis, *B. subtilis* DnaD was revealed to interact with DnaA and DnaD itself. A mutated DnaD, which is a *danD23* mutant (thermos-sensitive mutant) and has a single mutation at residue A166T in the C-terminal domain of the DnaD protein, was

found to lose the ability to interact with DnaA but was capable of interacting with the wild-type DnaD (Ishigo-oka et al., 2001).

In *E. coli*, the initiator of replication, DnaA, binds to specific sequences (DnaA boxes) within oriC and opens dsDNA at AT-rich sequences in the region of these boxes (Figure 5.2). DnaA protein also interacts with the helicase, DnaB (Seitz et al., 2000), and is essential for prepriming complex formation on DNA (Sutton et al., 1998). DnaC creates a stable hexameric complex with DnaB, which then provokes ssDNA binding activity of DnaC (Learn et al., 1997).

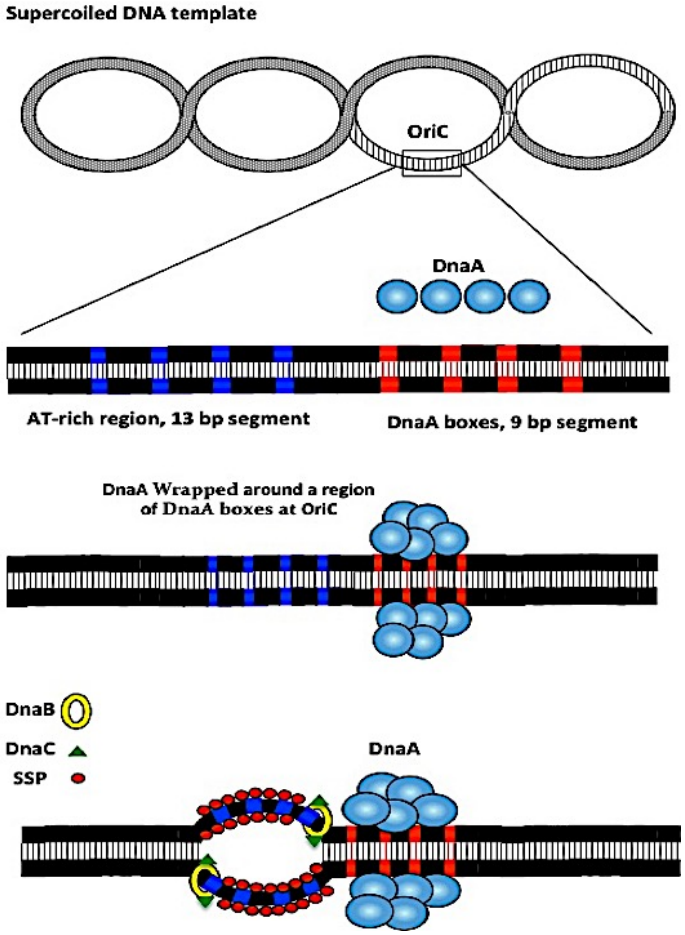


Figure 5.2. Initiation stage of DNA replication in *E. coli*.

The *oriC* region (the origin of replication) in double-stranded DNA is processed through the function of the initiator protein, DnaA (blue), which recognizes and binds 9bp repeats in *oriC* (consensus sequence of 5'-TTATCCACA-3') to form a complex of negatively supercoiled DNA wrapped around a central core producing single-stranded DNA substrates. DnaA protein guides a DnaB (helicase) and DnaC (gyrase) complex into a melted region to form the prepriming complex. DnaB, a helicase unwinds the DNA in the prepriming complex in both directions. SSB (single strand binding proteins) and gyrase proteins are necessary for DnaB activity. The helicase/gyrase further unwinds the helix to allow entry of the RNA polymerase, which activates the primase to begin RNA primer synthesis.

These primosomal effectors i.e. DnaD, DnaB and DnaI are indispensable for cell sustainability (Bruand et al., 2005).

The DNA replication fork is activated by nucleoprotein clusters (primosomes), which are primarily responsible for loading the replicative helicase onto ssDNA ((Marsin et al., 2001). In *B. subtilis*, genetic analysis has shown that the primosomal proteins (DnaB and DnaD) have no clear homologues in *E. coli*. (Ogasawara et al. 1986; C Bruand et al. 1995; C. Bruand et al. 1995). However, the DnaI sequence is marginally similar to DnaC of *E. coli* (Koonin, 1992). Similarly, the recognition of PriB, PriC and DnaT proteins of *E. coli* could not be found in the *B. subtilis* genome (Kunst and Al., 1997).

There is an essential requirement for the presence of DnaA to initiate the DNA replication by binding to specific sites within the oriC. However, DnaA also binds to several sites throughout the chromosome as a transcription factor outside the oriC region (Smits et al., 2011). It was found that the DnaD and DnaB proteins are essentially required in the DNA replication for loading the replicative helicase outside the oriC region without a need of DnaA. However, chromatin immunoprecipitation analysis revealed a possible association of DnaD and DnaB, which is not the replicative helicase in *B. subtilis*, at several chromosomal sites within *B. subtilis* DNA that are bound by DnaA (Smits et al., 2011). This association requires DnaA with the same order of recruitment as found at *oriC,* but independently of a functional *oriC* and without the need for open complex formation to stabilize the association with DNA. Interaction of DnaB and DnaD at different regions of the chromosome is indicative of a function for these proteins outside oriC and separate from the DnaB and DnaD function in replication restart (Smits et al., 2011).

5.6 DnaD protein

Many genes mediate DNA replication in *B. subtilis*. These genes were determined using thermo-sensitive mutations like the *dnaD23* mutant. This mutant has an effect on the initiation stage of DNA replication in *B. subtilis* via an undiscovered mechanism at an appropriate temperature (Gross et al., 1968); (Karamata and Gross, 1970). It was found to affect both the synthesis of the lagging strand and the assembly site activity of the

primosome which suggests that the *dnaD* gene product plays an important role in DNA replication as an essential part of the *B. subtilis* primosome (Bruand et al., 2005).
In 1970, the *dnaD* gene was located as being next to the *trp* gene cluster in the *B. subtilis* chromosome (Karamata and Gross, 1970), when a yeast artificial chromosome was used to clone a part of the *B. subtilis* chromosome, which includes the *trp* operon, for sequence determination (Azevedo et al., 1993). The *dnaD* gene encodes a 232 amino acid protein. Two ORFs in the bacterial genome following the *dnaD* sequence are perhaps co-transcribed with *dnaD*. A 46% sequence identity with the *E. coli nth* gene (endonuclease III protein) was reported for one of these ORFs. Thus, the involvement of two genes possibly in the same operon was suggested to be a part of DNA metabolism (Bruand et al., 1995).

An additional function of DnaD was suggested based on the production of DnaD in the cell which is estimated at 3,000 to 5,000 molecules per cell (Bruand et al., 2005). An activity of DnaD protein for remodelling DNA structure was discovered (Turner et al., 2004). Although, it operates by an unknown mechanism, it offers a potential connection between DNA remodelling and the initiation stages during chromosome replication (Zhang et al., 2006). Additionally, the abundance of DnaD may imply the protein acts as a stimulator for the topoisomerase I activity to alleviate additional negative supercoiling (Sikder et al., 2001). Direct interaction between these two proteins (DnaD and topo I) was not observed although this could still possibly be correct based on analysis of topo I activity, which can be modified indirectly by single stranded DNA-binding protein (SSB) without a direct interaction (Sikder et al., 2001). Keeping DnaD in the cytoplasm has been found to inhibit unnecessary over-initiation of DNA replication, but its regulated recruitment to the membrane by DnaB triggers DNA replication initiation (Rokop et al., 2004).

5.6.1 DnaD structure

DnaD reveals a complex range of DNA recognition and protein oligomerization activities. It has the capability of binding single stranded and double stranded DNA, and of remodeling supercoiled DNA by removing writhe (Zhang et al., 2008). It was found that DnaD aids the untwisting of supercoiled DNA based on the results of atomic force microscopy analysis (Turner et al., 2004).

Structurally, DnaD protein is composed of two main functional domains: the N-terminal domain approximately 16 kDa (residues of 1–128) and is a DNA- independent oligomerase (Figure 5.3); and a C-terminal domain, which is 14 kDa (residues of 129-232), and is classified as a DNA-induced oligomerase with DNA binding activity (Carneiro et al., 2006). In order to achieve the DNA-remodeling activity by DnaD, the N-terminal domain must be covalently connected with the C-terminal domain in the same molecule (Zhang et al., 2005). It was suggested that the aggregation of the N-terminal domain starts only when the C-terminal domain binds to DNA because the unbound C-terminal domain deactivates an aggregation surface of the N-terminal domain (Schneider et al., 2008). The cooperative interaction between the two DnaD domains induces the DNA-remodeling function due to the create of the scaffold forming activity. The structural determination for the N-terminal domain of DnaD revealed that it has a motif of a winged helix-turn-helix fold. This motif is commonly found to bind DNA in other proteins but in the DnaD N-terminal domain is proposed to assist the formation of an extensive protein scaffold (Schneider et al., 2008). The DNA-binding properties of DnaD were characterized to be restricted to the C-terminal domain (Zhang et al., 2008). A structural study of the N-terminal domain confirmed its ability to aggregate and generate dimerization and tetramerization interfaces comprising a helix-strand-helix (H1 -S1 -H2) unit extended at the N-terminal end. Two beta strands formed additional extensions from the N-terminal domain. A third attached helix at the C-terminal end of the winged helix-turn-helix core was found to be involved cross-tetramer oligomerization within the crystal structure (Schneider et al., 2008).

DnaD is reported to behave as a dimer in solution that forms two different conformations based on the relative orientation of its domains. One of these conformations is capable of binding DNA and becomes stabilized (Figure 5.3). This DnaD-DNA complex is proposed to mediate the formation of a stable tetramer complex by attracting another dimer molecule. The final DnaD complex generates a seed to promote formation of scaffold activity (Schneider et al., 2008).

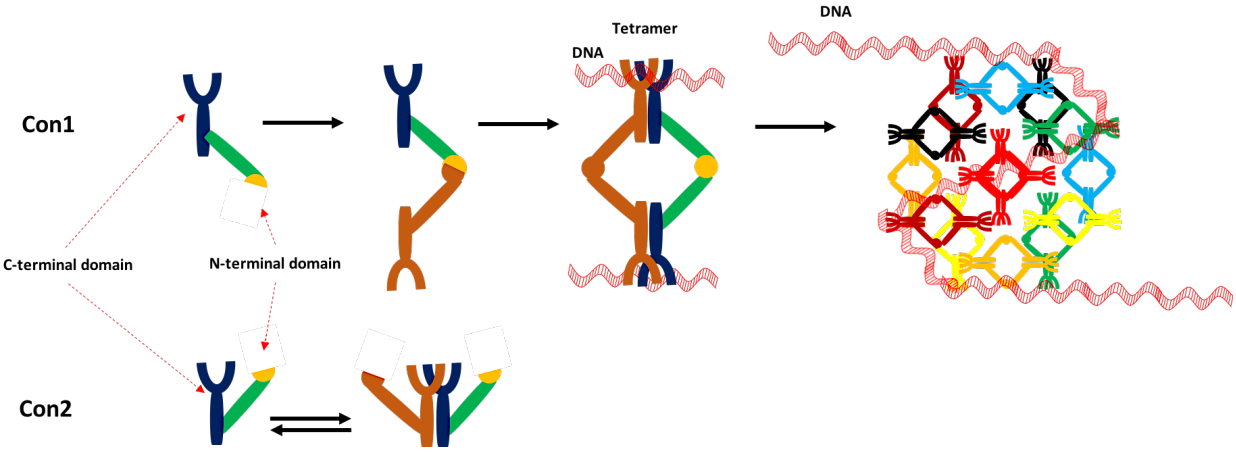


Figure 5.3. Scaffold formation of DnaD structure.

Formation stages of Scaffold activity relies strictly on the relative orientation of the two DnaD domains (to bind DNA. In solution, native DnaD is found in two dimeric conformations (Con1 and Con2). It is proposed that only Con1 is capable to create a tetramer due to a high affinity to bind DNA in comparison with Con2. The presence of DNA increases the stability of Con1 to form a tetramer by attracting another DnaD dimer. The tetramer: DNA complex is suggested to act as a nucleus for a scaffold formation.

5.6.2 Importance of DnaD protein

Recently, studies revealed that the DnaD and DnaB proteins have individual or combined activity for ssDNA interaction in the initial stages of chromosomal DNA replication (Bruand et al., 2005). Furthermore, a central structural role for DnaD has been demonstrated in the multiprotein assemblies built during the two fundamental processes of initiation of DNA replication (DnaA and PriA) (Bruand *et al.,* 2005). Purified DnaD was found to be highly active in binding ssDNA (Marsin et al., 2001). This activity was increased in the presence of DNA-bound PriA that indicated functional interaction between PriA and DnaD. It was also found that the presence of DnaD protein increases the ssDNA binding affinity of DnaB (Bruand et al., 2005). Purified DnaD has been reported to interact physically with PriA and with DnaB. It was found that the loading of SSB onto ssDNA, which is necessary for the replicative helicase loading, depends essentially on the interaction between DnaD and DnaB proteins with ssDNA (Bruand et al., 2005). Adding to the roles they play in the loading of the replicative helicase, the DnaA function may also be modulated by DnaD and DnaB proteins (Smits et al., 2011).

In addition to its role in the initiation of DNA replication as an essential component in the primosomal cascades, DnaD forms a variety of large nucleoprotein structures that can affect universal DNA architecture, and produce an open circular form by modifying a supercoiled plasmid, this is a similar activity as a HU protein in *E coli* (Turner et al., 2004). Functionally, the stability of interaction between the helicase and DnaI can be distracted by DnaD perhaps during helicase loading (Turner et al., 2004). According to the mentioned properties of DnaD, it suggests that the DnaD protein is an important component of the association between the initiation of DNA replication and the nucleoid reorganization in *B. subtilis* (Turner et al., 2004).

5.6.3 Function of DnaD protein

DnaD was suggested to convert the twisted DNA structure to an untwisted form (Zhang et al., 2006). Furthermore, a probable association was found for DNA remodelling with a considerable untwisting of the duplex based on atomic force microscopy results, which proved the ability of DnaD protein to produce a twist by altering all of the writhe of supercoiled plasmids (Zhang et al., 2005). The substantial energy required for opening up supercoiled plasmids without nicking can be compensated by DnaD. DnaD molecules hold the plasmid tightly by formation of a circular protein scaffold (Zhang et al., 2006). The sum of three distinct oligomerization and DNA-binding activities were proposed to generate the DnaD remodelling activity depending on two connected but separate domains of DnaD protein. The N-terminal domain has a DNA-independent oligomerization activity whereas a C-terminal domain has DNA binding and DNA-dependent oligomerization activities (Carneiro *et al.,* 2006). However, it was shown that binding of the C-terminal domain of DnaD to plasmid is not enough to convert writhe into negative twist, suggesting that scaffold formation by intact DnaD is required for duplex untwisting (Zhang et al., 2006). These separate activities must be linked to each other on the same polypeptide to convert DNA (Zhang *et al.,* 2006).

A DnaI and DnaC complex can be interrupted by DnaD to facilitate its binding to chromosomal DNA and its loading onto the *oriC* region for replication proficient initiation via forming a scaffold to open up the DNA strand (Marsin et al., 2001).

5.7 The aim of the project.

The two DnaD domains structures (N-terminal and C-terminal domains) were solved individually by using crystallography and NMR techniques. The aim was to solve the structure of a more complete form of the DnaD protein.

6 Chapter 6. Results of DnaD protein

6.1 general introduction

The most essential process in all biology is DNA replication. It is built on three main stages, which are initiation, elongation, and termination. The remodeling of a replication origin (oriC) is the initiation stage of the DNA replication that occurs in bacteria through the action of the main initiator protein DnaA and primosomal multi-protein cascades that load two replicative ring helicases on each strand of the DNA duplex. The DNA primase, which is a type of DNA polymerase and catalyses the synthesis of a primer (a short DNA fragment of RNA or DNA) complementary to a ssDNA template, is recruited by associated DNA helicases and signals the switch from initiation to elongation when the two replication forks migrate in opposite directions, one on each strand (Marszalek and Kaguni, 1994). Homologues of DnaA are found in all bacteria but a form of the primosomal cascade involving the DnaD, DnaB, and DnaI proteins is found only in some low G-C content Gram-positive bacteria, including *Bacillus subtilis*. Although DnaI is the Gram-positive functional homologue of the *Escherichia coli* helicase-loader DnaC, both DnaD and DnaB of *Bacillus subtilis* have no homologues in Gram-negative bacteria. They are essential for viability and required for both DnaA and PriA-mediated initiation of DNA replication (Bruand et al., 2001); Postow et al. 2004). DnaD and DnaB have global DNA-remodeling activities. DnaD forms scaffolds and converts supercoiled DNA to an open circular form (Figure 5.2), whereas DnaB compacts laterally supercoiled and linear DNA (Turner et al. 2004; (Zhang et al., 2005).

DnaD is a 232 amino acid primosomal protein that binds to supercoiled forms of DNA and converts them to open forms without nicking. During this remodeling process, all the writhe is converted to twist and the plasmids are held around the periphery of large scaffolds made up of DnaD molecules. This DNA-remodeling function is the consequence of a scaffold-forming activity on the N-terminal domain of DnaD and a DNA-dependent oligomerization activity on the C-terminal domain. (Ishigo-oka et al., 2001); (Marsin et al., 2001).

6.2 Background of DnaD structures

In *Bacillus subtilis,* the full-length version of DnaD has 232 residues that consists of two distinct domains. The N-terminal domain (residues 1-115) was solved by crystallography (pdb: 2V79) (Schneider, S. *et al*. 2007) (Figure 6.1), while the C-terminal domain (residues 129-232), was solved by NMR (Marston et al., 2010) (Figure 6.2). No full-length structure has yet been solved for DnaD.

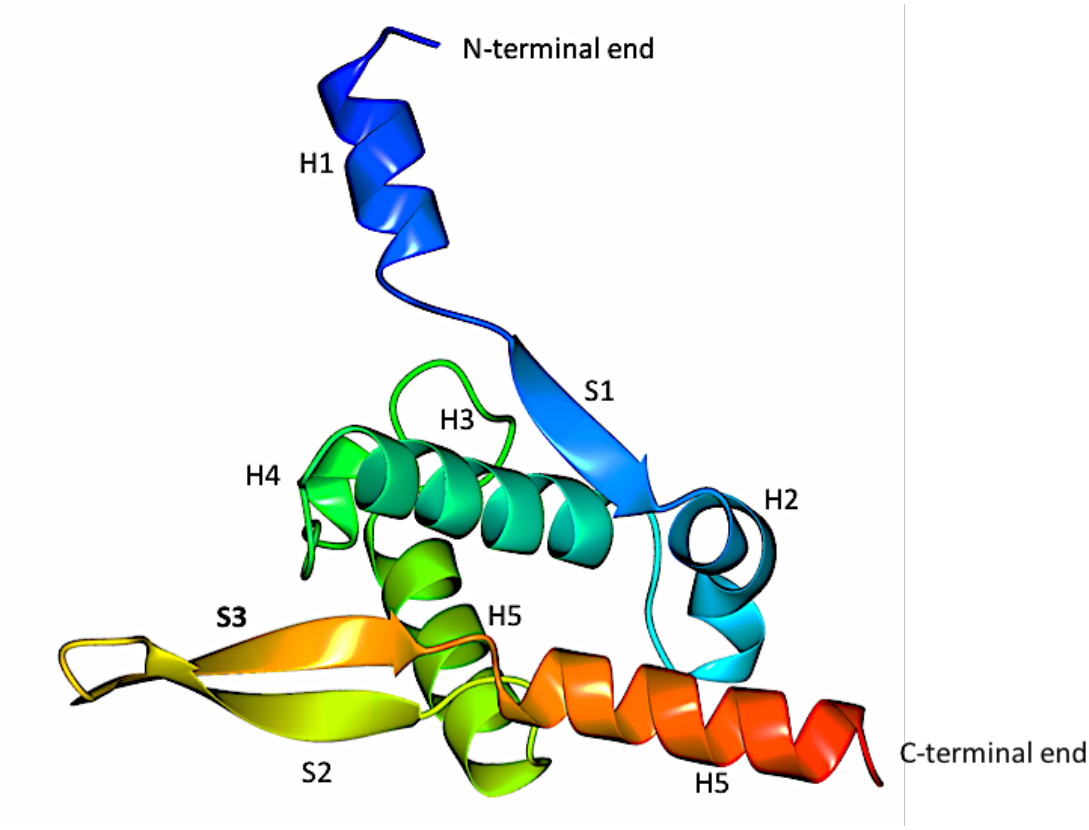


Figure 6.1. N-terminal domain of DnaD structure.

The structure of the N-terminal Domain of DnaD protein was solved in 2008 by Schneider. The N-terminal domains structure (pdb: 2V79) (115 residues) was built as one unit, which showed five distinct helices (H1-5) with two anti-parallel beta strands (S2-3).

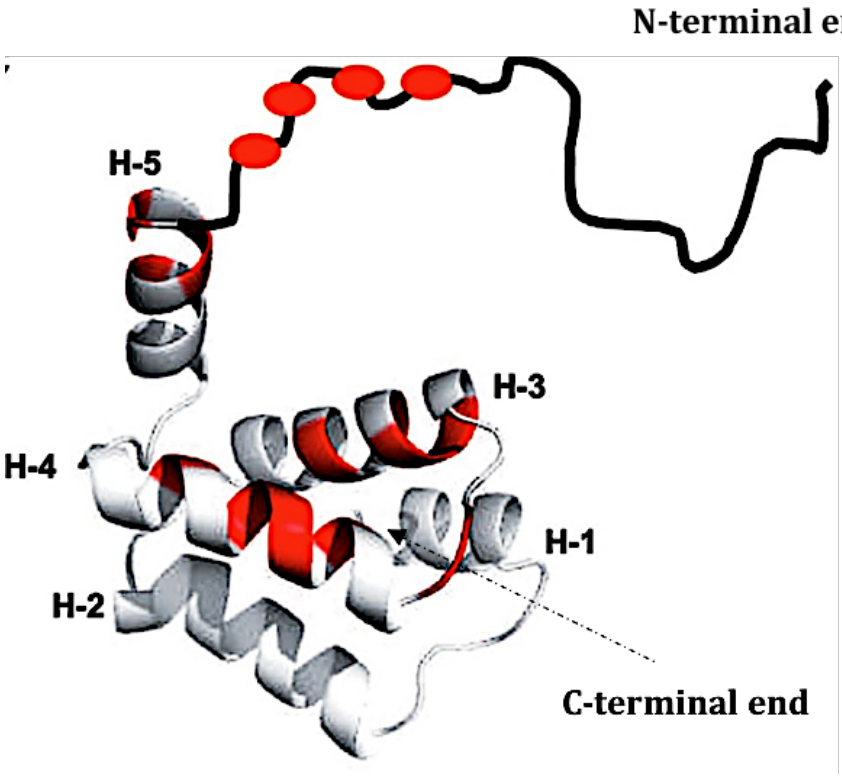


Figure 6.2. C-terminal domain of DnaD structure.

The C-terminal domain of DnaD protein was solved by NMR. The structure (104 residues) was built as one unit, which is made up of five consecutive helices (H1-5) toward the C-terminal end. The mobile part is shown in sketch form at the N-terminal end. The location of DNA binding region is colored in red (Marston, F. Y. *et al.* 2010).

6.3 DnaD cloning

DnaD196 (residue 1-196) is a soluble and stable version of the *Bacillus subtilis* DnaD protein (Prof. P. Soultanas, University of Nottingham, personal communication). The truncated DnaD gene encoding residues 1 to 196 was inserted into the 5493bp pET-22b vector between the NdeI and NotI restriction sites. The transcribed product includes the His-tag residues AAALEHHHHHH immediately after V196 (Figure 6.3).

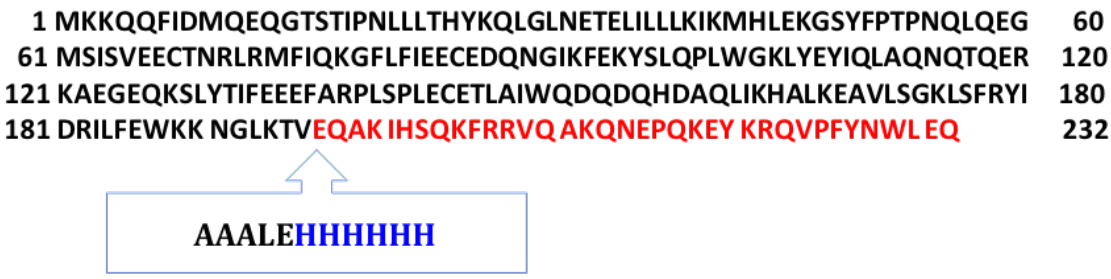


Figure 6.3. Bacillus subtilis DnaD full-length protein sequence.

The sequence of DnaD196 protein from residue Met1 to V196 (black) was used in this study. The remaining sequence of the full-length version is shown (red). The region 1-196 was cloned to include a His-tag fragment (blue).

6.4 Protein expression and solubility

BL21 (DE3) cells carrying the cloned vector were received from the collaborator (Prof. Soultanas) and cultivated in 10 ml LB media 37 ˚ C overnight. A glycerol stock culture of 1ml volume was stored in -70 ˚C until used. A small-scale culture was tested to determine the optimum condition for producing a high level of soluble protein (Chapter 2). A pellet from 1 ml culture of C-terminal His-Tag DnaD196 protein was suspended in 200 µl of 50 mM Tris-HCL buffer pH8 with 0.5 M NaCl and lysed by sonication (3 × 4 sec). The suspension was centrifuged (20 K rpm for 10 min) to obtain the supernatant, which contained all soluble proteins. The results showed that the highest level of expression of soluble protein of interest was under conditions of 1 mM IPTG induction followed by growth at 25 ˚C overnight. The level of the soluble part of the 6xHis-DnaD196 protein was very similar in size to the insoluble level. However, it was in sufficient quantity to be applied in a large-scale production experiment (Figure 6.4).

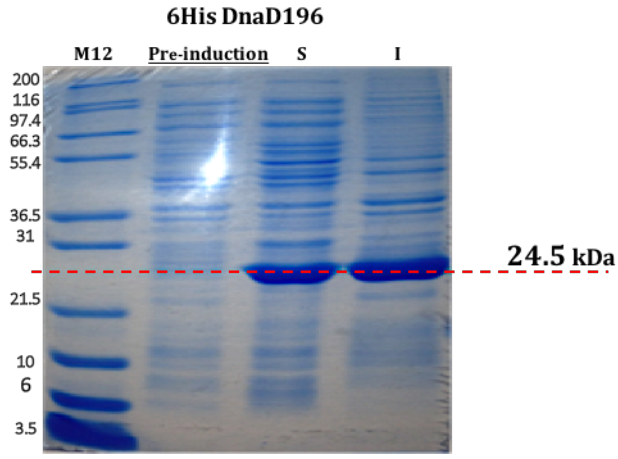


Figure 6.4. Small-scale expression of 6His DnaD196 protein.

SDS gel shows a significant amount of (24.5 kDa) 6His DnaD196 production via overexpression with 1 mM IPTG, 25 ˚C Overnight. However, the solubility ratio of the target protein was approximately 50% in comparison with the insoluble part. S= soluble, I= Insoluble. 10 mg/ml of protein was loaded.

6.5 Protein purification

The large-scale culture and expression was optimized from small scale conditions using 3 litres LB media inoculated with 1% overnight culture, and addition of 1 mM IPTG for induction, 100 mg/ml Ampicillin, and further growth at 25 ˚C overnight (Chapter2).

The harvested pellet of 3 gram cells was suspended in 25 ml of 50 mM Tris-HCl pH 8 with 0.5 M NaCl, and then cell walls were disrupted by sonication (3x20 seconds) and the same protein preparation procedure was followed as described earlier in section 2.12. The collected supernatant had a volume of 22 ml with 440 mg total protein content at a concentration at 20 mg/ml. The cell free extract was loaded onto a His-Trap Ni-HP 5 ml column (GE healthcare^{TM}) at 4 ml/minute flow rate on an AKTA purifier machine (GE-Healthcare), the column was washed with 1 CV of binding buffer, and the protein was eluted by gradient 0-0.35 M imidazole in the same buffer (Section 2.14.4). Analysis of the chromatography showed that the 6HisDnaD196 protein was eluted after 55 ml at 200 mM Imidazole from a 50 mM to 500 mM gradient in fractions 14-16 (Figure 6.5). This result was confirmed by 12% acrylamide SDS-PAGE, which revealed a major band at 24 kDa corresponding to the 6xHis DnaD196 protein after combining the eluted peak fractions. At around 48 kDa a minor band was seen (Figure 6.6). The high purity level of the protein was suitable for crystallization trials. The final volume of purified protein was 200 µl at 16 mg/ml concentration and it was used in successful crystallization trials.

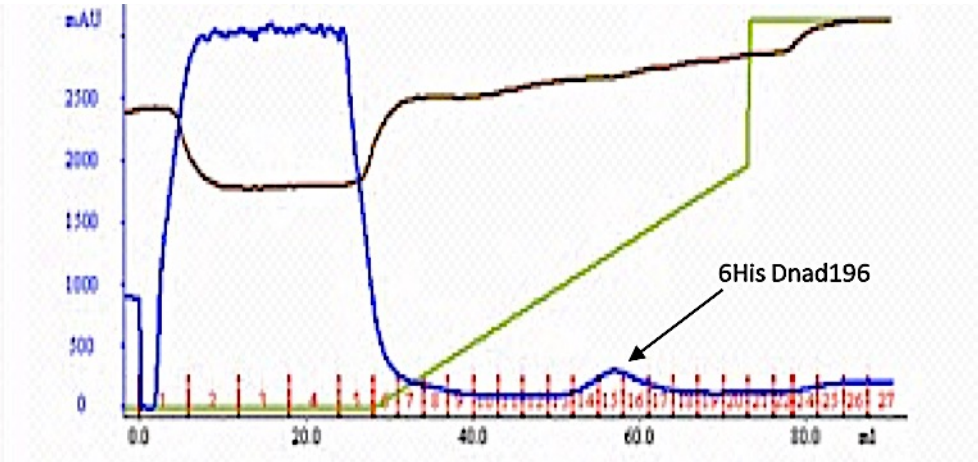


Figure 6.5. His-tag purification assay for 6His DnaD196 protein sample.

Chromatogram shows the 6xHis DnaD196 protein eluted through a His-Trap Ni-HP 5 ml column (GE healthcareTM) purification stage. The blue line represents the UV absorption at 280 nm, grean line refers to imidazole ingredient, and the brown line refers to the conductivity, at a 4 ml/minute flow rate on the AKTA machine. The results showed that the 6His DnaD196 protein eluted after 55 ml at 0.2M imidazole in fractions of 14-16, which were pooled and run on SDS-PAGE and used for crystallization trials.

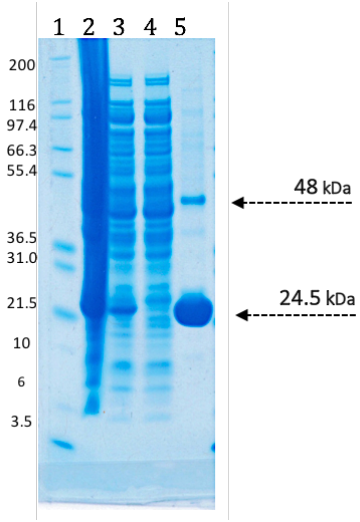


Figure 6.6. 6His DnaD196 SDS-PAGE.

The result on 12% acrylamide SDS-PAGE revealed a large band at 24 kDa, which would correspond to 6xHis DnaD196 protein after combining the eluted fractions of the protein with a minor contaminant at around 48 kDa. 15 µg of the protein was loaded in each lane. 1= Mark12, 2= Debris, 3= CFE, 4= Unbound, 5= Final prep of purified 6His DnaD196 protein

6.6 Crystallization trials

The purified 6xHis DnaD196 protein was concentrated to a volume of 200 µl at 16 mg/mL using a 1 mL 10 kDa Vivaspin column. The buffer was also exchanged into 5 mM Tris-HCl, pH 8 and 50 mM NaCl using a Zepa column. The protein showed signs of precipitation and thus the buffer was changed again to 10mM Tris-HCl pH8 and 0.1 M NaCl.

The protein solution was then used to set up crystal trials containing the various crystallization conditions from different screens (PACT, MPD, JCSG, Ammonium Sulfate, Proplex and Classic). Crystals were observed after 7 days of incubation at 17 ˚C (Figure 6.7).

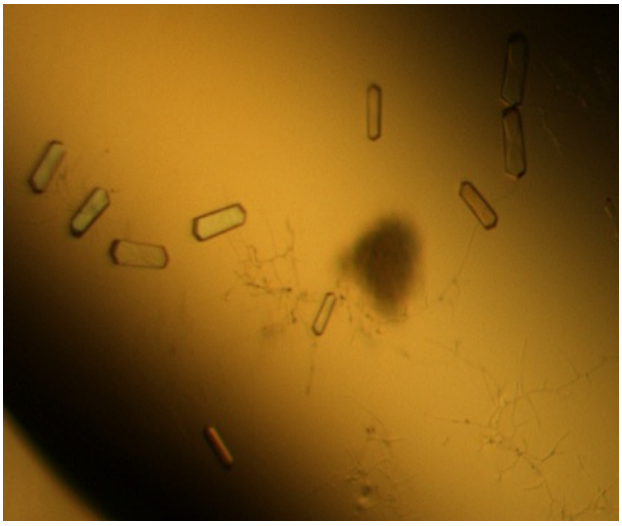


Figure 6.7. 6xHis DnaD196 crystal.

The image shows one of the observed crystals of 6xHis DnaD196 grown from condition H7- PACT (0.2 M Na acetate, 0.1 M Bis-Tris propane pH 8.5, 20% PEG3350). Crystals were observed after 7 days of incubation at 17 ˚C

6.7 Structure determination for 6xHis DnaD195 protein

A wide-ranging variety of 6xHis DnaD196 crystals were cryo-protected using 20% ethylene glycol to test their X-ray diffraction properties at the Diamond Synchrotron, UK. The highest resolution diffraction was collected to 1.7 Å from a crystal grown in 0.2 M Na acetate, 0.1 M Bis-Tris propane pH 8.5, 20% PEG3350 (H7-PACT) (Figure 6.7). A total of 1800 images were collected with an Oscillation of 0.10˚ per image and transmission of 40% on beamline I03 on visit mx12788-51.

6.8 Native data processing

An X-ray diffraction crystal dataset was analyzed by using the statistics of the AIMLESS program via the Xia2 system and using xia2 3d mode, which identified the space group as I222 with unit cell a= 49.9 Å, b= 57.2 Å, c= 77.8 Å, α= 90, β= 90, γ= 90. Strikingly, the Matthews coefficients calculations did not work to give an estimate for the number of molecules in the asymmetric unit (AU) when the molecular weight of the full-length protein (24 kDa) was used. However, it suggested one molecule (1.91 Matthews coeff, 35.72% solvent) in the asymmetric unit based on an 14 KDa estimated molecular weight, which is the size of N-terminal domain of DnaD protein (Figure 6.8). All the data statistics are described in table 6.1.

For estimated molecular weight 14000.

Nmol/asym	Matthews Coeff	%solvent	P(1.51)	P(tot)
1	1.91	35.72	1.00	1.00

Figure 6.8. Matthews's coefficients calculations.

The statistics for the estimate of solvent content based on the number of subunits in the AU of a 1.7 Å crystal dataset suggest one molecule in the AU according to an 14 kDa estimated molecular weight, which is the size of N-terminal domain of DnaD protein.

Table 6.1. Data set collection of 6xHis DnaD196 crystal.

The table shows the processed merging statistics for native data of 6xHis Dnad196 crystal.

SAD	Overall	Low	High
High resolution limit	1.68	4.56	1.68
Low resolution limit	41.97	41.99	1.71
Completeness	96.1	99.7	73.6
Multiplicity	5.4	5.9	2.3
I/sigma	17.2	52.4	1.1
Rmerge (I)	0.065	0.030	0.769
Rmerge (I+/-)	0.058	0.028	0.944
Rmeas (I)	0.072	0.033	0.976
Rmeas (I+/-)	0.07	0.033	1.285
Rpim (I)	0.029	0.014	0.589
Rpim (I+/-)	0.038	0.018	0.867
CC half	0.999	0.999	0.583
Wilson B factor	14.966		
Anomalous completeness	90.3	99.2	2.7
Anomalous multiplicity	3	3.4	1.4
Anomalous correlation	0.17	0.021	-0.070
Anomalous slope	1.032	0.000	0.000
dF/F	0.075		
dI/s (dI)	0.868		
Total observations	67797	4225	1095
Total unique	12541	721	482
Space group	I222		
Unit cell dimensions	a	49.9	
	b	57.2	
	c	77.8	
	α	90	
	β	90	
	γ	90	

$$R_{merge} = \frac{\sum_h \sum_{i=1}^{N} |I_{(h)i} - \bar{I}_{(h)}|}{\sum_h \sum_{i=1}^{N} I_{(h)i}} \qquad R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{j=1}^{n} |I_{hkl,j} - \langle I_{hkl} \rangle|}{\sum_{hkl} \sum_j I_{hkl,j}}$$

6.9 Molecular replacements for DnaD196 structure

The published N-terminal domain structure of DnaD replicase protein (pdb: 2V79) (Shaneider, S. *et al*. 2008) was used for molecular replacement to determine the structure from the new dataset. The published N-terminal domain structure of DnaD had a space group of $P3_121$ with cell dimensions of a= 78.67, b= 78.67, c= 124.56, α= 90, β= 90, and γ= 120, which was totally different in comparison with the new data set presented here. Additionally, it had two molecules in the symmetric unit of the lattice (Schneider et al., 2008).

The initial electron density map from the native diffraction data was poduced by using PHASER-MR via the CCP4 suite to search for one subunit of the N-terminal domain in the asymmetric unit based on Matthews coefficients results, which provided the initial phases to use with the new collected data set. The resulting electron density map revealed noticeable differences in the density in comparison with the phasing model. The structure was of higher resolution to that seen in the published structure and was used to build the first 114 residue of DnaD196 (N-terminal domain). The PHASER MR output file was refined initially for ten cycles using the REFMAC5 program within CCP4 suite and the pdb coordinates were analyzed in COOT (Emsley et al., 2010) (Figure 6.9 and 6.10).

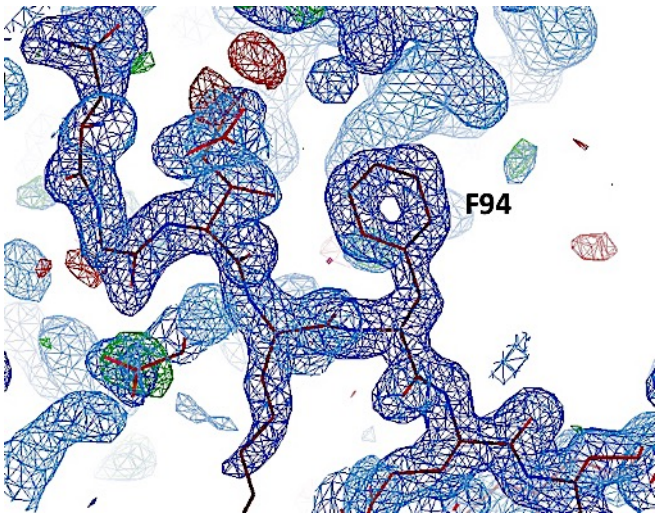


Figure 6.9. Molecular replacement and refinement.

Electron density map via COOT for the DnaD196 structure (114 residues N-terminal domain) after the final refinement (R factor= 15.34 and R free= 20.21). The high-resolution dataset made the electron density very clear to follow and observe residues such as shown phenylalanine 94 in an anti-parallel beta strand.

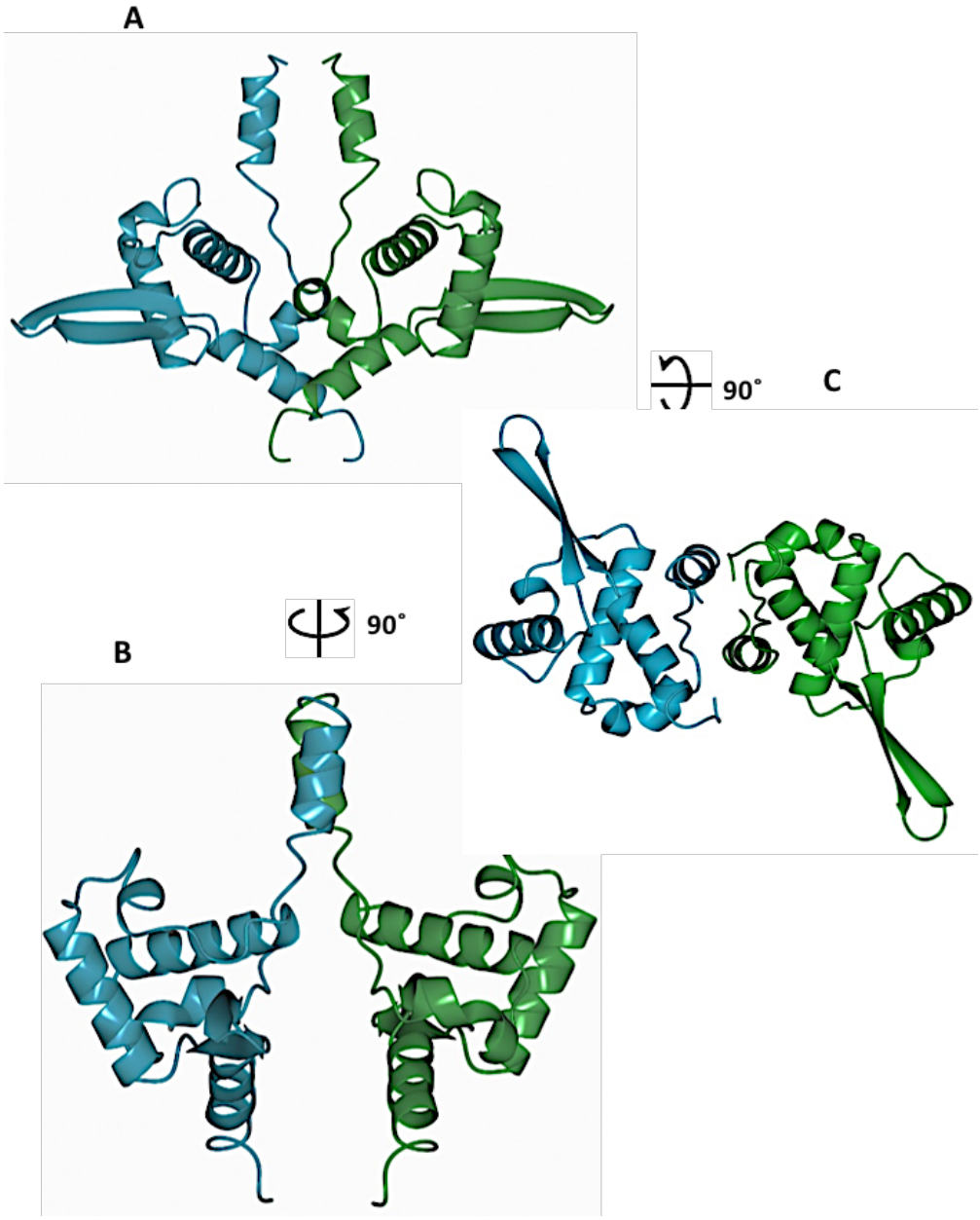


Figure 6.10. N-terminal DnaD196 structure.

The N-terminal domain of DnaD196 structure is shown for two subunits to see the interface in between. They are shown in different colors as a cartoon form via CCP4mg program and in different orientations (front view A, side view B, top down view C). Each subunit contains 6 α-helices and two anti-parallel β-strands.

6.10 Structure validation for DnaD196 protein

A validation analysis was carried out using the MolProbity server for the final refined model structure of DnaD196 N-terminal domain as shown in tables 6.2 and 6.3. A Ramachandran plot analysis revealed all the residues within the favored regions (Figure 6.11).

Table 6.2. Validation statistics of DnaD196 structure.

MolProbity analysis output (Chen *et al.,* 2010) shows the statistics of the validation scores of DnaD196 structure after the final refinement.

All-Atom Contacts	Clashscore, all atoms:	6.2		90th percentile* (N=792, 1.68Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	1	0.93%	Goal: <0.3%
	Favored rotamers	104	96.30%	Goal: >98%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	114	100.00%	Goal: >98%
	MolProbity score^	1.34		96th percentile* (N=8853, 1.68Å ± 0.25Å)
	Cβ deviations >0.25Å	1	0.92%	Goal: 0
	Bad bonds:	0 / 979	0.00%	Goal: 0%
	Bad angles:	1 / 1314	0.08%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 4	0.00%	Expected: ≤1 per chain, or ≤5%

In the two column results, the left column gives the raw count, right column gives the percentage.

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst. For clashscore the comparative set of structures was selected in 2004, for MolProbity score in 2006.

^ MolProbity score combines the clashscore, rotamer, and Ramachandran evaluations into a single score, normalized to be on the same scale as X-ray resolution.

Table 6.3. The statistics of the final refinement cycle.

Value	Final
R factor	0.1534
R free	0.1840
Rms Bond length	0.0180
Rms Bond Angle	1.8775
Rms Chir Volume	0.1237

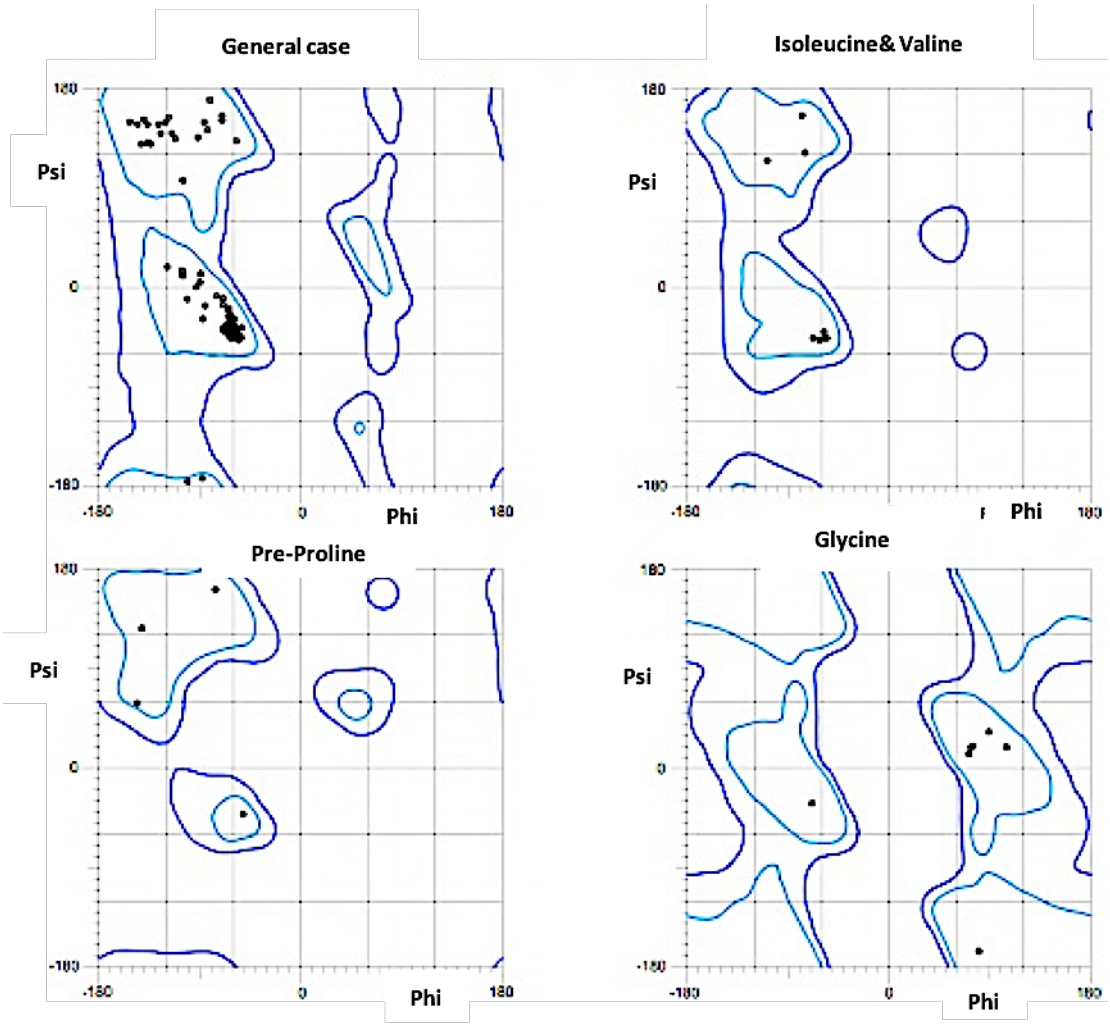


Figure 6.11. Ramachandran analysis of the final structure of Ea10.

The Ramachandran plot is shown for the final refined model of the DnaD196 structure generated using the MolProbity server. http://molprobity.biochem.duke.edu/. The general case of each amino acid in the structure is represented as a plot as well as those for isoleucines and valines, pre-proline residues, and glycines. MolProbity Ramachandran analysis shows the islands of favored Ramachandran angles in light blue and the acceptable Ramachandran angles in dark blue. For DnaD196 crystal, all residues (100%) were in preferred regions. MolProbity analysis output (Chen et al., 2010).

6.11 Tetramer form of DnaD196

The structure of the DnaD196 N-terminal domain was built in a tetramer form by an application of the crystal symmetry, as also suggested by its solution state after gel filtration, to understand the interfaces in the structure (Figure 6.12).

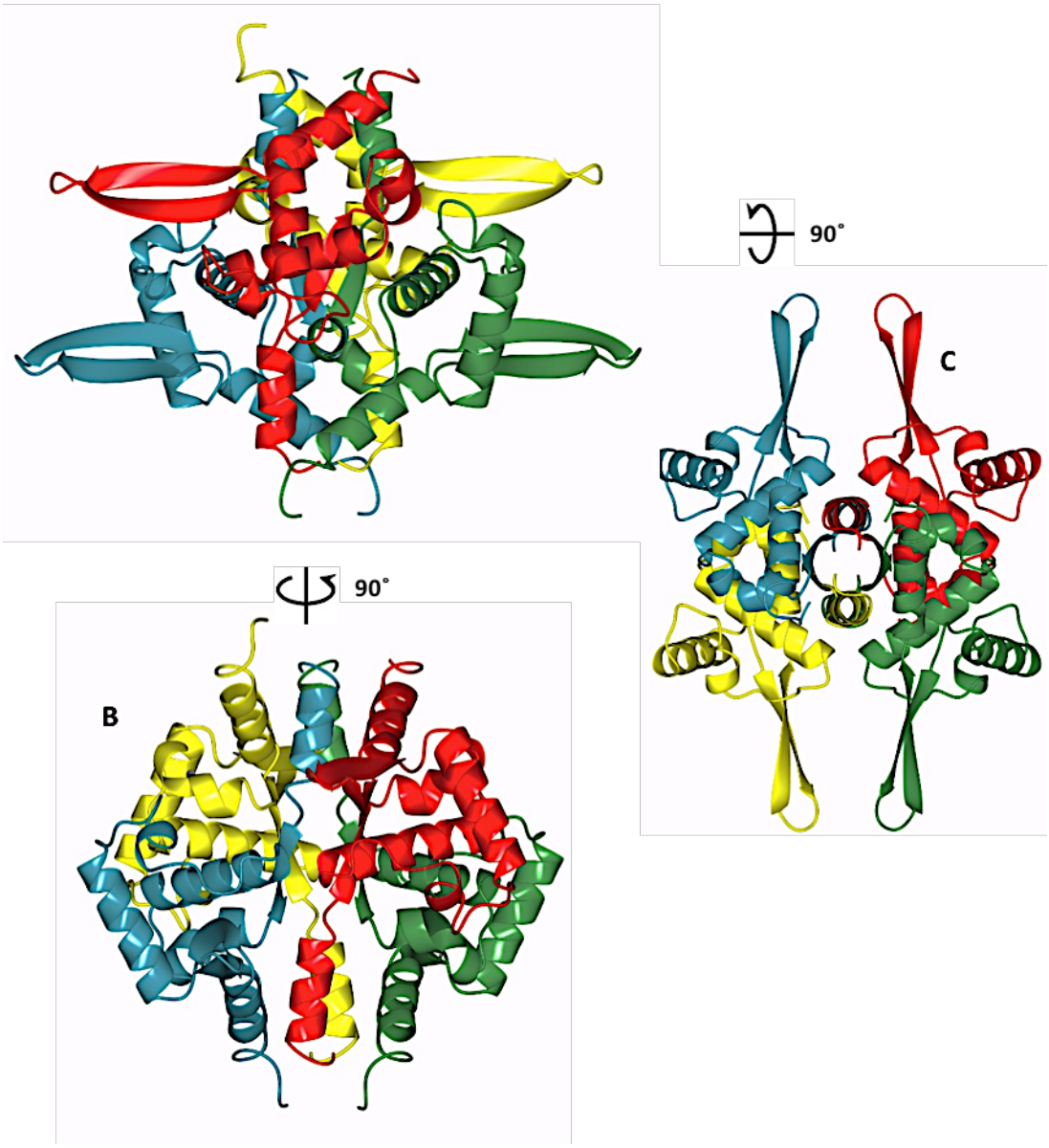


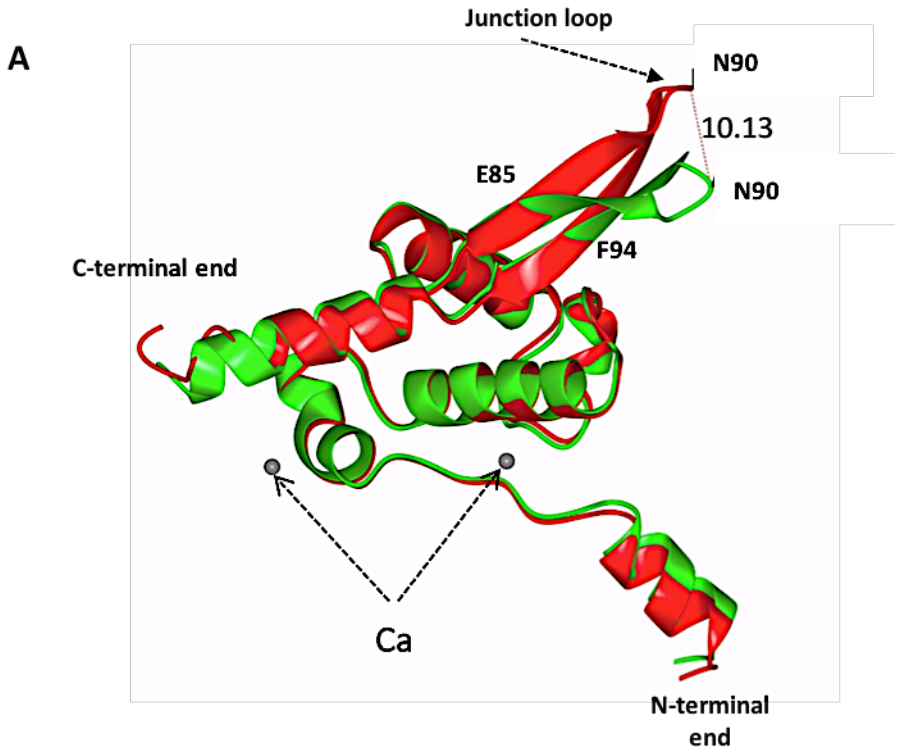
Figure 6.12. Tetramer form of N-terminal DnaD196 structure.

The N-terminal domain of DnaD196 structure is shown for four subunits to see the interfaces in between the subunits, which are shown in different colors as a ribbon form via CCP4mg program and in different orientations, (front on view A, side view B, and top down view C).

6.12 Superposition for N-terminal domain of DnaD196 structure

The monomer and the whole tetramer form of DnaD196 structure were superimposed on the published N-terminal DnaD model (monomer and tetramer structures from pdb) to compare and understand the relationships and differences between the structures (Krissinel and Henrick, 2007). The results of the superposition strategy showed a very good overlay of these two models with high concordance between them that scored RMSD of 1.1 Å for 111 alpha carbons of the monomer and RMSD of 1.17 Å for 429 alpha carbons of the overall tetramer structure. However, they also showed remarkable differences in some regions (Figure 6.13 A, B).

The junction loop, which connects the two anti-parallel beta strands in the DnaD196 chains, was found to be in a different orientation in the two structures. The main clear conformational changes were in the top part of the anti-parallel Beta strands from residue E85 to F94 in each chain, which reported a maximum deviation for the position of the Cα of N90 of range 10.13-12.63 Å. As well as with a slight diversion in the N-terminal helix (Figure 6.13 A, B)



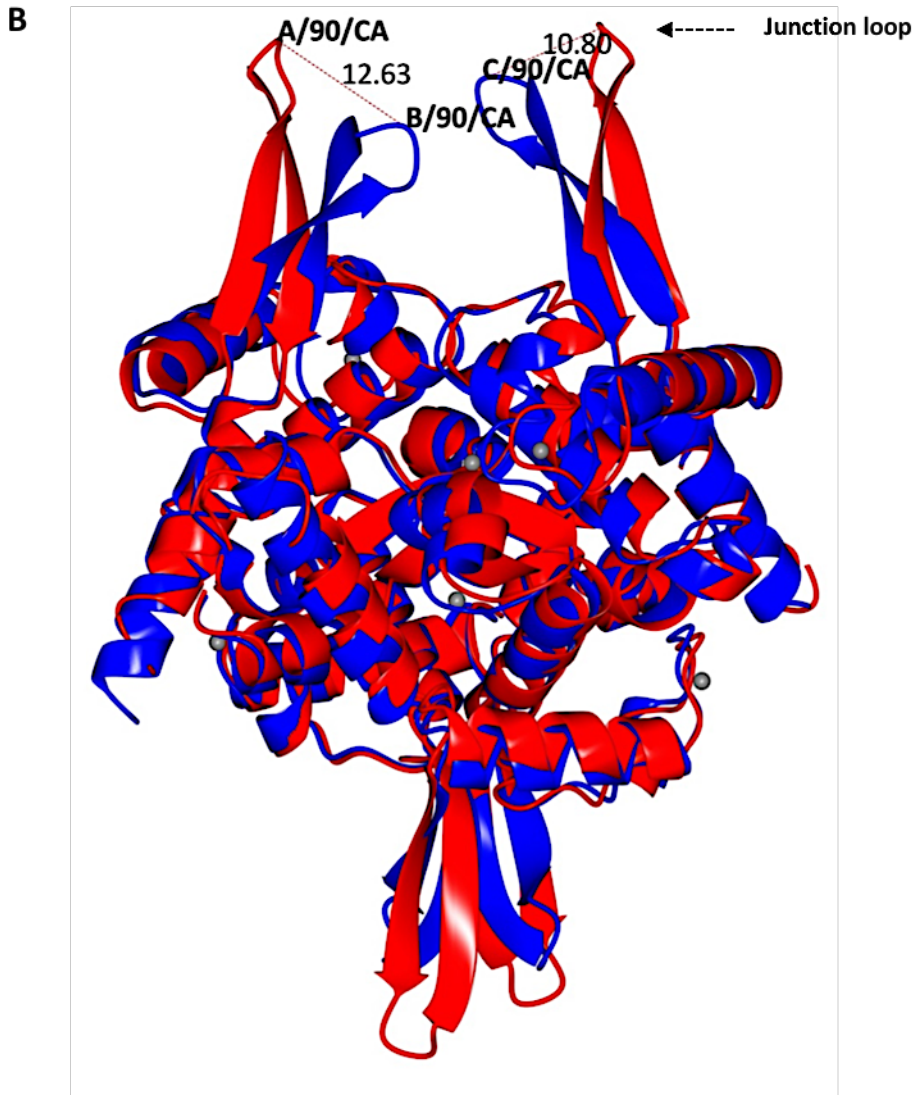


Figure 6.13. superposition for N-terminal domain of DnaD196 structure.

A) The monomer form of the N-terminal DnaD196 structure in red color superimposed with monomeric structure of the published N-terminal DnaD structure model (green) by CCP4mg program. A maximum deviation was observed for the position of the Cα of N90, with a slight diversion in the N-terminal helix. **B**) The tetramer of the N-terminal DnaD196 structure in red showing the orientation and interfaces between its subunits. The N-terminal domain tetramer of DnaD196 structure was superimposed with the tetramer form of the published N-terminal DnaD model (blue), which is built using the suggested biological assembly.

6.13 Mass-spectrometry for DnaD196 crystals

The sample of DnaD196 used in the crystal trials was sent for mass spectrometry (MS) analysis (Chemistry department, University of Sheffield). The result revealed a short version of the DnaD196 protein, which is identical with the N-terminal domain of DnaD196 model as built into the map. Many samples of DnaD196 protein, including the remaining protein solution after crystallization, crystal samples and fresh purified protein, were sent to mass-spectrometry to check their sizes. The investigation revealed that the full length DnaD196 protein was produced in the purification but cleaved during the crystallization trials, possibly because of a contaminating protease present within the conditions (Figure 6.14).

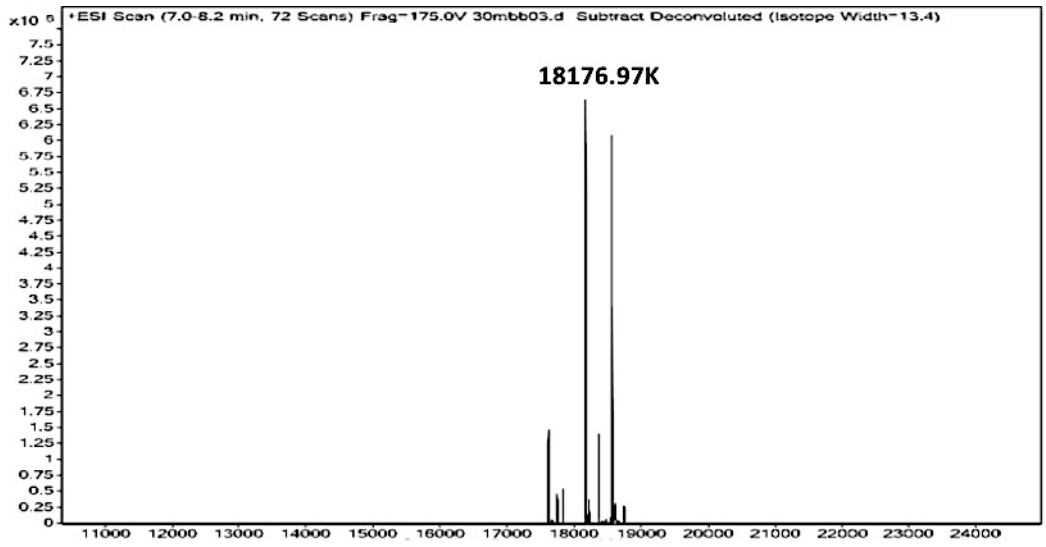


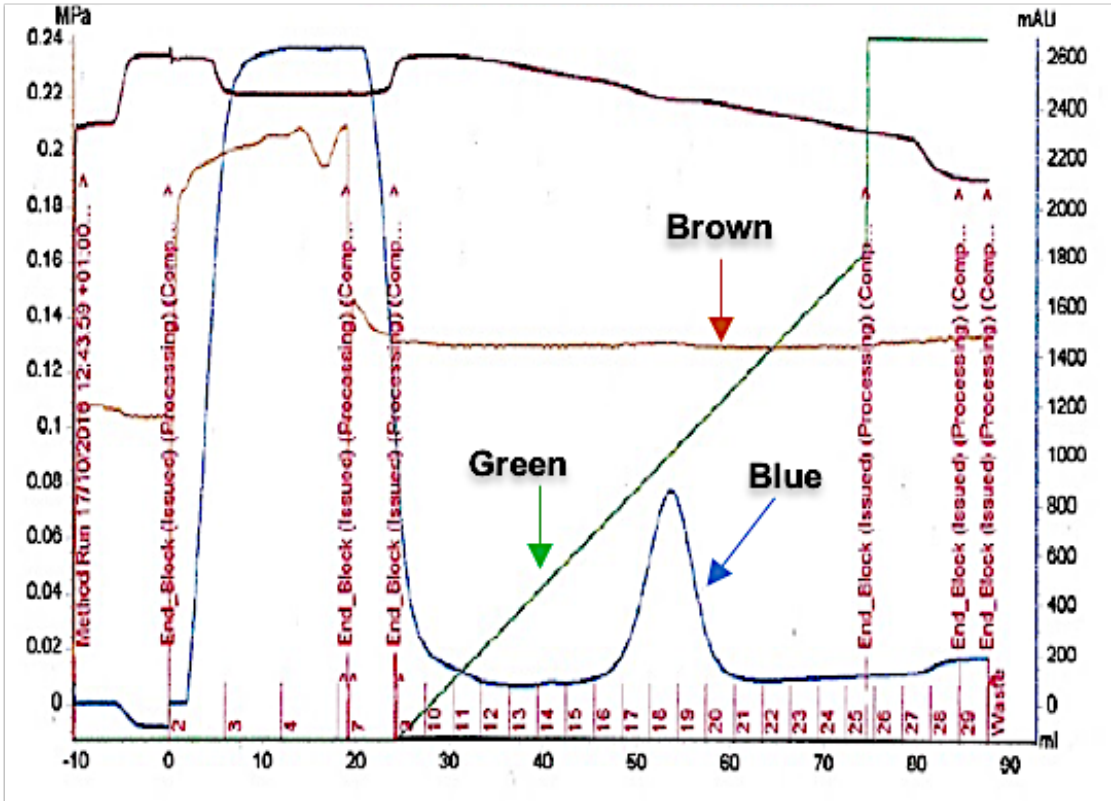
Figure 6.14. The mass-spectrum for DnaD196 crystals.

The spectrum shows the DnaD196 crystals of the same condition used for structure determination. The result shows that the crystals were exactly the short version of DnaD (114 amino acids) observed in the structure.

6.14 Purifying 6xHis DnaD protein with protease inhibitor

A new purification of DnaD196 protein was performed with a protease inhibitor cocktail (complete, EDTA-free, Roche Diagnostic-Germany) added to the cell paste suspension to prevent protein cleavage, and then the protein was purified using the same strategy as described before (see section 6.5). The chromatography result showed that 6His DnaD196 protein was eluted in 55 ml at 0.25 M Imidazole concentration in fractions 16-19 that was approximately similar to the pervious purification. This result was confirmed by 12% acrylamide SDS-PAGE, which also revealed a purity level of the protein, which required an additional run on a gel filtration column (Figure 6.15).

A



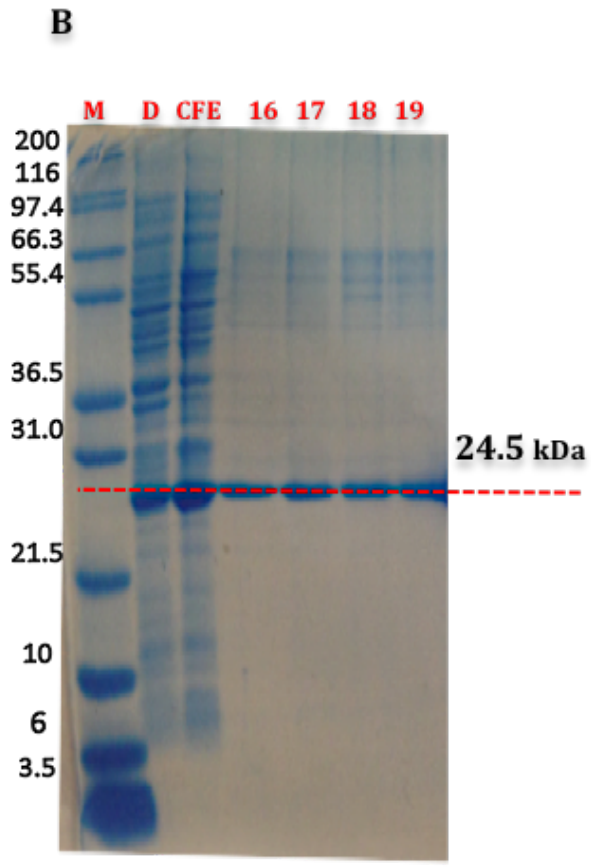


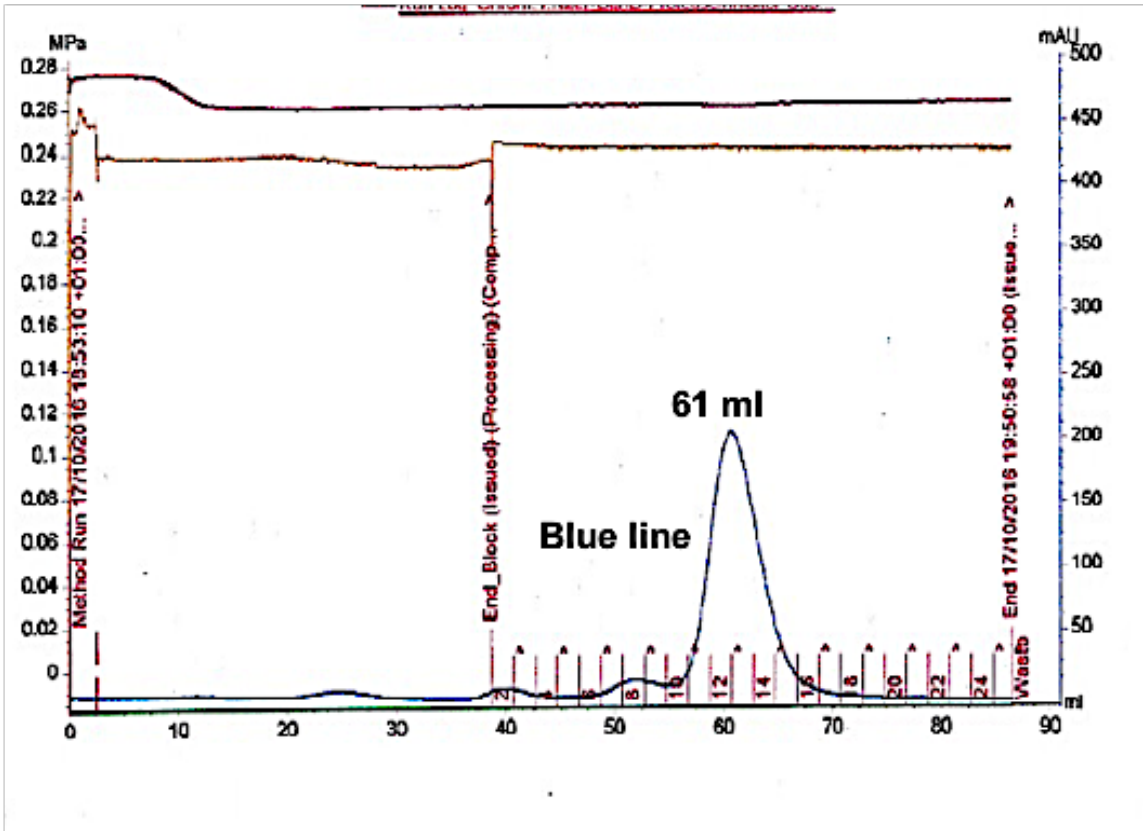
Figure 6.15. Purification strategy for 6xHis-tag DnaD196 protein sample, protease inhibitor added.

The chromatogram shows the 6xHis DnaD196 protein from a His-Trap Ni-HP 5 ml column (GE healthcare™) purification step. The blue line represents the UV absorption at 280 nm, green line represents the Imidazole ingredient**,** and the brown line refers to the conductivity**. A**) The results showed that the DnaD196 protein eluted in fractions 16-19 (7.5 ml volume 2.6 mg/ml concentration, 19.8 mg total). **B)** The SDS PAGE gel shows a band (24 kDa) consistent with the full-length size of DnaD196 protein in lanes 16-19. 15 µg of protein was loaded in each lane.

The fractions were merged and concentrated down to 1 ml volume at 25 mg/ml concentration, and then applied onto a 1.6 x 60 cm Superdex™ 200 gel filtration column (GE healthcare) equilibrated with 50 mM Tris-HCl pH 8.0, 0.5 M NaCl at a flow rate of 1 ml/min. 2 ml fractions were collected and the protein concentration for each determined using the BioRad protein assay. The purification was confirmed using SDS-PAGE with fractions across the peak from the column (Figure 6.16). Based on elution volume, calculations for the highest peak suggest that the DnaD196 protein may be

forming a hexameric assembly, using the following equation: Kav = (Ve-41)/75, where Ve = the eluted volume, 41 = Vo = void volume, 75 = V_t = total column volume, and Kav represents the elution volume function of a molecule. The Ve in this case for the highest peak on the chromatogram is 61.3, giving a Kav of around 0.27. From the Kav, the molecular weight of the protein was then estimated by comparison to protein standards plotted on a logarithmic scale. A molecular weight of around 150 kDa was calculated for the peak, close to the expected 148 kDa size of a hexamer. The eluted fractions with more than 20% of the maximum protein concentration were pooled and a sample was analysed for molecular size and folded conformations via Mass-spectrometry and NMR, respectively. The protein was used for setting up more crystallization trials. The remaining sample was stored at 4 ˚C (up to 1 week) or -70 ˚C (for long-term storage) until needed.

A



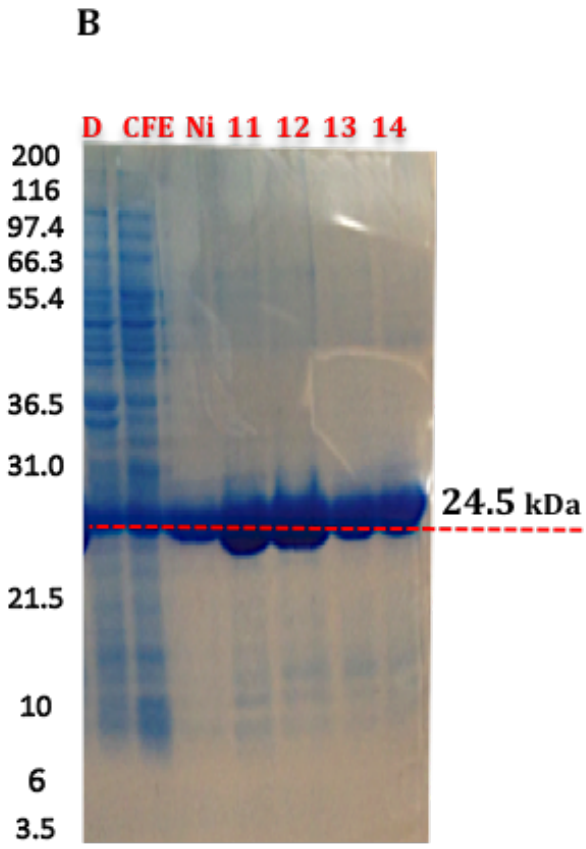


Figure 6.16. Gel filtration purification for 6xHis DnaD196 protein.

A) Chromatogram shows 6xHis DnaD196 protein eluted through the final gel filtration Superdex^{TM} 200 column where the blue line represents the UV absorption at 280 nm. Peak fractions 11-15 have an elution volume of 61.3 ml, which suggested formation of a hexamer assembly based on the calibration plot for the column. **B)** SDS PAGE gel shows the purified 6xHis protein after elution from the Superdex^{TM} 200 column. The eluted fractions were of the expected molecular weight for the intact protein and pure enough for setting crystallization trials. Therefore, they were combined and pooled for crystallization, and also for NMR and mass-spectrum investigations. 15 µg of protein was loaded in each lane.

6.15 Mass spectrometry assay for the full length DnaD196 protein

The purified DnaD protein samples were analyzed by mass spectrometry (Simon Thorpe, University of Sheffield) showing that the protein sample after His-Trap Ni-HP column gave two major products representing the expected full-length size of DnaD196 protein and a double size molecular weight protein product, which could be a dimeric form of the protein. The result of the sample following the gel filtration column revealed a dominant band of a large molecular weight species belonging to the full-length molecular weight of the target protein, without any sign of degradation to the N-terminal domain size itself (Figure 6.17 and 6.18). The mass spectrometry results confirmed that the purification strategy did not contribute to the clipping process for the full-length target protein.

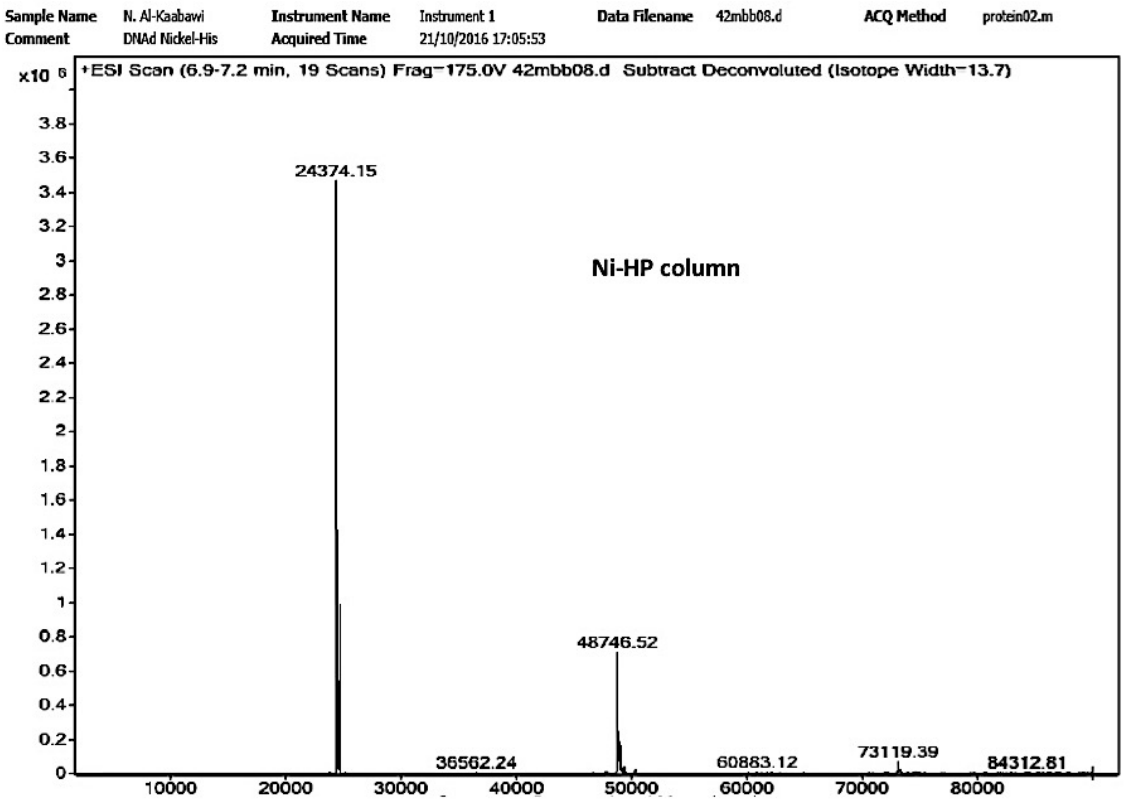


Figure 6.17. The mass-spectrum result of the DnaD196 after Ni column.

The spectrum shows that the DnaD protein sample after Hi-Trap Ni-HP column was exactly in the right molecular weight of the full-length (24 kDa) product without proteolysis, as well as a small peak from a 48 kDa species, which might refer to a dimeric form of the protein.

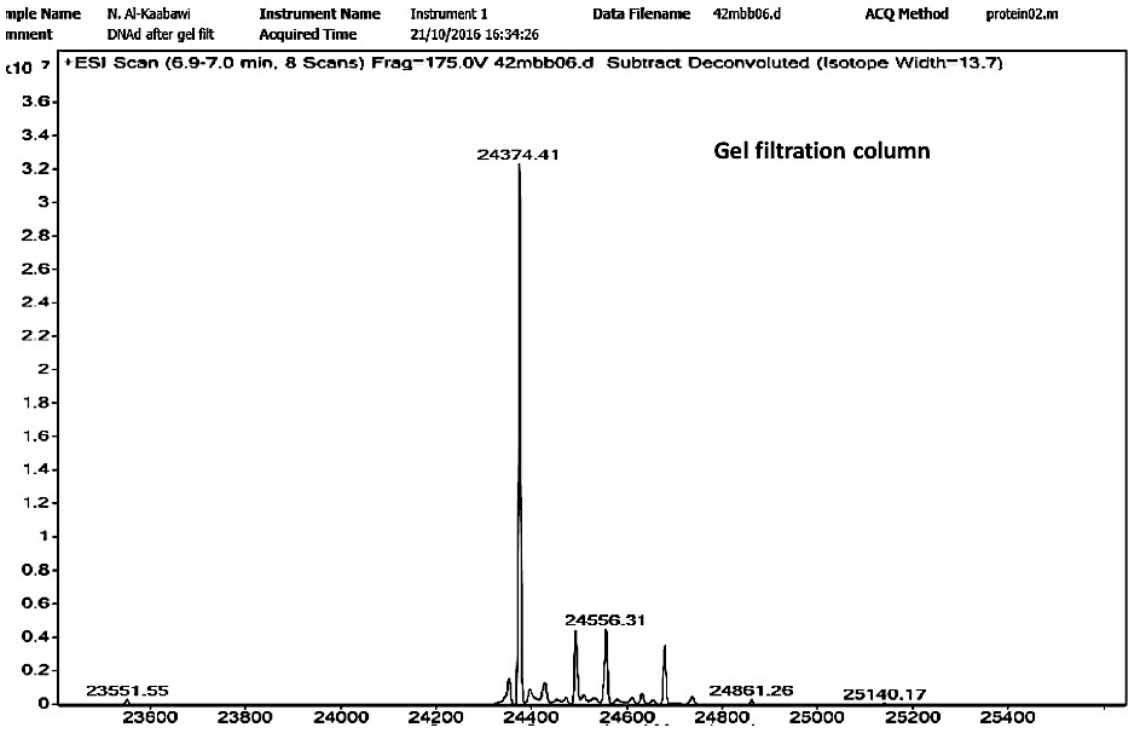


Figure 6.18. The mass-spectrum result of the DnaD196 after Gel filtration.

The chromatogram confirmed the presence of the full-length DnaD196 protein (24 KDa after Gel filtration column without signs of a small proteolysed version of it (18 KDa N-terminal domain).

6.16 NMR (nuclear magnetic resonance)

The purified DnaD196 protein sample after gel filtration procedure was diluted to contain lower salt in a buffer of 10 mM Tris-HCl pH 8.0, 50 mM NaCl, and 10% D_2O (0.5 ml of 0.3 mM concentration of protein), and then sent for analysis via NMR (NMR group, MBB department, University of Sheffield) to analyze the nature of the full length DnaD196 protein (Figure 6.19). The target protein was folded based on the NMR results interpretation (Page et al., 2005) (Section 2.15).

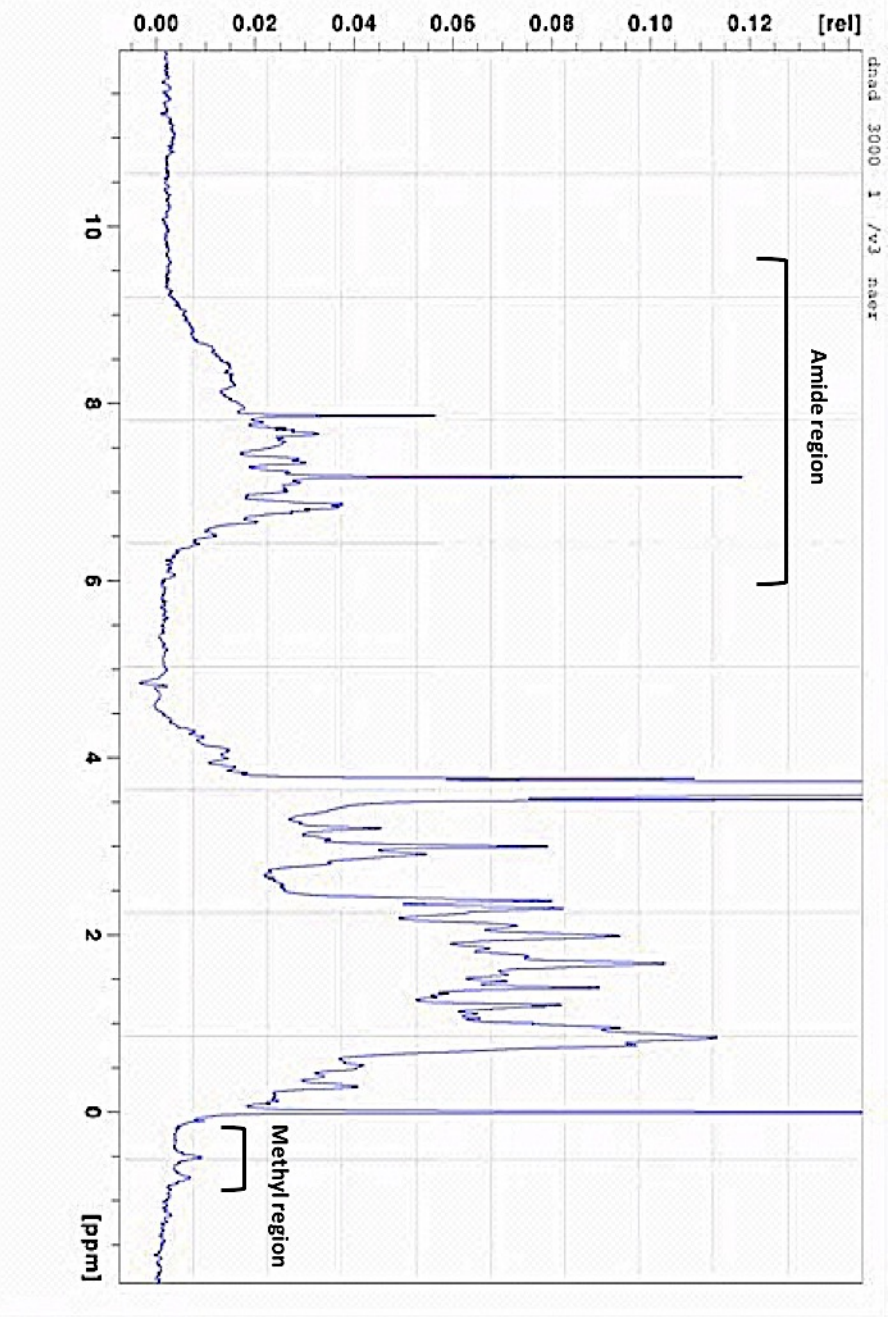


Figure 6.19. The NMR for DnaD196 protein.

The NMR spectrum of DnaD196 protein confirms that the full length DnaD protein was well folded, due to the presence of single sharp peaks in the methyl region (before 0 ppm), and also the dispersed and sharp peaks in the amide region (from 6 to 10 ppm), which refer to folded protein features (Page R *et al*,. 2005).

6.17 Crystallization trials

The purified 6xHis DnaD196 protein was concentrated down to a volume of 500 µl at 15 mg/mL using a 1 mL 10 kDa vivaspin column (same strategy was followed as described in section 6.6. The buffer was also exchanged into 5 mM Tris-HCl, pH8 and 50 mM NaCl using a Zeba column. The stability of purified protein in low salt buffer was better than the first purification as judged by less precipitation.

The DnaD196 protein solution was then used to set up crystal trials containing the various crystallization conditions (PACT, MPD, JCSG, Ammonium Sulfate, Proplex and Classic). The crystals were observed after a few days of incubation at 7 ˚C. Many crystals were observed of different sizes and shapes (Figure 6.20). The crystals were cryo-protected and sent to the Diamond synchrotron (near Oxford) for X-ray diffraction tests.

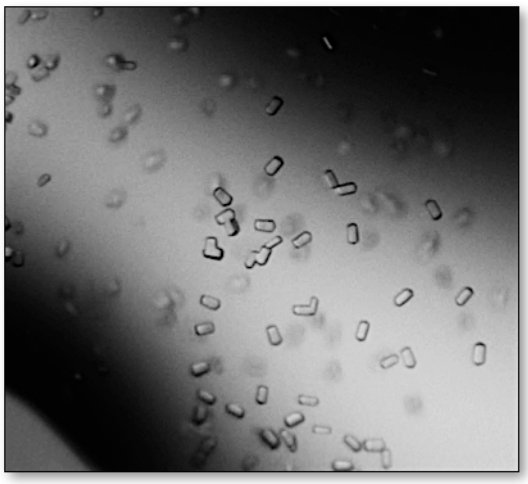


Figure 6.20. 6xHis DnaD196 crystal.

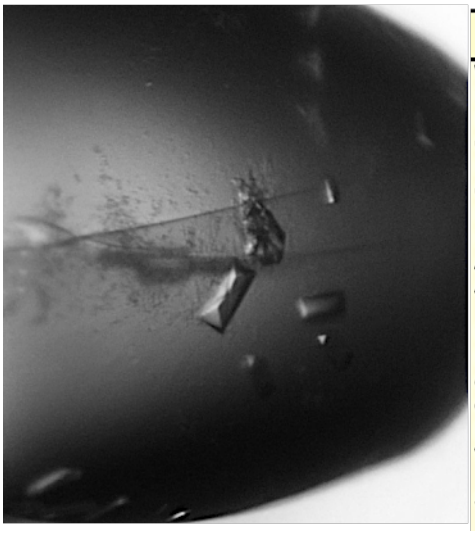
Crystals of 6xHis DnaD196 (prepared with protease inhibitor) after a few days of incubation in 7 ˚C grown in 0.45 M Na/K tartrate tetrahydrate, 0.1 M MES pH 6.3.

6.18 Structure determinations for 6xHis DnaD196 protein

The X-ray diffraction data results revealed that there were protein crystals and the data were collected and processed by XDS and scaled and merged by Aimless. The best resolution data (1.5 Å) were from a crystal grown in condition A1-PACT (0.1 M SPG pH 4, 25% PEG1500) that belonged to space group of I222 with cell dimensions of a= 50.4 Å, b= 56.7 Å, c= 75.2 Å. The statistics of the full dataset are shown (Figure 6.21). However, this data analysis revealed the same results to the previously determined short length version of the N-terminal domain DnaD196 protein.

The results suggested that the N- terminal and C-terminal domains are not stable enough to be crystalized together. Therefore, a new experiment was set up to co-crystallize the full length DnaD protein with the N-terminal part of the DnaA protein, which interacts with the C-terminal domain of DnaD196 protein (Prof. Pannos Soultanas,

unpublished data). The aim was to stabilize the two domains of DnaD196 protein and permit the full-length protein to be crystallized.



	SAD
Wavelength (Å)	0.97625
High resolution limit	1.51 (1.53 - 1.51)
Low resolution limit	41.84 (1.51 - 1.53)
Completeness	70.75
Multiplicity	5.52
CC-half	0.9989
I/sigma	26.58
Rmerge(I)	0.0373
Anomalous completeness	65.29
Anomalous multiplicity	3.03
Unit cell dimensions: a (Å)	50.350
b (Å)	56.560
c (Å)	75.210
α (°)	90.000
β (°)	90.000
γ (°)	90.000
Space group	I 2 2 2

Figure 6.21. Crystal diffraction of DnaD196.

The diagram shows the best dataset were collected from a DnaD196 crystal grown in 0.1 M SPG pH4, 25% PEG1500, which diffracted 1.5 Å. It belongs to I222 space group with a= 50.4 Å, b= 56.7 Å, c= 75.2 Å cell dimensions as analyzed by Aimless. It is near isomorphous with the previous crystal form.

6.19 Investigation of stability for DnaD196 domains by using protein–protein interaction with DnaA protein.

The lyophilized 10 kDa N-terminal domain of the *B. subtilis* DnaA protein (provided by Prof. Soultanas, University of Nottingham) was used to set up a new experiment, to increase the stability of the full-length Dna196 due to the interaction of DnaA with the C-terminal domain of DnaD196 protein. The lyophilized DnaA protein was re-suspended in 1 ml at 10 mg/ml concentration in Buffer A (50 mM Tris-HCl pH 8.0, 0.5 M NaCl), and then run on 1.6x60 cm Superdex-200 gel filtration column (GE healthcare) (following the same procedure, as described on section 2.14.6) to measure the elution volume of the protein peak and to check the stability.

The DnaA protein was combined with 6xHis DnaD196 using a 1:1 ratio then incubated for 2 hours in room temperature. It was then run on a 1.6x60 cm Superdex-200 gel

filtration column to see the volume of complex elution and the strength of the interaction. The results showed no interaction between DnaD and DnaA protein based on the individual peaks in the chromatogram (Figure 6.22) or possibly a very weak interaction, which could be separated by Gel filtration. The two eluted peaks were run on a 15% SDS-PAGE gel to confirm their content.

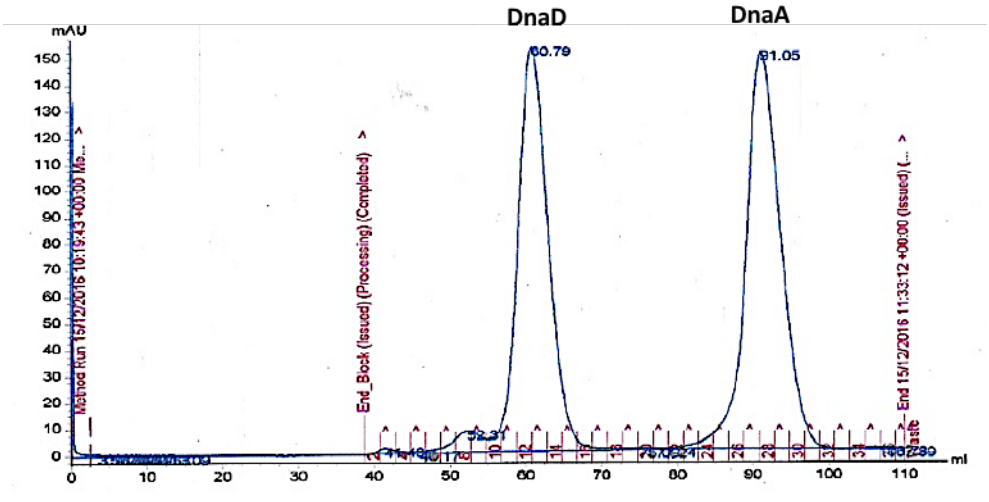


Figure 6.22. Gel filtration of the DnaD and DnaA complex.

Chromatogram shows individual DnaD and DnaA proteins peaks on gel filtration column, which suggested no interaction had happened between these two proteins.

6.20 Protein complex formation on native gel

The DnaD196 and DnaA proteins, and the complex (DnaAD) were run on 12% polyacrylamide native gel (without SDS solution in the buffer and the loading dye) (Section 2.19.2), to see the interaction between the proteins in the natural folded form in a pH8 buffer for both proteins (PI= 6.1 DnaD196, and 5.2 DnaA). In contrast to gel filtration, the result showed a band of protein complex was formed on the native gel that proved the presence of the interaction between DnaD196 and DnaA (Figure 6.23 A). The DnaAD complex solution then was changed to lower salt buffer and crystal trials set up against a range of crystallization screens. Many crystal hits were formed after a few days of incubation in 7 ˚C. The crystals were observed in different sizes and shapes (Figure 6.23 B). The crystals diffractions were then examined via X-ray diffraction at the Diamond synchrotron near Oxford.

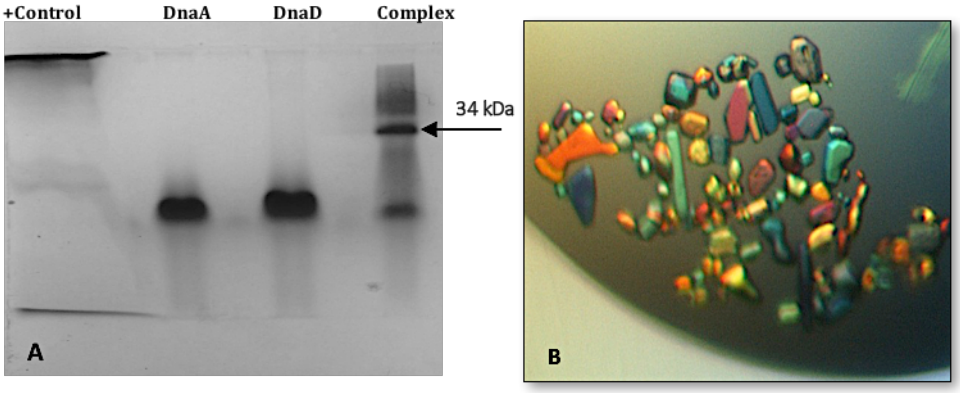


Figure 6.23. 12% Polyacrylamide native gel and Dna complex crystal.

A) The native gel shows the formation of the 34 KDa complex band of DnaD and DnaA proteins (after two hours incubation and Coomassie stain without SDS in the loading dye), which runs heavier than the two proteins individually and suggests the possible interaction between both. 15 µg sample was loaded on gel per lane. **B**) Crystals of a possible DnaAD complex with added protease inhibitor after a few days of incubation at 7 ˚C grown in 0.2 M Na malonate dibasic-hydrate, 0.1 M Bis-Tris propane pH 8.5, 20% PEG3350.

6.21 Structure determinations for DnaAD crystal

Many crystals from a range of different conditions were cryo-protected using 25% ethylene glycol and sent to the Diamond Synchrotron, UK. The highest resolution diffraction data was from a crystal grown from condition C9-PACT (0.2 M Lithium chloride, 0.1 M HEPES pH 7, 20% PEG6000) that gave a 3.8 Å resolution dataset. The X-ray diffraction dataset was analyzed via the Xia2 system and using xia2 3d mode, which identified the space group as I422 with unit cell a= 58.4 Å, b= 58.4 Å, c= 310.8 Å, α= 90, β= 90, γ= 90, which was different from both the earlier study of DnaD196 (see section 6.8), and the published structure (pdb entry 2v79).

One protein molecule in the AU was suggested by Matthews's coefficients calculations for both the full length molecular weight (24.5 KDa), while two protein molecules in the AU was suggested of the short length molecular weight (14 KDa) of DnaD (Figure 6.24). All the data statistics are described in table 6.4. A total of 3600 images were collected with 0.10˚ Oscillation and 0.01 sec exposure time per image.

For estimated molecular weight 14000.

Nmol/asym	Matthews Coeff	%solvent	P(3.83)	P(tot)
1	4.63	73.47	0.02	0.01
2	2.32	46.95	0.97	0.98
3	1.54	20.42	0.00	0.00

For estimated molecular weight 24500.

Nmol/asym	Matthews Coeff	%solvent	P(3.83)	P(tot)
1	2.65	53.58	0.99	1.00
2	1.32	7.16	0.01	0.00

Figure 6.24. Matthews's coefficients calculations for DnaAD complex crystal.

The statistics for the 3.8 Å crystal dataset of a possible DnaAD complex suggest one molecule in the AU for either a 24.5 or 14 kDa estimated molecular weight for the DnaD protein.

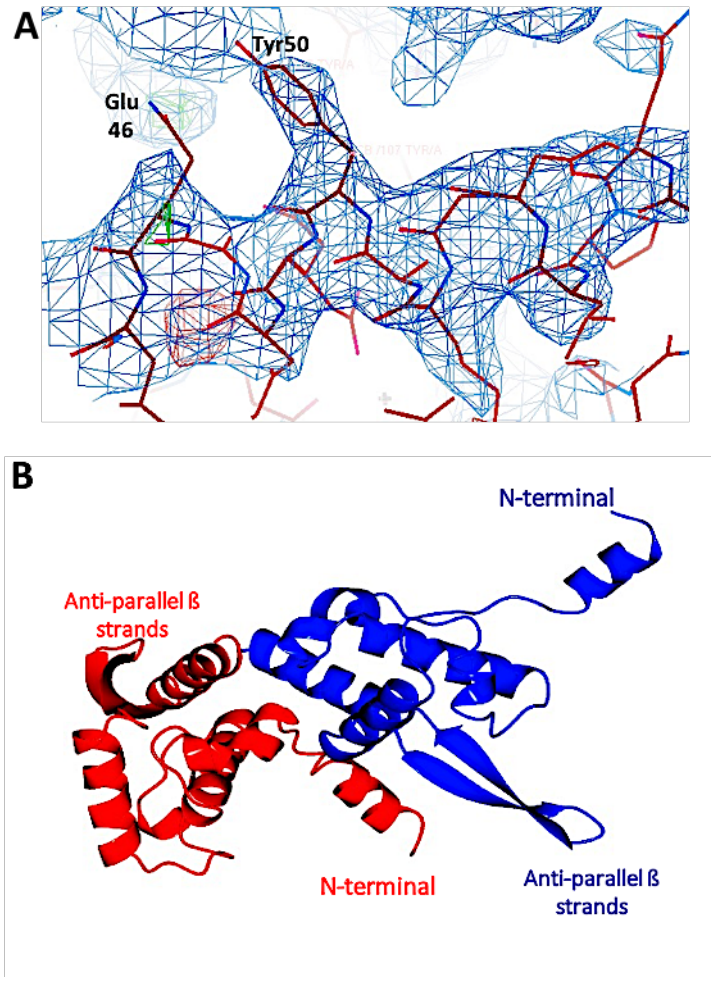
Table 6.4. Data set statistics for a potential DnaAD crystal.

The table shows the processed statistics for the potential DnaAD complex crystal.

SAD	Overall	Low	High
Low resolution	29.54	29.54	3.93
High resolution	3.83	17.12	3.83
Multiplicity	23.3	13.9	19.1
I/sigma	12.50	25.90	1.40
Completeness	99.3	84.2	92.5
Rmerge (I)	0.248	0.097	5.468
Rmerge (I+/-)	0.250	0.098	5.401
Rmeas (I)	0.258	0.102	5.733
Rmeas (I+/-)	0.256	0.102	5.545
Rpim (I)	0.070	0.030	1.690
Rpim (I+/-)	0.053	0.026	1.215
CC half	99.7	99.5	81.6
Wilson B factor	14.966		
Anomalous completeness	99.3	100.0	91.2
Anomalous multiplicity	13.8	12.5	10.8
Anomalous correlation	-5.7	-29.7	8.0
Anomalous slope	0.797		
dF/F	0.075		
dI/s (dI)	0.657		
Total observations	66345	584	3583
Total unique	2847	42	188
Space group	I422		
Unit cell dimensions	a	58.4	
	b	58.4	
	c	310.8	
	α	90	
	β	90	
	γ	90	

6.22 Molecular replacements for DnaAD structure

The N-terminal domain structure of DnaD196 replicase protein was used for molecular replacement with a 3.8 Å resolution DnaAD crystal. The structure was solved by using PHASER-MR from CCP4 suite to search for two subunits in the asymmetric unit based on the solvent estimation by the Matthews coefficients calculations. All choices of related space groups for I422 were tested (McCoy, 2006). The resulting electron density map revealed noticeable differences to the model used. All 115 residues of N-terminal DnaD196 protein were assigned. Subsequent rounds of rebuilding and refinement were performed using COOT and REFMAC5 until no more improvement could be made to the model (Emsley et al., 2010). The results revealed that the resulting structure was just the N-terminal DnaD196 structure and in an orientation with some conformational changes in comparison with the previous structures, based on interaction between the subunits in the dimer and tetramer (R factor= 0.33, R free= 0.378) (Figure 6.25).



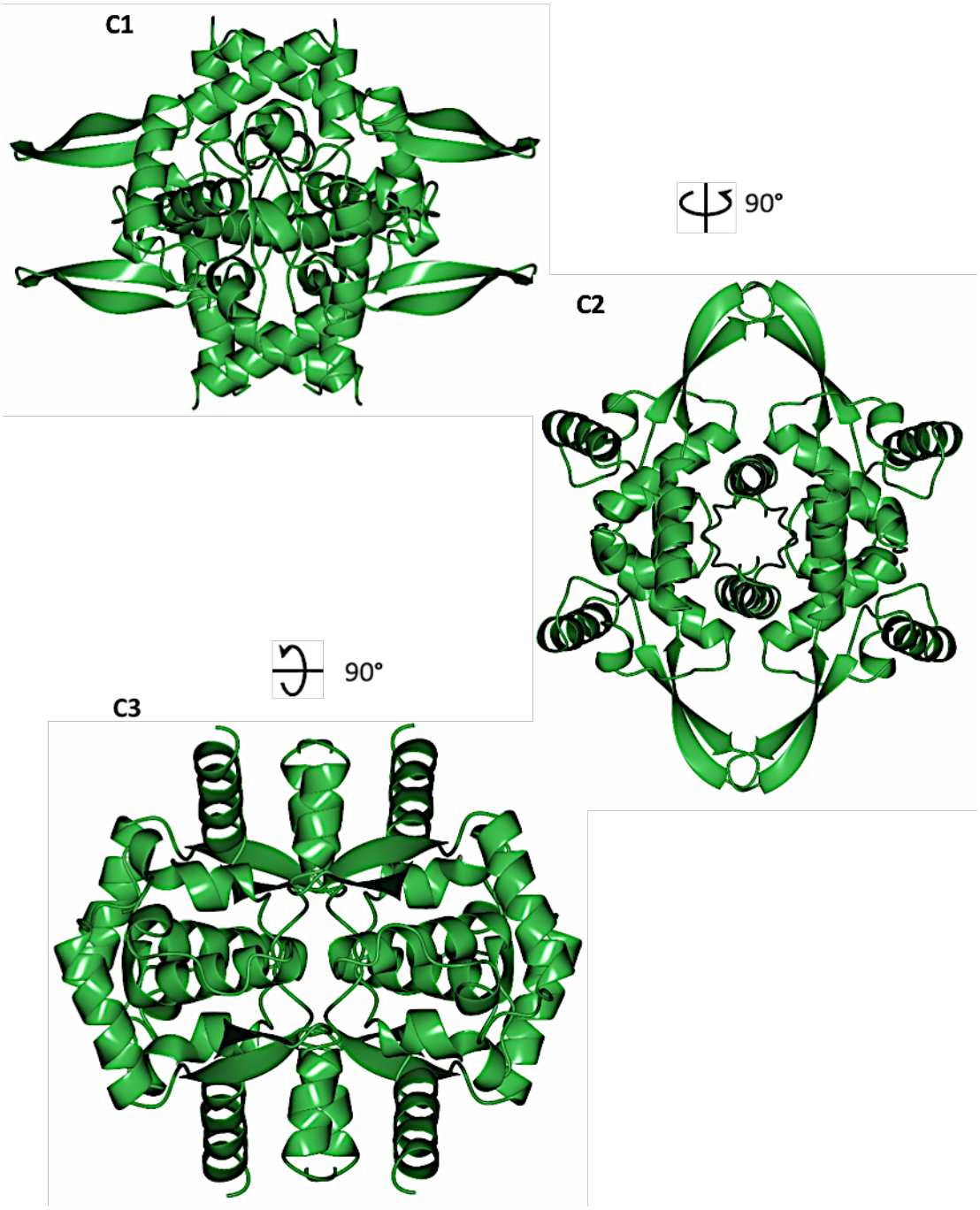


Figure 6.25. Structure of DnaD from the potential DnaAD complex crystal dataset.

A) Part of the electron density map around the 2nd helix of DnaD (residues 45-57) after the final refinement. The electron density was difficult to follow because of the low resolution (3.8 Å) as shown for Glutamic Acid 46 and Tyrosine 50. **B**) A dimeric form (red and blue) of the N-terminal domain of DnaD structure was observed as seen in the previous structures. **C)** The N-terminal domain of DnaD structure is shown as a tetramer (green) in a ribbon form via CCP4mg program and in different orientations (front on view C1, side view C2, top down view C3).

6.23 Superposition for N-Terminal domain of DnaD structure

The monomeric form of DnaD, which belongs to a crystal formed from conditions containing both DnaD and DnaA, were superimposed on the published N-terminal DnaD model (Schneider et al., 2008). The results of the superposition strategy showed a general agreement in the structure with a very good overlay of these two models with high concordance between them that scored an RMSD of 1.1 Å for 111 alpha carbons in the monomer, but some changes were seen. The distinct conformational changes were in the top part of the anti-parallel Beta strands (junction loop). A value of 3.05 Å was reported for a deviation in the position of residue Q89. This suggests that this region is dynamic, which may be related to the function of the protein. A slight shift in the N-terminal helix was seen (2.12 Å at residue K3) in the N-terminal end (Figure 6.26 A). Additionally, the superposition results of DnaD monomer with the previous structure presented in this work and the published structures revealed a range of conformations, which could be relevant to the function of protein, or may have occurred due to crystal packing (Figure 6.26 B)

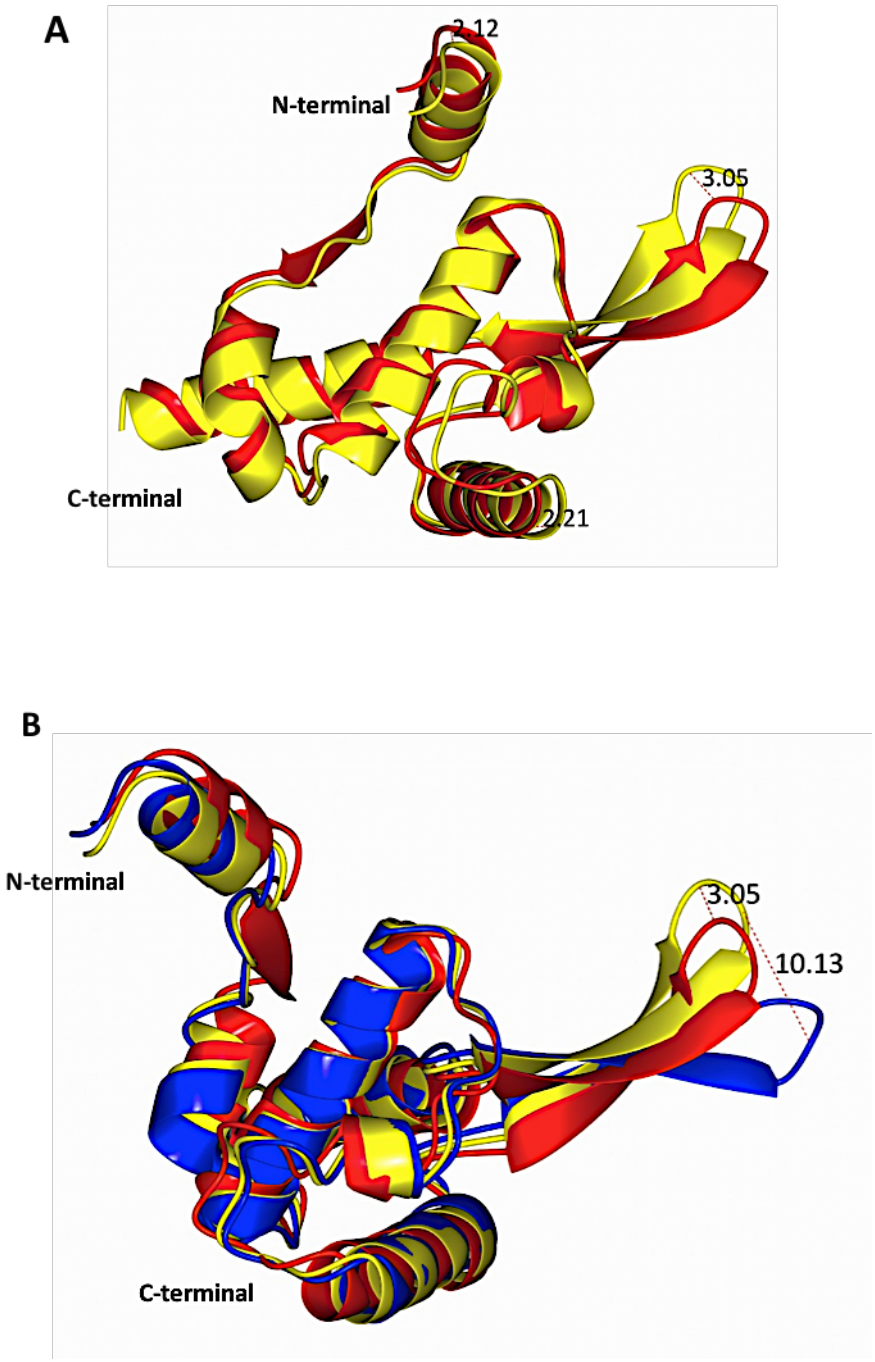


Figure 6.26. Superposition for N-terminal domain of DnaAD structure.

A) The monomer form of the N-terminal DnaD196c structure in red superimposed with monomeric structure of the published N-terminal DnaD structure model (pdb: 2V79) in yellow by CCP4mg program. **B**) The monomer of the N-terminal DnaAD structure in red is superimposed with previously solved N-terminal DnaD196 structure (blue) and the monomer structure of the published N-terminal DnaD model (yellow). The result shows a maximum displacement for the Cα of Q89 against the previously solved N-terminal DnaD196 structure

6.24 Discussion

The structure determination of the full length DnaD protein was attempted, involving the N- and C-terminal domains which have never been solved together.

The full-length product of DnaD protein (232 residues) had a solubility problem. Therefore, a DnaD196 protein was designed to sort out this issue by removing the last 34 residues from the C-terminal end and adding a 6x Histidine-tag for a purification reason. The DnaD196 product was well expressed with a reasonable level of solubility. Crystals of DnaD196 protein were observed after setting up many crystallization trials. Structural determination using a model of the published 18 kDa N-terminal structure (pdb:2v79) (Schneider et al., 2008). in molecular replacement with the DnaD196 crystals revealed that all the crystals contained only the N-terminal domain. However, the result showed that the solved structure was of higher resolution (1.68 A) and in a different space group (I222) and cell dimensions (a= 49.99, b=57.2, c=77.8, α=90, β= 90, and γ= 90) in comparison with the published model as mentioned in section (6.9).

The reason behind a cleavage of the two domains of DnaD196 protein in the crystallization solutions is still unknown. It is more likely to be due to the presence of a protease activity that can cut in a protease sensitive region in the linker loop, which is hidden within the protein sequence. Mobility of DnaD domains would increase the possibility of cleavage by opening an access for the protease enzymes to bind the protease sensitive region of the DnaD linker loop. Mutations could be made in the linker loop sequence to avoiding a protease cleavage. Although, a cocktail protease inhibitor was used in the protein solution to inhibit the protease activity, only N-terminal DnaD196 crystals were formed. The type or the quantity of the protease inhibitor were clearly not suitable to stop cleavage of the DnaD196 protein. Alternative protease inhibitors in different titrations could be used to find a more suitable inhibitor to stop the protease activity. Many proteases need metal ions to accelerate the enzyme activity. Therefore, using EDTA may help in this issue. The used inhibitor cocktail was EDTA –free.

A more extensive analysis could involve a screen by mass spectrometry of all the crystallization conditions for determining anywhere the full-length size of the protein

was found and comparing these conditions with those where the truncated N-terminal domain size is found. This might reveal activation factors for the protease activity.

Crystallization trials were also set down using a protein complex of DnaD196 protein and DnaA protein. It was hoped that using a protein complex of DnaAD in crystallization would give the possibility of a full-length structure of DnaD196 and an analysis of the protein–protein interaction between DnaD196 and DnaA proteins. Gel filtration purification results showed that there was no or only very weak interaction between the proteins (Section 6.19). However, the acrylamide native gel confirmed the presence of this interaction (Section 6.20). The failure to get crystals of the complex and the determination of a structure of DnaD N-terminal domain alone suggests that the complex was indeed not very stable and the protease sensitive site in DnaD was still accessible.

Identifying a small molecule that has the ability to stabilize the two proteins together (*i.e.* increase their affinity for each other) could be helpful in producing co-crystal structures. However, without knowledge of any such molecule, this would require screening of a compound library to identify possible stabilizers and using a binding assay to compare the affinity of DnaD to DnaA for each molecule to the control (without any molecule). In reality, such a strategy would likely be time-consuming as well as extremely costly, especially without recognition of a medical use to justify identifying such a compound for this purpose.

It was found during the purification that the DnaD196 appears in solution to be in a hexameric state with a size of 150 kDa and the full-length form of DnaD196 is not cleaved in the solution. However, this data is contradictory to the published data by Schneider (2008), which mentioned that DnaD assembles into a dimeric state in solution.

Regarding with hexameric state of DnaD, the protein solution can possibly be used to solve the structure of DnaD196 by either using a Small Angle X-ray scattering technique (SAXS) or non-crystallographic techniques such as Cryo-Electron Microscopy strategies, which use the protein solution rather than crystals as well.

Since DnaD is a DNA binding protein, adding appropriate DNA substrates to the protein solution during crystallization may also help to stabilize the full-length form, thereby assisting the two domains to be crystallized together.

Regarding the N-terminal domain solved structure, conformational change was seen, especially at the position of the junction loop of the anti-parallel Beta-strands (residue E85 to F94) in comparison with the published structure (pdb: 2V79) (Section 6.23).
This flexible region might be involved in locating/interacting with the C-terminal domain or involved in forming a scaffold by increasing the stability to bind another tetramer. It possibly also interacts with other proteins such as DnaA or plays a role with DNA binding activity.

7 Chapter 7. Reference

Ash, C., Farrow, J.A.E.A.E., Wallbanks, S., and Collins, M.D.D. (1991). Phylogenetic heterogeneity of the genus Bacillus revealed by comparative analysis of small-subunit-ribosomal RNA sequences. Lett. Appl. Microbiol. *13*, 202–206.

Azevedo, V., Alvarez, E., Zumstein, E., Damiani, G., Sgaramella, V., Ehrlich, S.D., and Serror, P. (1993). An ordered collection of Bacillus subtilis DNA segments cloned in yeast artificial chromosomes. Proc. Natl. Acad. Sci. U. S. A. *90*, 6047–6051.

Benson, F.E., Stasiak, A., and West, S.C. (1994). Purification and characterization of the human Rad51 protein, an analogue of E. coli RecA. EMBO J. *13*, 5764–5771.

Berry, J., Rajaure, M., Pang, T., and Young, R. (2012). The spanin complex is essential for lambda lysis. J. Bacteriol. *194*, 5667–5674.

Blasche, S., Wuchty, S., Rajagopala, S. V., and Uetz, P. (2013). The Protein Interaction Network of Bacteriophage Lambda with Its Host, Escherichia coli. J. Virol. *87*, 12745–12755.

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.

Bruand, C., Ehrlich, S.D., and Jannière, L. (1995). Primosome assembly site in Bacillus subtilis. EMBO J. *14*, 2642–2650.

Bruand, C., Farache, M., McGovern, S., Ehrlich, S.D., and Polard, P. (2001). DnaB, DnaD and DnaI proteins are components of the Bacillus subtilis replication restart primosome. Mol. Microbiol. *42*, 245–255.

Bruand, C., Velten, M., McGovern, S., Marsin, S., Serena, C., Dusko Ehrlich, S., and Polard, P. (2005). Functional interplay between the Bacillus subtilis DnaD and DnaB proteins essential for initiation and re-initiation of DNA replication. Mol. Microbiol. *55*, 1138–1150.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003). Prophage Genomics. Microbiol. Mol. Biol. Rev. *67*, 238–276.

Carneiro, M.J.V.M., Zhang, W., Ioannou, C., Scott, D.J., Allen, S., Roberts, C.J., and Soultanas, P. (2006). The DNA-remodelling activity of DnaD is the sum of oligomerization and DNA-binding activities on separate domains. Mol. Microbiol. *60*, 917–924.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J.,

Murray, L.W., Richardson, J.S., Richardson, D.C., and IUCr (2010). *MolProbity* : all-atom structure validation for macromolecular crystallography. Acta Crystallogr. Sect. D Biol. Crystallogr. *66*, 12–21.

Churion, K.A., and Bondos, S.E. (2012). Identifying solubility-promoting buffers for intrinsically disordered proteins prior to purification. Methods Mol. Biol. *896*, 415–427.

Clark, a. J., and Margulies, a. D. (1965). Isolation and characterisation of recombination-deficient mutant of Escherichia coli K12. Proc. Natl. Acad. Sci. U. S. A. *53*, 451–459.

Costa, S., Almeida, A., Castro, A., and Domingues, L. (2014). Fusion tags for protein solubility, purification, and immunogenicity in Escherichia coli: The novel Fh8 system. Front. Microbiol. *5*, 63.

Court, D., and Oppenheim, A.B. (1983). Phage λ's accessory genes in Lambda II . Lambda II 251–277.

Court, D.L., Sawitzke, J.A., and Thomason, L.C. (2002). Genetic Engineering Using Homologous Recombination. Annu. Rev. Genet. *36*, 361–388.

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr. Sect. D Biol. Crystallogr. *62*, 1002–1011.

Dahms, S.O., Kuester, M., Streb, C., Roth, C., Sträter, N., Than, M.E., and IUCr (2013). Localization and orientation of heavy-atom cluster compounds in protein crystals using molecular replacement. Acta Crystallogr. Sect. D Biol. Crystallogr. *69*, 284–297.

Deng, S., Stein, R.A., and Higgins, N.P. (2004). Transcription-induced barriers to supercoil diffusion in the Salmonella typhimurium chromosome. Proc. Natl. Acad. Sci. U. S. A. *101*, 3398–3403.

Didelot, X., and Maiden, M.C.J. (2010). Impact of recombination on bacterial evolution. Trends Microbiol. *18*, 315–322.

Didelot, X., Méric, G., Falush, D., and Darling, A.E. (2012). Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics *13*, 256.

van Dijl, J.M., and Hecker, M. (2013). Bacillus subtilis: from soil bacterium to super-secreting cell factory. Microb. Cell Fact. *12*, 3.

Dixon, D.A., and Kowalczykowski, S.C. (1993). The recombination hotspot χ is a regulatory sequence that acts by attenuating the nuclease activity of the E. coli RecBCD enzyme. Cell *73*, 87–96.

Dorman, C.J. (2010). 12. Bacterial Chromatin and Gene Regulation. In Bacterial

Chromatin, (Dordrecht: Springer Netherlands), pp. 245–250.

Drahos, D.J., and Hendrix, R.W. (1982). Effect of bacteriophage lambda infection on synthesis of groE protein and other Escherichia coli proteins. J. Bacteriol. *149*, 1050–1063.

Duplay, P., Bedouelle, H., Fowler, A., Zabin, I., Saurin, W., and Hofnung, M. (1984). Sequences of the malE gene and of its product, the maltose-binding protein of Escherichia coli K12. J. Biol. Chem. *259*, 10606–10613.

Eisen, H., Brachet, P., Pereira Da Silva, L., and Jacobt, F. (1970). Regulation of Repressor Expression in X*. Proc. Natl. Acad. Sci. *66*, 855–862.

Eisen, H., Georgiou, M., Georgopoulos, C.P., Selzer, G., Gussin, G., and Herskowitz, I. (1975). The role of gene cro in phage development. Virology *68*, 266–269.

Ellman, G.L. (1959). Tissue sulfhydryl groups. Arch. Biochem. Biophys. *82*, 70–77.

Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, K., and IUCr (2010). Features and development of *Coot*. Acta Crystallogr. Sect. D Biol. Crystallogr. *66*, 486–501.

Erni, B., Zanolari, B., and Kocher, H.P. (1987). The mannose permease of Escherichia coli consists of three different proteins. Amino acid sequence and function in sugar transport, sugar phosphorylation, and penetration of phage lambda DNA. J. Biol. Chem. *262*, 5238–5247.

Esposito, D., and Gerard, G.F. (2003). The Escherichia coli Fis protein stimulates bacteriophage λ integrative recombination in vitro. J. Bacteriol. *185*, 3076–3080.

Evans, G., and Pettifer, R.F. (2001). CHOOCH: A program for deriving anomalous-scattering factors from X-ray fluorescence spectra. J. Appl. Crystallogr. *34*, 82–86.

Feil, E.J. (2004). Small change: keeping pace with microevolution. Nat. Rev. Microbiol. *2*, 483–495.

Ferenci, T., Schwentorat, M., Ullrich, S., and Vilmart, J. (1980). Lambda receptor in the outer membrane of Escherichia coli as a binding protein for maltodextrins and starch polysaccharides. J. Bacteriol. *142*, 521–526.

Fox, J.D., Kapust, R.B., and Waugh, D.S. (2001). Single amino acid substitutions on the surface of Escherichia coli maltose-binding protein can have a profound impact on the solubility of fusion proteins. Protein Sci. *10*, 622–630.

Franklin, N.C. (1974). Altered Reading of Genetic Signals Fused to the N Operon of Bacteriophage 3L : Genetic Evidence for Modifica- tion of Polymerase by the Protein Product of the N gene. J. Mol. Biol *89*, 33–48.

Fruciano, E., and Bourne, S. (2007). Phage as an antimicrobial agent: d'Herelle's heretical theories and their role in the decline of phage prophylaxis in the West. In Canadian Journal of Infectious Diseases and Medical Microbiology, (Hindawi), pp. 19–26.

Georgiou, M., Georgopoulos, C.P., and Eisen, H. (1979). An Analysis of the Tro Phenotype of Bacteriophage A. Virology *94*, 38–54.

Glassberg, J., Meyer, R.R., and Kornberg, A. (1979). Mutant single-strand binding protein of Escherichia coli: genetic and physiological characterization. J. Bacteriol. *140*, 14–19.

Gonda, D.K., and Radding, C.M. (1983). By searching processively RecA protein pairs DNA molecules that share a limited stretch of homology. Cell *34*, 647–654.

Gross, J.D., Karamata, D., and Hempstead, P.G. (1968). Temperature-sensitive mutants of B. subtilis defective in DNA synthesis. Cold Spring Harb. Symp. Quant. Biol. *33*, 307–312.

Haber, J.E. (1999). DNA recombination: The replication connection. Trends Biochem. Sci. *24*, 271–275.

Hames, B., and Rickwood, D. (1998). Gel electrophoresis: a practical approach (Oxford: Oxford University Press).

Hanahan, D. (1983). Studies on Transformation of Escherichia coli with Plasmids. J. Mol. Biol *166*, 557–580.

Handa, N., Ichige, A., and Kobayashi, I. (2009). Contribution of RecFOR machinery of homologous recombination to cell survival after loss of a restriction-modification gene complex. Microbiology *155*, 2320–2332.

Hatfull, G.F., and Hendrix, R.W. (2011). Bacteriophages and their genomes. Curr. Opin. Virol. *1*, 298–303.

Hershey, A.D., and Burgi, E. (1965). Complementary structure of interacting sites at the ends of lambda. Proc. Natl. Acad. Sci. U. S. A. *53*, 325–328.

Hiom, K. (2009). DNA Repair: A RIDDLE at a Double-Strand Break. Curr. Biol. *19*.

Ishigo-oka, D., Ogasawara, N., and Moriya, S. (2001). DnaD Protein of Bacillus subtilis Interacts with DnaA, the Initiator Protein of Replication. J. Bacteriol. *183*, 2148–2150.

Karamata, D., and Gross, J.D. (1970). Isolation and genetic analysis of temperature-sensitive mutants of B. subtilis defective in DNA synthesis. Mol. Gen. Genet. *108*, 277–287.

Katsura, I. (1983). Structure and Inherent Properties of the Bacteriophage Lambda Head Shell IV.? Small-head Mutants. J. Mol. Biol *171*, 297–317.

Koonin, E. V (1992). DnaC protein contains a modified ATP-binding motif and belongs to a novel family of atpases including also DnaA. Nucleic Acids Res. *20*, 1997.

Koonin, E. V, and Galperin, M.Y. (2003). Genomics: From Phage to Human.

Kovall, R., and Matthews, B.W. (1997). Toroidal structure of lambda-exonuclease. Science *277*, 1824–1827.

Kowalczykowski, S.C., Dixon, D.A., Eggleston, A.K., Lauder, S.D., and Rehrauer, W.M. (1994). Biochemistry of homologous recombination in Escherichia coli. Microbiol. Rev. *58*, 401–465.

Krasin, F., and Hutchinson, F. (1981). Repair of DNA double-strand breaks in Escherichia coli cells requires synthesis of proteins that can be induced by UV light. Proc. Natl. Acad. Sci. U. S. A. *78*, 3450–3453.

Kriiger, M., and Hobom, G. (1982). A chain of interlinked genes in the ninR region of bacteriophage lambda. Gene *20*, 25–38.

Krissinel, E., and Henrick, K. (2007). Inference of Macromolecular Assemblies from Crystalline State. J. Mol. Biol. *372*, 774–797.

Kunst, F., and Al., E. (1997). The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature *390*, 249–256.

Kuzminov, A. (1995). Collapse and repair of replication forks in Escherichia coli. Mol. Microbiol. *16*, 373–384.

Kuzminov, A. (1999). Recombinational repair of DNA damage in Escherichia coli and bacteriophage lambda. Microbiol. Mol. Biol. Rev. *63*, 751–813, table of contents.

Laitinen, O.H., Airenne, K.J., Hyotönen, V.P., Peltomaa, E., Mähönen, A.J., Wirth, T., Lind, M.M., Mäkelä, K.A., Toivanen, P.I., Schenkwein, D., et al. (2005). A multipurpose vector system for the screening of libraries in bacteria, insect and mammalian cells and expression in vivo. Nucleic Acids Res. *33*, 1–10.

Learn, B.A., Um, S.J., Huang, L., and McMacken, R. (1997). Cryptic single-stranded-DNA binding activities of the phage lambda P and Escherichia coli DnaC replication initiation proteins facilitate the transfer of E. coli DnaB helicase onto DNA. Proc. Natl. Acad. Sci. U. S. A. *94*, 1154–1159.

Little, J.W. (1967). An exonuclease induced by bacteriophage lambda. II. Nature of the enzymatic reaction. J. Biol. Chem. *242*, 679–686.

Mahdi, a a, Sharples, G.J., Mandal, T.N., and Lloyd, R.G. (1996). Holliday junction resolvases encoded by homologous rusA genes in Escherichia coli K-12 and phage 82. J.

Mol. Biol. *257*, 561–573.

Maniloff, J., and Ackermann, H.W. (1998). Taxonomy of bacterial viruses: Establishment of tailed virus genera and the order Caudovirales. Arch. Virol. *143*, 2051–2063.

Marians, K.J. (1999). PriA: at the crossroads of DNA replication and recombination. Prog. Nucleic Acid Res. Mol. Biol. *63*, 39–67.

Marsin, S.S., McGovern, S., Ehrlich, S.D., Bruand, C., and Polard, P. (2001). Early Steps of Bacillus subtilis Primosome Assembly. J. Biol. Chem. *276*, 45818–45825.

Marston, F.Y., Grainger, W.H., Smits, W.K., Hopcroft, N.H., Green, M., Hounslow, A.M., Grossman, A.D., Craven, C.J., and Soultanas, P. (2010). When simple sequence comparison fails: The cryptic case of the shared domains of the bacterial replication initiation proteins DnaB and DnaD. Nucleic Acids Res. *38*, 6930–6942.

Marszalek, J., and Kaguni, J.M. (1994). DnaA protein directs the binding of DnaB protein in initiation of DNA replication in Escherichia coli. J. Biol. Chem. *269*, 4883–4890.

McCoy, A.J. (2006). Solving structures of protein complexes by molecular replacement with Phaser. In Acta Crystallographica Section D: Biological Crystallography, (International Union of Crystallography), pp. 32–41.

Morimatsu, K., and Kowalczykowski, S.C. (2003). RecFOR proteins load RecA protein onto gapped DNA to accelerate DNA strand exchange: a universal step of recombinational repair. Mol. Cell *11*, 1337–1347.

Mosberg, J.A., Lajoie, M.J., and Church, G.M. (2010). Lambda red recombineering in Escherichia coli occurs through a fully single-stranded intermediate. Genetics *186*, 791–799.

Motamedi, M.R., Szigety, S.K., and Rosenberg, S.M. (1999). Double-strand-break repair recombination in Escherichia coli: Physical evidence for a DNA replication mechanism in vivo. Genes Dev. *13*, 2889–2903.

Mount, D.W., Harris, A.W., Fuerst, C.R., and Siminovitch, L. (1968). Mutations in bacteriophage lambda affecting particle morphogenesis. Virology *35*, 134–149.

Niesen, F.H., Berglund, H., and Vedadi, M. (2007). The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat. Protoc. *2*, 2212–2221.

Oppenheim, A., Belfort, M., Katzir, N., Kass, N., and Oppenheim, A.B. (1977). Interaction of cII, cIII, and cro gene products in the regulation of early and late functions of phage λ. Virology *79*, 426–436.

Otto, S.P., and Michalakis, Y. (1998). The evolution of recombination in changing environments. Trends Ecol. Evol. *13*, 145–151.

Page, R., Peti, W., Wilson, I.A., Stevens, R.C., and Wüthrich, K. (2005). NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. Proc. Natl. Acad. Sci. U. S. A. *102*, 1901–1905.

Postow, L., Hardy, C.D., Arsuaga, J., and Cozzarelli, N.R. (2004). Topological domain structure of the Escherichia coli chromosome. Genes Dev. *18*, 1766–1779.

Poteete, A.R. (2001). What makes the bacteriophage ?? Red system useful for genetic engineering: Molecular mechanism and biological function. FEMS Microbiol. Lett. *201*, 9–14.

Poteete, A.R., Fenton, A.C., and Nadkarni, A. (2004). Chromosomal duplications and cointegrates generated by the bacteriophage lamdba Red system in Escherichia coli K-12. BMC Mol. Biol. *5*, 22.

Proux, C., Van Sinderen, D., Suarez, J., Garcia, P., Ladero, V., Fitzgerald, G.F., Desiere, F., and Brüssow, H. (2002). The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. J. Bacteriol. *184*, 6026–6036.

van de Putte, P., Zwenk, H., and Rörsch, A. (1966). Properties of four mutants of Escherichia coli defective in genetic recombination. Mutat. Res. Mol. Mech. Mutagen. *3*, 381–392.

Rajagopala, S. V, Casjens, S., and Uetz, P. (2011). The protein interaction map of bacteriophage lambda. BMC Microbiol. *11*, 213.

Ramagopal, U.A., Dauter, Z., Thirumuruhan, R., Fedorov, E., Almo, S.C., and IUCr (2005). Radiation-induced site-specific damage of mercury derivatives: phasing and implications. Acta Crystallogr. Sect. D Biol. Crystallogr. *61*, 1289–1298.

Reeve, J.N., and Shaw, J.E. (1979). Lambda encodes an outer membrane protein: The lom gene. Mol. Gen. Genet. *172*, 243–248.

Roberts, J.W. (1969). Termination factor for RNA synthesis. Nature *224*, 1168–1174.

Rocha, E.P.C. (2003). An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: From duplications to genome reduction. Genome Res. *13*, 1123–1132.

Rocha, E.P.C., Cornet, E., and Michel, B. (2005). Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems. PLoS Genet. *1*, e15.

Rohwer, F., and Edwards, R. (2002). The Phage Proteomic Tree: a genome-based

taxonomy for phage. J. Bacteriol. *184*, 4529–4535.

Rokop, M.E., Auchtung, J.M., and Grossman, A.D. (2004). Control of DNA replication initiation by recruitment of an essential initiation protein to the membrane of Bacillus subtilis. Mol. Microbiol. *52*, 1757–1767.

Sandler, S.J. (2000). Multiple genetic pathways for restarting DNA replication forks in Escherichia coli K-12. Genetics *155*, 487–497.

Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. (1982). Nucleotide sequence of bacteriophage ?? DNA. J. Mol. Biol. *162*, 729–773.

Sawitzke, J.A., and Stahl, F.W. (1992). Phage X Has an Analog of Escherichia coli rec0, recR and recF Genes James. Genet. Soc. Americ *16*, 7–16.

Sawitzke, J.A., Thomason, L.C., Costantino, N., Bubunenko, M., Datta, S., and Court, D.L. (2007). Recombineering: In Vivo Genetic Engineering in E. coli, S. enterica, and Beyond. Methods Enzymol. *421*, 171–199.

Sayers, J.R., and Eckstein, F. (1991). A single-strand specific endonuclease activity copurifies with overexpressed T5 D15 exonuclease. Nucleic Acids Res. *19*, 4127–4132.

Schägger, H., Cramer, W.A., and von Jagow, G. (1994). Analysis of molecular masses and oligomeric states of protein complexes by blue native electrophoresis and isolation of membrane protein complexes by two-dimensional native electrophoresis. Anal. Biochem. *217*, 220–230.

Schneider, S., Zhang, W., Soultanas, P., and Paoli, M. (2008). Structure of the N-Terminal Oligomerization Domain of DnaD Reveals a Unique Tetramerization Motif and Provides Insights into Scaffold Formation. J. Mol. Biol. *376*, 1237–1250.

Schwartz, M. (1975). Reversible interaction between coliphage lambda and its receptor protein. J. Mol. Biol. *99*, 185–201.

Seigneur, M., Bidnenko, V., Ehrlich, S.D., and Michel, B. (1998). RuvAB acts at arrested replication forks. Cell *95*, 419–430.

Seitz, H., Weigel, C., and Messer, W. (2000). The interaction domains of the DnaA and DnaB replication proteins of Escherichia coli. Mol. Microbiol. *37*, 1270–1279.

Sharples, G.J., Corbett, L.M., and Graham, I.R. (1998). Rap protein is a structure-specific endonuclease involved in phage recombination. Proc. Natl. Acad. Sci. *95*, 13507–13512.

Sharples, G.J., Curtis, F.A., McGlynn, P., and Bolt, E.L. (2004). Holliday junction binding and resolution by the Rap structure-specific endonuclease of phage λ. J. Mol. Biol. *340*, 739–751.

Sikder, D., Unniraman, S., Bhaduri, T., and Nagaraja, V. (2001). Functional cooperation between topoisomerase I and single strand DNA-binding protein. J. Mol. Biol. *306*, 669–679.

Smits, W.K., Goranov, A.I., and Grossman, A.D. (2010). Ordered association of helicase loader proteins with the Bacillus subtilis origin of replication in vivo. Mol. Microbiol. *75*, 452–461.

Smits, W.K., Merrikh, H., Bonilla, C.Y., and Grossman, A.D. (2011). Primosomal proteins DnaD and DnaB are recruited to chromosomal regions bound by DnaA in Bacillus subtilis. J. Bacteriol. *193*, 640–648.

Spies, M., and Kowalczykowski, S.C. (2005). Homologous Recombination by the RecBCD and RecF Pathways. Bact. Chromosom. 389–404.

Stahl, F.W., Kobayashi, I., and Stahl, M.M. (1985). In phage lambda, cos is a recombinator in the red pathway. J. Mol. Biol. *181*, 199–209.

Stahl, M.M., Thomason, L., Poteete, A.R., Tarkowski, T., Kuzminov, A., and Stahl, F.W. (1997). Annealing vs. invasion in phage ?? recombination. Genetics *147*, 961–977.

Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev. *24*, 814–826.

Studier, F.W., and Moffatt, B.A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol. *189*, 113–130.

Sutton, M.D., Carr, K.M., Vicente, M., and Kaguni, J.M. (1998). Escherichia coli DnaA protein. The N-terminal domain and loading of DnaB helicase at the E. coli chromosomal origin. J. Biol. Chem. *273*, 34255–34262.

Szczepańska, A.K. (2009). Bacteriophage-encoded functions engaged in initiation of homologous recombination events. Crit. Rev. Microbiol. *35*, 197–220.

Takahashi, N., and Kobayashi, I. (1990). Evidence for the double-strand break repair model of bacteriophage lambda recombination. Proc. Natl. Acad. Sci. *87*, 2790–2794.

Tarkowski, T.A., Mooney, D., Thomason, L.C., and Stahl, F.W. (2002). Gene products encoded in the ninR region of phage λ participate in red-mediated recombination. Genes to Cells *7*, 351–363.

Thaler, D.S., Stahl, M.M., and Stahl, F.W. (1987). Double-chain-cut sites are recombination hotspots in the red pathway of phage λ. J. Mol. Biol. *195*, 75–87.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised genome dynamics in the

Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. *5*, e1000344.

Trotter, A., and John (2015). DNA binding proteins of the phage λ ninR region. Durham University.

Turner, I.J., Scott, D.J., Allen, S., Roberts, C.J., and Soultanas, P. (2004). The Bacillus subtilis DnaD protein: A putative link between DNA remodeling and initiation of DNA replication. FEBS Lett. *577*, 460–464.

Umezu, K., Chi, N.W., and Kolodner, R.D. (1993). Biochemical interaction of the Escherichia coli RecF, RecO, and RecR proteins with RecA protein and single-stranded DNA binding protein. Proc. Natl. Acad. Sci. U. S. A. *90*, 3875–3879.

Vedadi, M., Niesen, F.H., Allali-Hassani, A., Fedorov, O.Y., Finerty, P.J., Wasney, G.A., Yeung, R., Arrowsmith, C., Ball, L.J., Berglund, H., et al. (2006). Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. Proc. Natl. Acad. Sci. *103*, 15835–15840.

Wang, L., Wuerffel, R., Feldman, S., Khamlichi, A.A., and Kenter, A.L. (2009). S region sequence, RNA polymerase II, and histone modifications create chromatin accessibility during class switch recombination. J. Exp. Med. *206*, 1817–1830.

Watt, V.M., Ingles, C.J., Urdea, M.S., and Rutter, W.J. (1985). Homology requirements for recombination in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *82*, 4768–4772.

White, M.F., Phe, M.-J.Á., Giraud-Panis, E., Richard, J., Po, G., Hler, È., and Lilley, D.M.J. (1997). Review Article: Recognition and Manipulation of Branched DNA Structure by Junction-resolving Enzymes. J. Mol. Biol *269*, 647–664.

Williams, P.H. (1980). Black rot: a continuing threat to world crucifers. Plant Dis. *64 No. 8*, 736-.

Winship, P.R. (1989). An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. Nucleic Acids Res. *17*, 1266.

Wu, R., and Kaiser, A.D. (1968). Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J. Mol. Biol. *35*, 523–537.

Wu, A.M., Kahn, R., Dasgupta, C., and Radding, C.M. (1982). Formation of nascent heteroduplex structures by RecA protein and DNA. Cell *30*, 37–44.

Zhang, W., Carneiro, M.J.V.M., Turner, I.J., Allen, S., Roberts, C.J., and Soultanas, P. (2005). The Bacillus subtilis DnaD and DnaB proteins exhibit different DNA remodelling activities. J. Mol. Biol. *351*, 66–75.

Zhang, W., Machón, C., Orta, A., Phillips, N., Roberts, C.J., Allen, S., and Soultanas, P. (2008). Single-Molecule Atomic Force Spectroscopy Reveals that DnaD Forms Scaffolds and Enhances Duplex Melting. J. Mol. Biol. *377*, 706–714.

Zhang, Y.W., Hunter, T., and Abraham, R.T. (2006). Turning the replication checkpoint on and off. Cell Cycle *5*, 125–128.