Structural studies on
Glutamine de-amidase enzymes

George William Mobbs
MBiolSci

A thesis submitted to the University of Sheffield in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Molecular Biology and Biotechnology
September 2014

Abstract:

The recent discovery of Burkholderia lethal factor 1 (BLF1) a potent glutamine de-amidase toxin, provides the third example in the putative glutamine de-amidase superfamily comprising two toxin examples and a key regulator of chemotaxis. Both toxin members inactivate disparate but essential cellular processes; yet have diverged at a genetic level to such a degree that identification of additional family members presented a substantial challenge. Structural comparisons between the 3 current members of the super-family: BLF1, Cytotoxic necrotising factor 1 (CNF1) and CheD have led to the design of a 1^{o} sequence fingerprint, with which other members of the family might be identified. Sequence searches using this motif have led to the identification of a number of putative targets, which were rank ordered for structural studies.

Of these putative targets a single novel member of the super-family has been identified, called HCH_03101. This protein is produced by a recently discovered marine bacterium *Hahella. chejuensis*, which is implicated in the destruction of harmful algal blooms through the secretion of secondary metabolites. The structure of HCH_03101 shares striking similarity to the toxin BLF1, apart from at a long insertion protruding directly outwards from the globular body by 44 Å. This protrusion conserves a repeated proline rich motif mirrored across opposite strands, which forms a remarkable structural motif displaying a pseudo-2-fold axis of symmetry. Functional studies reveal that HCH_03101 binds to and de-amidates Eukaryotic initiation factor 4a (eIF4a) at the same site-specific glutamine position as BLF1. Furthermore, like BLF1, HCH_03101 also exhibits no signal peptides or additional domains capable of mediating cellular secretion. Thus, it is possible that *H. chejuensis* is also capable of occupying an intracellular niche as a pathogenic organism.

This thesis also presents the structure of BLF1 in complex with its native substrate human eIF4a-I, which is the first substrate bound structure for this novel family of toxins and the highest resolution data yet elucidated for eIF4a at 2.5 Å. Revealing that BLF1 co-ordinates its glutamine substrate through formation of a oxyanion hole, prior to forming an tetrahedral acyl intermediate via nucleophilic attack by a catalytic CYS – HIS dyad, a catalytic mechanism shared with the cysteine protease papain.

Acknowledgements:

First and clearly foremost my thanks and sincerest gratitude are extended to my supervisor Professor David Rice. Dave is a special kind of boss, his infectious enthusiasm follows him everywhere he goes and even on my off days I couldn't help but get caught up in it!

Thanks must also go to the unbelievable team that is the crystallography group, especially Dr Patrick Baker and Dr John Rafferty not only for all the help I have sponged off them, but also for the lighthearted zany atmosphere that is so prevalent around this group, the last 4 years have been wonderful fun. I would also like to thank Professor Pete Artymiuk for all the wacky and downright weird conversations held at coffee over the years, alas the struggle must go on, but you are missed. There are also many collaborators deserving of my appreciation amongst them: Dr Lynda Partridge, Dr Mark Dickman and Dr Jim Gilmour who have all directly contributed to the work presented in this thesis. On a technical side I also thank my constant collaborator and teacher Dr Svetlana Sedelnikova, from whom I have learnt an enormous amount. Thanks must also go to Dr Rosie Staniforth, who has always been willing to lend an ear particularly during this last year. I would also like to thank the BBSRC and Diamond light source for providing funding and an outrageously advanced facility full of the most amazing scientists and support.

Sheffield has been my home for the past 8 years and I have loved living in this city. A large part of that has been due to the amazing people I have had the good fortune to meet. Prominent amongst them: Oliver Woodman, Robert Davies and Joel Patterson you guys rock! The entire Sheffield University Bankers Mens 7th team, without whom I'm certain I would have lost my mind. With special mentions to both Shuo Jiang and Jim Wheelwright two close friends, who have all this to look forward to! Then it comes time to thank the crew, my co-conspirators in mass negligence. Special thanks must go to Sophie Bliss, Abi Williams, Jason Wilson, Hayley Owen and especially Claudine Bisson for all their help and support, especially this last year you have all been immense.

Last but never least, I would like to thank my incomparable family whom I love dearly and draw so much of my personal strength and better nature from. With this in mind I dedicate this thesis to both my grandfathers, who influenced me beyond measure and always emphasized the enjoyment derived from dedication, hard work and a job well done. Love to you all George. X

Contents:

List of figures and tables:

Chapter 1: Introduction

Chapter 2: Bioinformatics and target selection

Chapter 3: X-ray theory

List of abbreviations:

ASU – Asymmetric unit

BBSRC – Biotechnology and Biological sciences Research council

BLAST – Basic local alignment search tool

BPSL – *Burkholdria pseudomallei* Large chromosome

B. pseudomallei – Burkholderia pseudomallei

BSA – Bovine serum albumin

Bp – Base pair

CC – Correlation co-efficient

CCD – Charge coupled device

CCP4 – Collaborative computational project No. 4

CDC – Centre for disease control

CFE – Cell free extract

DEAE FF – Diethylaminoethyl fast flow

DMSO – Dimethyl sulphoxide

DNA – Deoxyribonucleic acid

DSLS – Diamond synchrotron light source

DTT – Dithiothreitol

E. coli - Escherichia coli

EDTA – Ethylenediaminetetraacetic acid

FT – Fourier transform

GF – Gel filtration

GUI – Graphical user interface

HEPES – 2-(2-hyrdroxyethyl)-1-piperazineethanesulfonic acid

ID – Identity

IPTG – Isopropyl β-D-1-thioglactopyranoside

LB – Lysogeny broth

MAD – Multi-wavelength anomalous dispersion

MALDI-TOF – Matrix assisted laser desorption / ionization time of flight

MBB – Molecular Biology and Biotechnology

MIR – Multiple isomorphous replacement

ML – Maximum liklihood

MPD – 2 – methyl-2-4-pentandiol

mRNA – Messenger RNA

MR – Molecular replacement

MRC – Medical research council

MSE – Measuring and scientific equipment

MW – Molecular weight

MQ-H_2O – MilliQ purified ionized water

NCBI – National centre for Biotechnological information

NCS – Non crystallographic symmetry

NEB – New England biolabs

NTA – Nitrolotriacetic acid

NTP – Nucleotide triphosphate

OD – Optical density

PCR – Polymerase chain reaction

PDB – Protein data bank

PEG – Polyethtyleneglycol

pI – isoelectric point

POW – Prisoner of war

RMSD – Root mean square deviation

SAD – Single-wavelength anomalous dispersion

SDS-PAGE – Sodium dodecyl sulphate polyacrylamide gel electrophoresis

Se-MET – Selenium derivative of methionine

T4 Ligase – Ligase from T4 bacteriophage

TAE – Tris-acetic acid EDTA buffer

Tris – Tris(hydroxymethyl)aminomethane

UV – Ultra violet

Chapter 1: Introduction

This chapter will offer a brief overview (Figure 1.1.1) of the three Glutamine de-amidase enzymes currently characterised, two of which are potent toxins with the third regulating chemotaxis across all prokaryotes.

1.1 - Toxins

The term toxin is defined in the Oxford Dictionary as: 'A poison of plant or animal origin, especially one produced by or derived from microorganisms and acting as an antigen in the body'. Bacteria commonly employ toxins to promote virulence, typically by modulating the host's cellular machinery in favour of the pathogen. There are two classes of toxin, endo and exo-toxins. Endotoxins are not secreted instead they are found associated with the membranes of many gram-positive bacteria. An example of an endo-toxin is the Lipopolysaccharide family found in pathogenic bacteria like *S. aureus*, where they interact with immune cell receptors causing increased cytokine secretion and local inflammation (Schumann, 1992). Exo-toxins on the other hand are secreted by the pathogen and then further classified based on how they interact with the host, into three distinct classes: The first class is cell surface active, interacting with receptors on the outer membrane of the host. The second is the membrane damaging class, which forms holes in the host cell membrane allowing either secondary secretions access to the cytoplasm or causing cell lysis. The final class of exo-toxin is the intracellular toxin, which requires access to the hosts cytoplasm in order to effect virulence.

In the case of pore-forming toxins the presented phenotype is clear. However many intracellular toxins are extremely subtle and challenging to characterise. Two members of the Glutamine de-amidase family are type III intracellular exotoxins, presenting a straightforward site-specific de-amidation phenotype. However, the knock on effects of this de-amidation are dramatic, as they both disrupt essential processes in the host cell, making them a fascinating topic to review.

1.2.1 – Cytotoxic necrotising factor – CNF1

Cytotoxic necrotising factor 1 (CNF1) was first characterised in the early eighties as a toxin capable of inducing significant cyto-skeletal reorganisation in Eukaryotes (Caprioli *et al.,* 1983, Fiorentini *et al.,* 1988). It is expressed by uropathogenic *E. coli* strains,

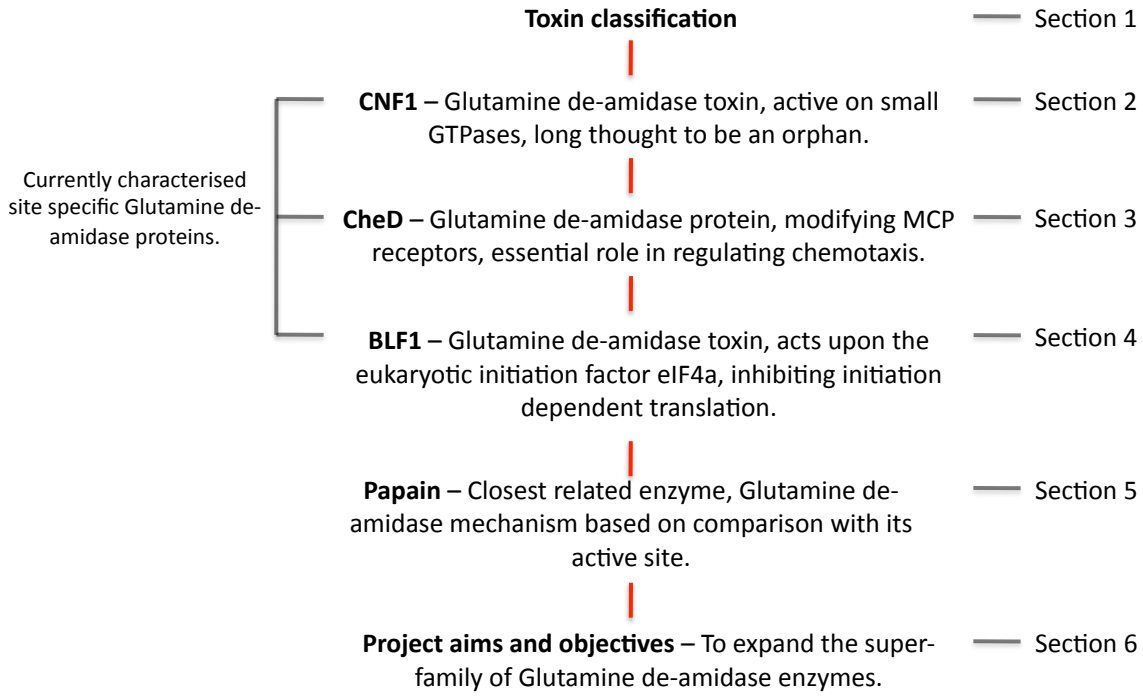


Figure 1.1.1 – Flow chart outlying the organisation of this introduction. This chapter introduces the Glutamine de-amidase protein super-family and has been split into 6 subsections. The first section will offer a brief overview of the toxin classification system, with sections 2, 3 and 4 examining the currently characterised Glutamine de-amidase enzymes CNF1, CheD and BLF1 respectively. Section 5 will then draw parallels between the Glutamine de-amidase family and the well characterised Cysteine protease Papain, with which they share a catalytic dyad. The chapter will then conclude with section 6 laying down the aims and objectives of this study.

implicated in harmful urinary tract infections (Boquet, 2001). CNF1 is encoded for by a 3042 bp gene (Falbo *et al.,* 1993), located within a pathogenicity island (Blum *et al.,* 1995). This gene encodes for a 1014 residue A/B type toxin, with three distinct domains (Falbo *et al.,* 1993) and a combined molecular weight of 108 KDa (figure 1.2.1).

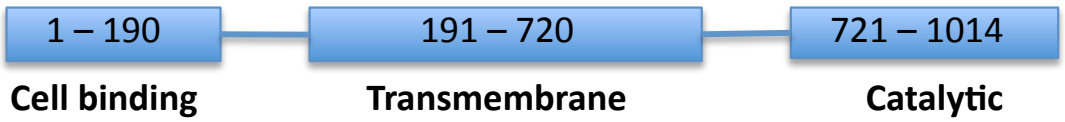


Figure 1.2.1 – CNF1 is an A/B type toxin composed of three domains. The N terminal domain is 190 residues long and responsible for cell binding (Lemichez *et al.,* 1997). The central domain extends from residues 191 to 720 and incorporates two spans of hydrophobic residues, organised into transmembrane helices separated by a hydrophilic loop, suggesting membrane translocation (Choe *et al.,* 1992). This transmembrane region is believed to function in a similar fashion to Diphtheria toxin, with the hydrophilic loop becoming protonated in low pH conditions allowing membrane translocation (Pei *et al.,* 2001). The catalytic domain (C-CNF1) is located at the C–terminus, between residues 720 and 1014 (Lemichez *et al.,* 1997) and de-amidates site specific GLN side chains.

1.2.2 - C-CNF1 function

C-CNF1 constitutively activates the small GTP binding proteins RhoA, Rac and Cdc42, leading to the formation of stress fibres and adhesions (figure 1.2.2) in most cell lines (Caprioli *et al.,* 1983). Typically RhoA activation is characterised by the formation of stress fibres, Rac with the formation of membrane ruffles and Cdc42 with membrane outcroppings called Filopodia. Cells infected with C-CNF1 have been observed exhibiting all the above phenotypes (Flatau *et al.,* 1997; Schmidt *et al.,* 1997) along with multi-nucleation and aneuploidy (D'Amato and Patarua, 1998).

These observed phenotypes are caused by the site-specific de-amidation of GLN 63 in RhoA and GLN 61 in both Rac and Cdc42 (Flatau *et al.,* 1997). De-amidation is the irreversible conversion of a side chain amide into a carboxylate moiety, which has the effect of switching a polar side-chain for a negatively charged polar acidic side chain (Washington *et al.,* 2013).

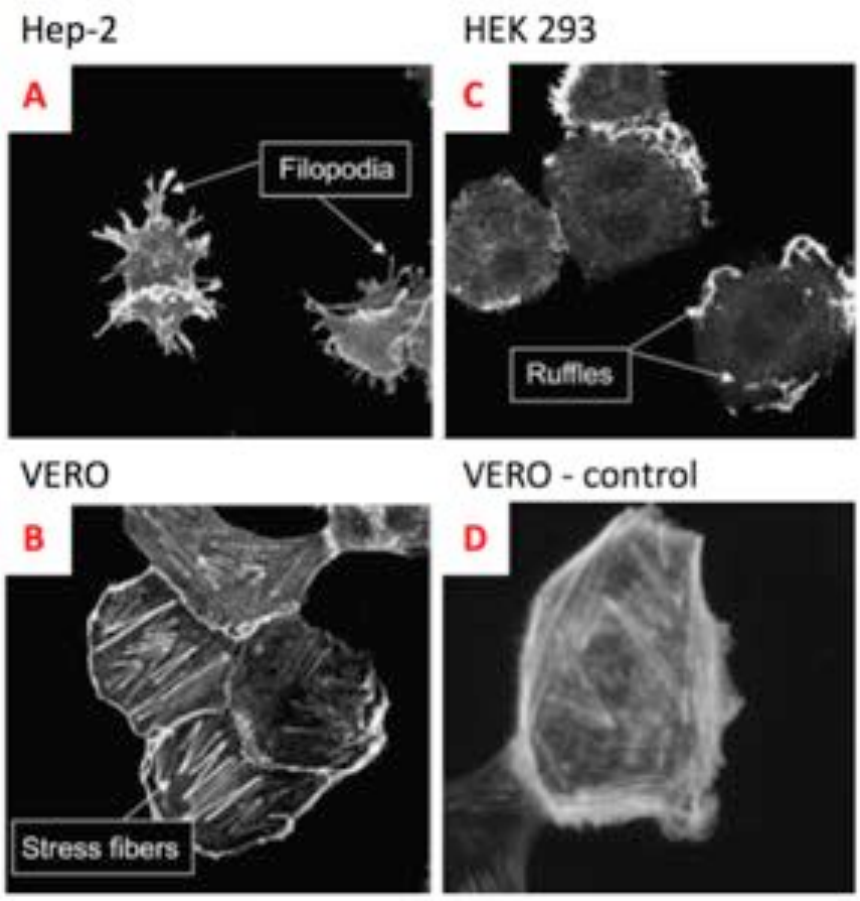


Figure 1.2.2 – Physiological effects of CNF1 infection on Eukaryotic cells. The above images are fluorescence micrographs of Eukaryotic cells, incubated with 10^{-9} M C-CNF1 for 24 hrs and then mixed with a fluorescent label targeting actin. Hep-2 cells (panel **A**) display actin-rich outgrowths called Filopodia caused by Cdc42 over-expression. The VERO cells (panel **B**) show high stress fibre content corresponding to RhoA activation, with the HEK293 cells (panel **C**) exhibiting membrane ruffles a hallmark of Rac1 over-expression. Panel **D** shows a non-infected VERO control cell, which does not exhibit any of the above morphological alterations. Adapted from (Lemonnier *et al.,* 2007).

The target GLN in all three small GTPases is conserved and involved in stabilising the transition state between Rho GTP and a water molecule interacting with the complex during GTP hydrolysis (Buetow *et al.,* 2001). Modifications at this location preclude GTPase interaction with the GTPase activating proteins (GAP), which mediate hydrolysis of the bound GTP into GDP. As a result the small GTPase remains constitutively active and its downstream effectors improperly regulated (figure 1.2.3).

Currently no indiscriminate bacterial glutamine de-amidase enzymes have been identified (Washington *et al.,* 2013). The apparent recognition site for C-CNF1 is fairly compact with a short oligo-peptide analogous to RhoA, with 5 amino acids either side of the modified GLN, the smallest viable substrate (Flatau *et al.,* 2000). The multiple substrate specificity exhibited by C-CNF1 is likely a consequence of this limited recognition site. However, site-specific de-amidation of GLN residues is not catalytically novel; for example, it is an important stage in several transglutaminase reactions such as with blood coagulation factor XIII (Yee *et al.,* 1994). Despite this toxins exhibiting site-specific glutamine de-amidase activity had not previously been reported. Early attempts at elucidating the catalytic mechanism of C-CNF1 involved broad site directed mutagenesis experiments, highlighting two essential residues CYS 866 and HIS 881 (Schmidt *et al.,* 1998). However, the possible roles these residues played in catalysis were not fully understood until the structure of C-CNF1 was solved (Buetow *et al.,* 2001).

1.2.3 - CNF1 structure

The C-CNF1 structure was solved using X-ray crystallography with a quadruple MET mutant, Se-MET derivative and MAD phase solution, then refined to a resolution of 1.83 Å. It forms a compact globular domain composed of a central β-sandwich, made from opposing β-sheets flanked by α-helices (figure 1.2.4). The β-sandwich portion of the protein is formed from two mixed β-sheets comprising 6 and 7 strands respectively. The active site contains a catalytic CYS – HIS dyad, supported by an underlying TYR residue (figure 1.2.4). CYS 866 is believed to be responsible for a nucleophilic attack at the δ-carboamide of the target GLN, with HIS 881 thought to hydrogen bond with CYS 866 forming a thiol-imidazolate ion pair. Mutation at the CYS position leads to a complete loss of catalysis, suggesting that CYS 866 is an essential catalytic residue (Buetow *et al.,* 2001). VAL 833 plays a role in orienting the imidazole ring of HIS 881 to interact with the thiol moiety of CYS 866, both VAL 833 and HIS 881 are located parallel to one another on opposing β-strands.

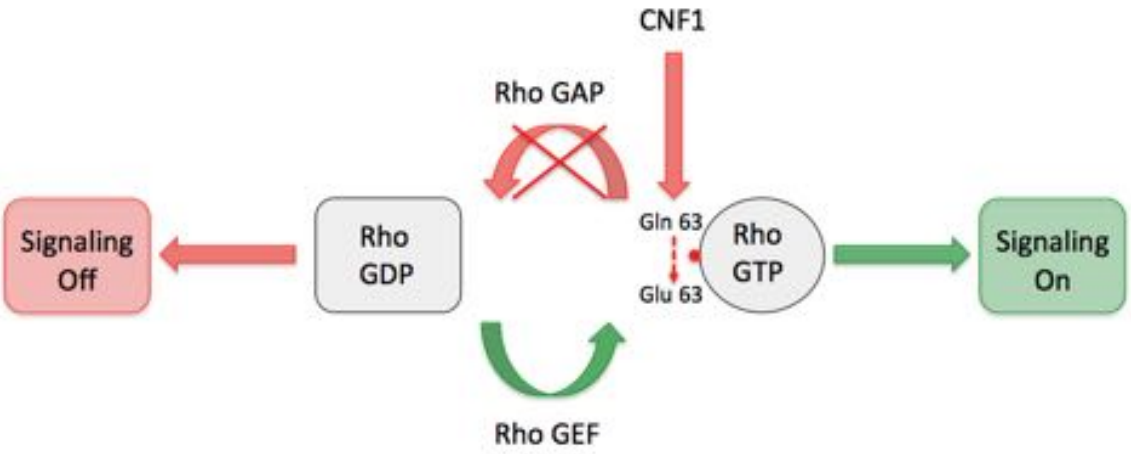


Figure 1.2.3 - Schematic of C-CNF1 activity on Rho GTP binding proteins. C-CNF1 acts upon the small GTPase regulation cycle. The step inhibited by C-CNF1 is labeled red, with the stages permitted highlighted in green. The de-amidation of GLN 63 in RhoA to GLU prevents the turn over of GTP into GDP by the GTPase activating proteins (GAP), whilst the activity of Guanine exchange factors (GEF) is un-impeded. This skews the equilibrium between active and inactive small GTPases towards an active state, which leads to the inflated expression of cytoskeleton components observed in infected host cells. Adapted from (Boquet, 2001).

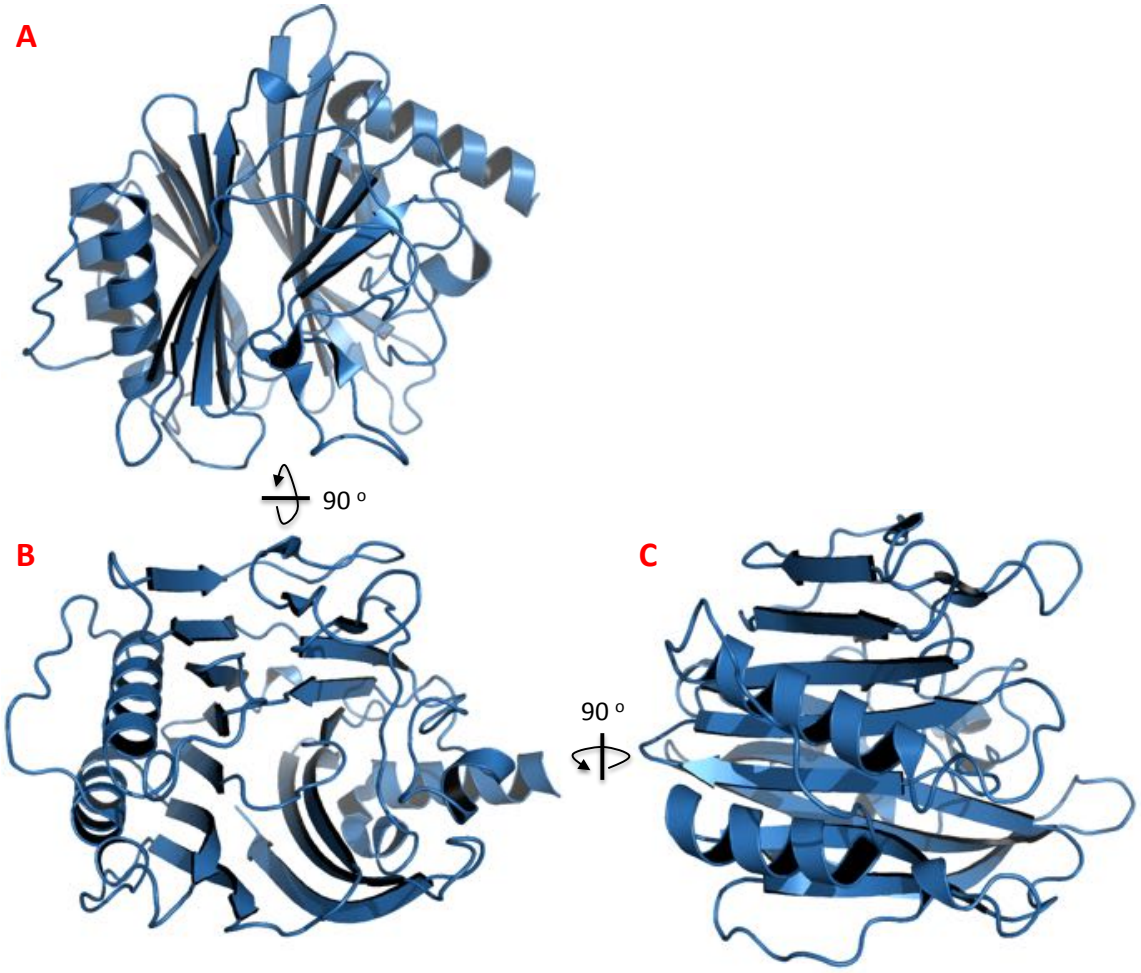


Figure 1.2.3 – 3D representation of C-CNF1. C-CNF1 is a type III exotoxin, exhibiting A/B type translocation into the host, expressed by uropathogenic *E. coli*. It is composed of three domains the C-terminal catalytic domain C-CNF1 is shown above. The most striking element of the C-CNF1 structure is the central β-sandwich, shown from the top down in panel **A**. This β-sandwich is composed of two mixed β-sheets comprising 7 strands on the left flank and 6 strands on the right, which is best shown in panel **B**. This central β-strand motif is then flanked by α-helices. Diagram constructed in Pymol using PDB ID: 1HQ0 (Buetow *et al.,* 2001).

Unlike the other catalytic residues CYS 866 is located on a loop between two β-strands, supported from underneath by TYR 962 through a hydrogen bond with the backbone carbonyl of CYS 866.

Previous studies had shown that the recognition surface of C-CNF1 for the small GTPases is relatively small (Flatau *et al.,* 2000). This observation is further corroborated by the shallow active site cleft exhibited in both surface and electrostatic maps of C-CNF1 (figure 1.2.5). Site directed mutagenesis of individual amino acids surrounding the active site, did not identify an obvious region for substrate recognition (Buetow *et al.,* 2001). However, crude loop knockouts (figure 1.2.5C) showed that the deletion of two loops proximal to the active site eliminates function, indicating that these two loops may play a role in substrate recognition (Beutow and Ghosh, 2003).

Sequence comparison has identified CNF1 homologues in a number of pathogens, but structural comparison detected no protein super families with a similar tertiary fold, which led to CNF1 initially being classed as an orphan. At this point in time, C-CNF1 was believed to have evolved convergently from the cysteine proteases, with which it shares the same catalytic dyad arrangement. However, recently several novel glutamine de-amidase enzymes have been uncovered, suggesting that CNF1 instead belongs to a larger super-family that has evolved divergently from a common ancestor (Washington *et al.,* 2013).

1.2.4 - CNF1 homologues and biological significance

Across the *E. coli* genus there are a variety of strains encoding homologues of CNF1. For example, CNF2 is a close homologue with 85 % sequence similarity and is produced by entero-pathogenic *E. coli* (Oswold *et al.,* 1994). CNF3 is another close homologue produced by necrotoxigenic *E. coli*, which displays 70 % sequence similarity (Orden *et al.,* 2007). This group of closely related homologues also extends further than the *E. coli* genus. For example, *Yersinia. Pseudotuberculosis* the causative agent of Far East scarlet-like fever (a disease closely resembling tuberculosis) produces a homologue with 61 % sequence similarity called CNFY (Lockman *et al.,* 2002). The final functional homologue of CNF1, sharing 40 % sequence similarity, is Dermonecrotic toxin (DNT) produced by *Bordetella. pertussis* (Masuda *et al.,* 2000). The homologues listed above all target glutamine residues in small GTPases.

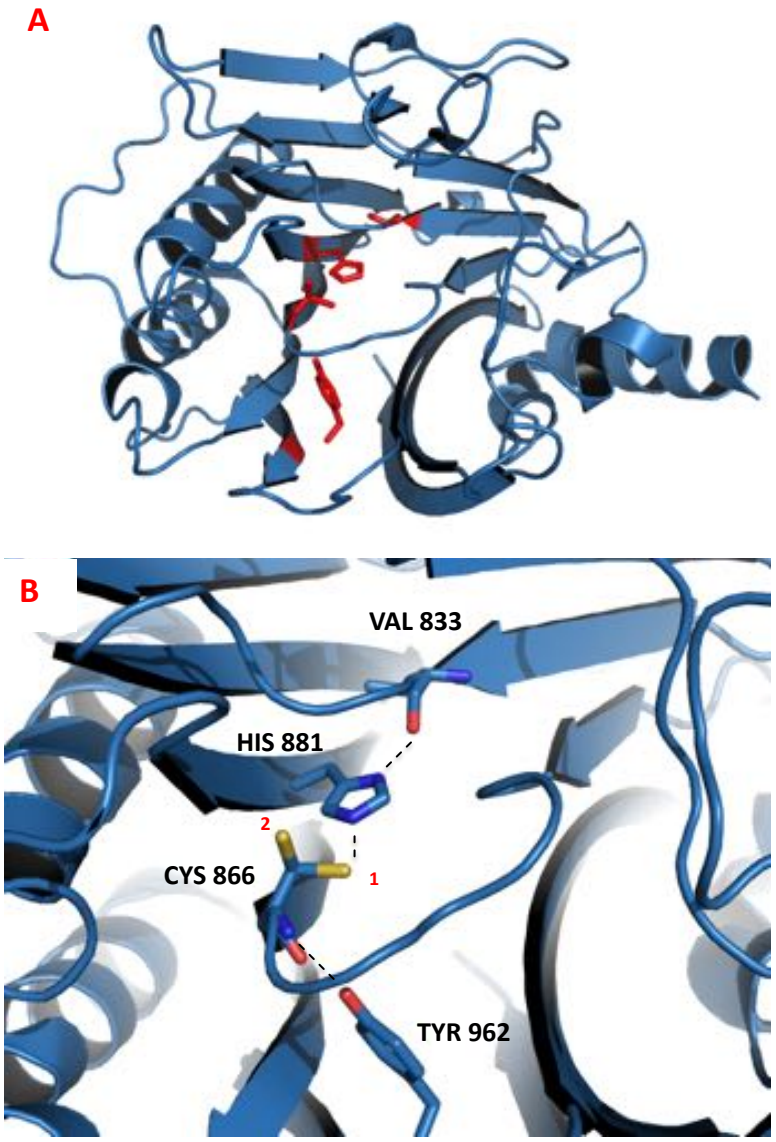


Figure 1.2.4 – C-CNF1 exhibits a catalytic dyad with supporting residues located on both sides of the central β-sandwich. C-CNF1 has four functionally important residues (panel **A**) with CYS 866 and HIS 881 forming an essential catalytic dyad, which is supported from either side by TYR 962 and VAL 833. The active site is located centrally, with key residues on both sides of the β-sandwich and CYS 866 located on a loop. Panel **B** displays the hydrogen bonds made between the catalytic dyad and their two stabilising residues. HIS 881 is stabilised in a location to co-ordinate with CYS 866 through a hydrogen bond between the imidazole rings Nε and the backbone carbonyl of VAL 833. The catalytic CYS is located on a loop and stabilised through a hydrogen bond between the side chain carbonyl of TYR 962 with backbone amide of CYS 866. Diagram constructed in Pymol using PDB: H1Q0 (Buetow *et al.,* 2001).

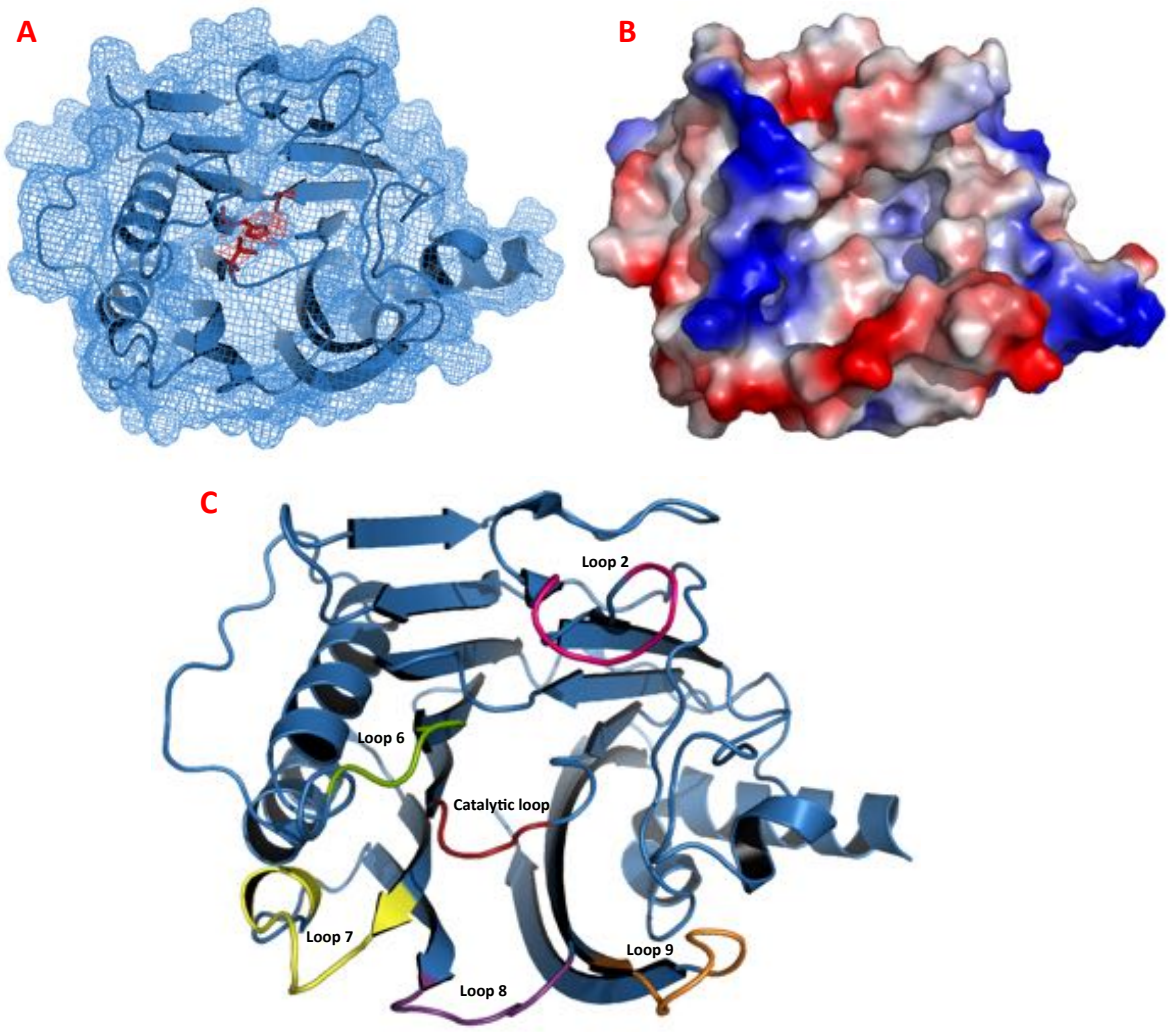


Figure 1.2.5 – Examination of the active site cleft of C-CNF1. It has previously been shown that a short, 10 amino acid, oligo-peptide analogous to RhoA can be de-amidated by C-CNF1. This indicates a narrow range of substrate specificity; which is borne out by C-CNF1s affinity for multiple small GTPases. Panel **A** displays a surface mesh model of C-CNF1 with the catalytic dyad (CYS 866 and HIS 881) and VAL 833 coloured red. Panel **B** is the same view of C-CNF1 but with the surface mesh substituted with an electrostatic charge diagram, with red corresponding to negative charges and blue relating to positive charges. This diagram demonstrates that the active site cleft is neutral, but flanked by positively charged ridges. Panel **C** displays the loops surrounding the active site cleft, mutagenesis experiments deleting each loop in isolation indicated that loops 8 (purple) and 9 (orange) are involved in substrate recognition. Diagram constructed in Pymol using PDB ID: H1Q0 (Beutow *et al.,* 2001; Buetow and Ghosh, 2003).

However, not all of these examples display identical affinity for all three of the small GTPase substrates. CNF3 binds far more tightly to RhoA (Stoll *et al.,* 2009), whilst CNF2 does not bind Rac at all (Oswold *et al.,* 1994).

De-regulation of the small GTPase signalling pathway and its effect on *E. coli* pathogenesis is presently under significant scrutiny. For example in mouse infection models ΔCNF1 knockouts display reduced colonisation in liver and bladder cells (Rippere-Lampe *et al.,* 2001). It was later shown that CNF1 is required for *E. coli* to gain access to the uroepithelial cells (Doye *et al.,* 2002) (figure 1.2.6A), via a mode of action believed to mimic pathogens from the salmonella genus (Galan and Zhou, 2000). These reports suggest that CNF1 is an essential virulence factor for the promotion of host invasion at endothelial locations.

However, this mode of action is complicated by reports of de-amidated RhoA becoming ubiquitylated in host cells, triggering degredation of RhoA and subsequent de-regulation of its downstream effectors (Landraud *et al.,* 2004). Therefore, the impact of CNF1 on *E.coli* virulence is clearly not as straightforward as had previously been expected. This trend is further examined in a variety of recent publications, with CNF1 shown to be involved in a multitude of non-uropathogenic infections. For example, CNF1 is an essential virulence factor in the K1 strain of *E.coli*, where along with Ferredoxin it promotes invasion of the endothelial cells in human brains. This colonisation of the brain is a critical stage in the pathogens progression to the central nervous system and the eventual development of *E. coli* mediated Meningitis (Yu and Kim, 2010). CNF1 is clearly an extremely versatile virulence factor (figure 1.2.7) and has even been shown to have a role in the onset of oncogenesis. This role in cancer is due to the downstream disruption of mitosis, where CNF1 has been shown to increase the occurrence of aneuploidy and multi-nucleation in infected cells (Malorni and Fiorentini, 2006).

1.3.1 - CheD

CheD was the next enzyme to be characterised with a tertiary fold and suggested catalytic mechanism in common with C-CNF1. CheD is a 16 KDa protein that unlike C-CNF1 is neither a toxin nor particularly isolated phylogenetically, as it is present in most Prokaryotes as an essential component in the chemotaxis signalling cascade. The latter has been shown by a lack of motility in CheD knockout mutants (figure 1.3.1) (Kirby *et al.,* 2001). CheD primarily functions through modifying the receptors responsible for sensing environmental stimuli, but it also plays a role in regulating the downstream effectors that mediate flagella rotation.

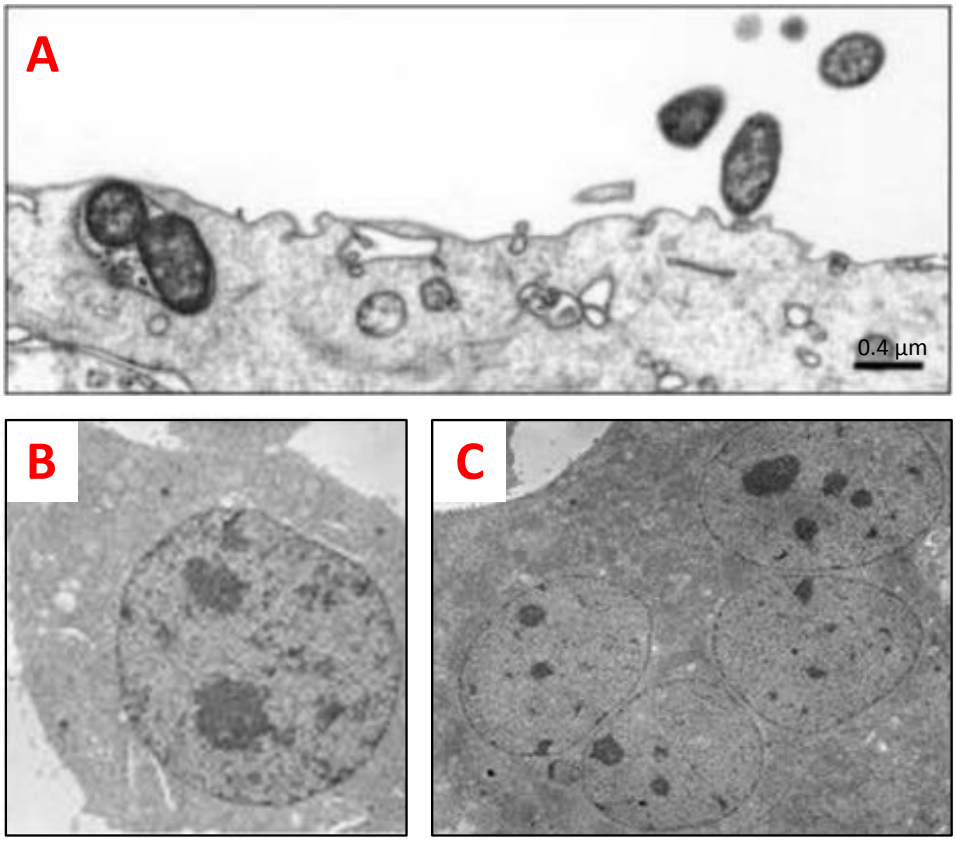


Figure 1.2.6 – C-CNF1 has a role in *E.coli* virulence through increased invasiveness and disruption of the host cell cycle. Panel **A** displays a transmission electron micrograph of *E. coli* strain UPEC 58A1 invading 804G epithelial bladder cells. The pathogenic UPEC cells are treated with 10^{-10} M C-CNF1 overnight and then incubated with the epithelial cells for 2 hours, prior to preparation for electron microscopy (Landraud *et al.,* 2004). Panel **B** and **C** display the multi-nucleation commonly exhibited by CNF1 infected host cells. Panel **B** is non-infected control with a single nucleus, whereas the infected cell shown in panel **C** exhibits 4 nuclei (Travaglione *et al.,* 2008).

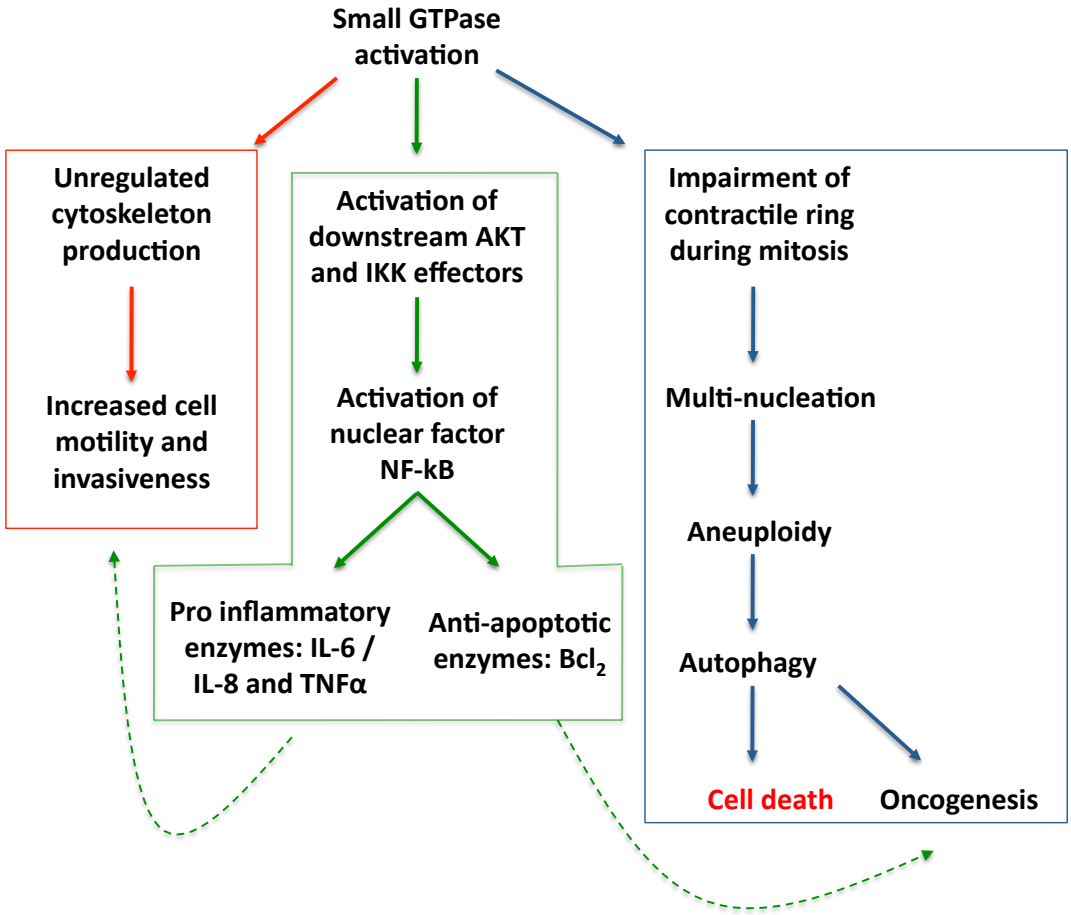


Figure 1.2.7 – Flow chart exploring the multifaceted impact of CNF1 upon *E.coli* virulence. Constitutive activation and downstream deregulation of the small GTPase enzymes RhoA, Rac and Cdc42 has been shown to induce cytoskeletal stress. However, further study has shown that disruption of this signalling pathway has wide spread effects, on both the hosts cell division and ability to resist invasion. The observed increase in *E. coli* invasion is directly linked to changes made to the host cells cytoskeleton (highlighted in red). Nevertheless, C-CNF1 is also hypothesised to play a role in oncogenesis, due to the tendency of infected cells to exhibit multi-nucleation and aneuploidy (highlighted in blue). However, there is also a third route of virulence (highlighted in green), which involves activating the AKT and IKK signalling cascade. This pathway leads to the activation of nuclear transcription factor NF-kB, which in turn up regulates transcription of the IL-8, IL-9 and TNFα genes involved in inflammation and Bcl_2 a prominent anti-apoptosis effector. These unregulated proteins have a direct influence on both of the previously established virulence pathways (indicated with dotted green arrows). Figure adapted from diagrams found in (Travaglione *et al.,* 2008; Malorni Fiorentini, 2006).

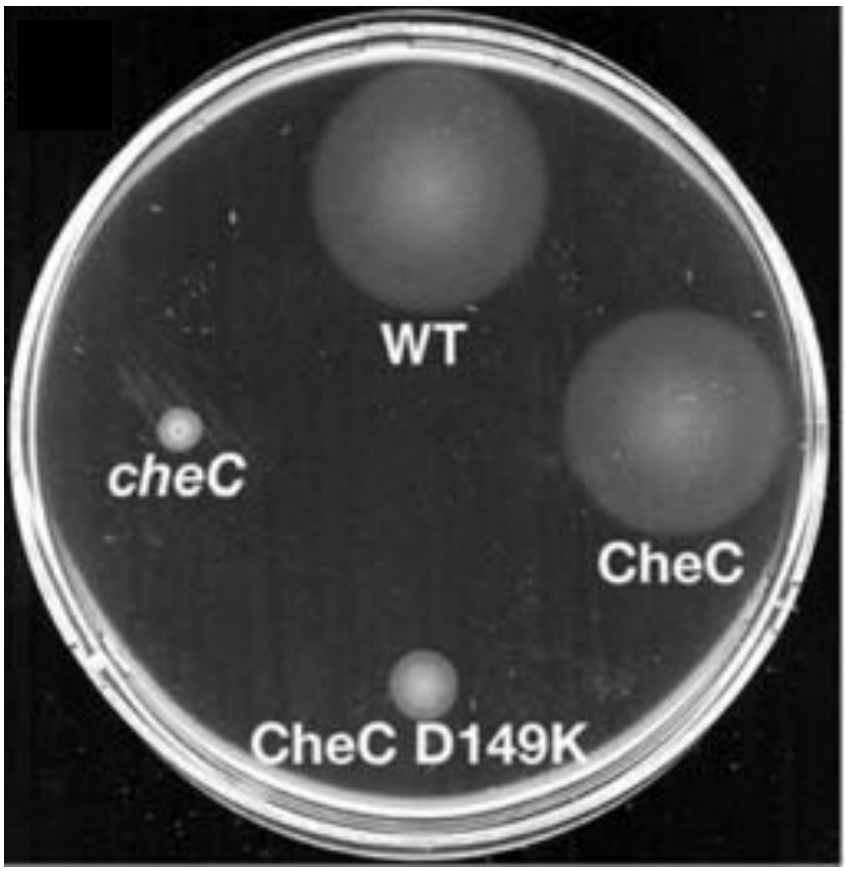


Figure 1.3.1 – *B. subtilis* strains with defective CheC-CheD heterodimers exhibit reduced motility. The swarm plate assay going from the top and around clockwise shows WT *B. subtilis*, a ΔCheC complement, a CheC D149K mutant and finally a ΔCheC deletion. The 6 o'clock D149K mutation in CheC is located at the CheD binding interface, preventing heterodimer formation between CheC and CheD. The swarms at 6 and 9 o'clock show that motility in the two strains with in-active CheC are reduced by ~70 % in comparison to the wild type, suggesting that a CheC: CheD heterodimer plays an essential role in regulating chemotaxis. Image taken from (Chao *et al.,* 2006).

1.3.2 - CheD function

CheD binds to both methyl accepting chemotaxis proteins (MCP), which are a family of transmembrane receptor proteins, and the downstream de-phosphorylase CheC; in both cases CheD catalyses the site-specific de-amidation of GLN residues. MCP receptors comprise an intracellular Histidine kinase domain and extracellular signal transducing domain, separated by a membrane-spanning domain made up of 4 α-helices (Falke and Hazelbauer, 2001). These MCP receptors and their downstream effectors mediate chemotaxis, a critical component for virulence in most pathogens (Foynes *et al,* 2000).

The feedback system mediating chemotaxis, is well characterised (figure 1.3.2) (Garrity and Ordal, 1997; Muff and Ordal, 2007). It is initiated by methylation of the extracellular MCP domain, in response to stimuli binding (Kehry *et al.,* 1985). This excitation is transmitted across the transmembrane portion of the MCP with the aid of CheD (Rosario *et al.,* 1995), triggering the cytoplasmic Histidine kinase CheA, which in turn phosphorylates a downstream effector CheY. The active phosphorylated CheY then interacts in a concentration dependent fashion with FliM, a flagella switch protein mediating flagella rotation (Toker and Macnab, 1997). This switch is crucial as the mode of motility observed in Prokaryotes is directly related to the direction their flagella are rotating. For example, in *B. subtilis* clockwise rotation corresponds to a tumbling motion, whereas counter-clockwise rotation is associated with swimming movements. The stimulation of *B. subtilis* MCP receptors causes the flagella to rotate in a counter-clockwise direction, leading to a swimming movement (Ordal *et al.,* 1993).

A critical stage in this cycle relates to the adaptation of the signalling cascade from an active state, into a stalled but sensitive condition ready to adapt to fresh stimuli. For this to occur the CheY-P signal requires negating and the cascade being returned to pre-excitation levels. In Prokaryotes there are two proteins responsible for the de-phosphorylation of active CheY-P, FliY a constitutively active low level de-phosphorylase and CheC a more efficient enzyme requiring activation (Szurmant *et al.,* 2003). CheC has previously been shown to form heterodimeric complexes with CheD (Rosario and Ordal, 1996) leading to the up regulation of CheC (Szurmant *et al.,* 2004). Further studies have shown that this dimerisation with CheD is essential to activate CheC's de-phosphorylase activity (Park *et al.,* 2004).

At the other end of the cycle, CheD has also been shown to activate the initial Histidine kinase CheA (Rosario *et al.,* 1995). Therefore, CheD has a dual role in both the production of active CheY-P and its eventual de-phosphorylation and deactivation to CheY. This suggests that it plays a regulatory role in bacterial chemotaxis, moderating the MCP receptors sensitivity to environmental stimuli (Saulmon *et al.,* 2004).

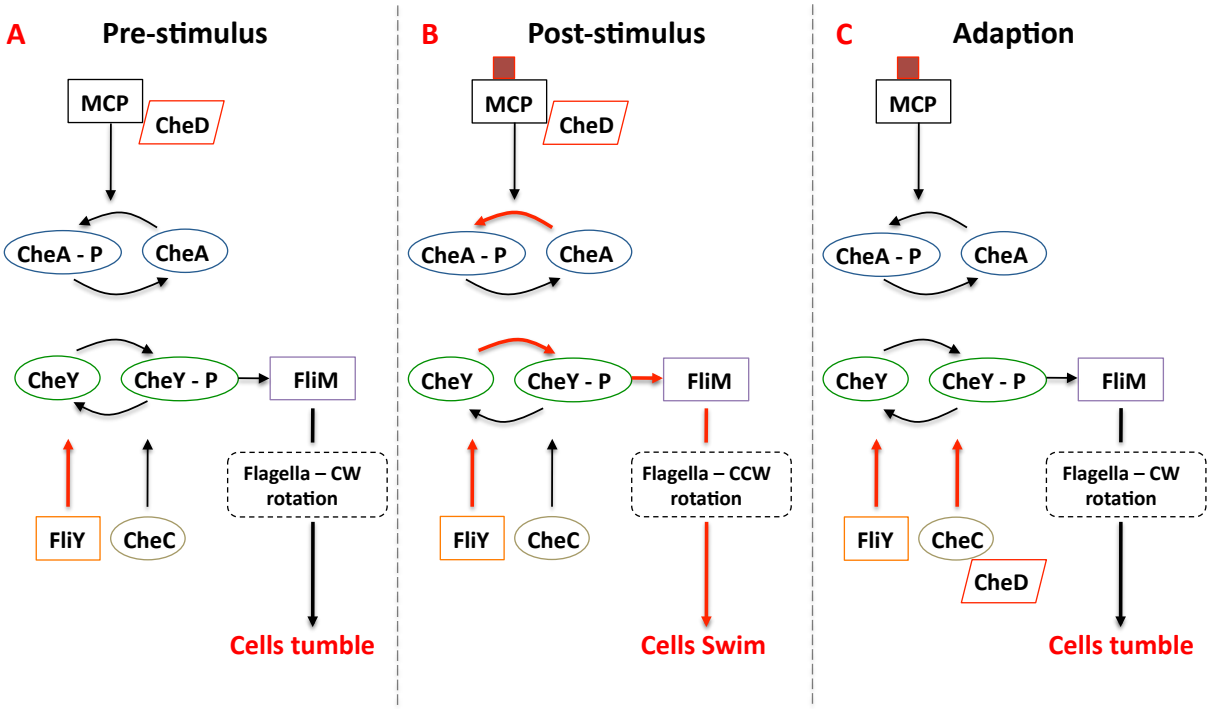


Figure 1.3.2 – Regulation of chemotaxis in *B. subtilis*. The mode of movement in bacteria is determined by the direction its flagella rotate clockwise denoting a tumbling pattern and counter-clockwise a swimming motion. Initially the system is in a balanced state with cells predominantly tumbling. The protein modulating flagella rotation is FliM, which is activated by increased levels of CheY in its active phosphorylated form. CheY is phosphorylated in response to MCP receptors interacting with external stimuli, activating an intracellular MCP domain a Histidine kinase called CheA. The balance between inactive CheY and active CheY-P is mediated by a de-phosphorylase FliY. Upon stimulant binding and MCP excitation (panel **B**) CheA phosphorylates CheY. CheY-P then interacts with FliM, which switches the flagella to rotate clockwise yielding a swimming motion. Panel **C** displays the adaptation required to reset the MCP so it can react to new stimuli. Elevated CheY-P levels lead to an increased affinity of CheC for CheD, which in turn leads CheD to dissociate from the MCP deactivating the receptor. Deactivation of both the MCP and CheA stalls the production of fresh CheY-P. CheD then forms a heterodimer with CheC, activating CheC and its de-phosphorylase activity, which along with FliY depletes the excess CheY-P back to pre stimulus levels. Diagram adapted from (Muff and Ordal, 2007).

1.3.3 - CheD structure

CheD is a 16 KDa protein, the first structure was solved using recombinant protein from *T. maritima* in complex with CheC, a 22 KDa protein with a previously determined structure (Park *et al.,* 2004). The CheD-CheC complex was solved using molecular replacement with a CheC exclusive ensemble (PDB ID: 1XKR) and refined to a resolution of 2.5 Å (Chao *et al.,* 2006).

CheD shares a similar tertiary fold with C-CNF1 (figure 1.3.3), composed of a 3 layered α, β, β sandwich, made up of 2 mixed β sheets of 5 β strands respectively, flanked on one side by 2 α helices. The active site cleft is located towards the top edge of the central β sandwich within a shallow cavity (figure 1.3.4). This active site contains a CYS 27 – HIS 44 catalytic dyad (figure 1.3.5), which is consistent with the dyad observed in C-CNF1. However, unlike C-CNF1 there is no supportive TYR residue below the essential CYS, despite it conserved location on a loop. There is also a difference observed in the pattern of hydrogen bonds made to the Imidazole ring of the catalytic HIS 44. In C-CNF1 the Nε of the catalytic HIS forms a hydrogen bond with a backbone carbonyl, whereas in CheD an equivalent hydrogen bond is formed with a side chain hydroxyl supplied by THR 21 (Chao *et al.,* 2006). The cleft in CheD is also far shallower than observed in C-CNF1, explaining how subtle conformational changes in CheC allow it to become a CheD binding partner.

This structure is especially relevant as it shows CheD in complex with one of its binding partners. CheD across both its binding partners interacts with a consistent consensus sequence, composed of an A/S – X(2) – Q/E – Q/E – X(2) – A/S motif, with the de-amidated GLN consistently followed by another GLN which is not modified. With both monomeric and heterodimeric structures of CheC available. It has been shown that the α helical CheD binding interface of CheC, is modified to favour interaction by mimicking the protrusion of small ALA or SER residues observed at the MCP-CheD binding site (Chao *et al.,* 2006). Therefore, CheD exhibits multiple substrate specificity through site-specific interaction with a weak consensus sequence, a trait shared with C-CNF1 and its specificity for multiple small GTPases. However, despite similarities at an active site and tertiary structure level C-CNF1 and CheD are sequentially extremely divergent sharing only 5 % sequence similarity, which only becomes apparent after structural alignment. The observed conservation in tertiary fold, partnered with very different binding partners and primary sequences, indicate that CheD and C-CNF1 have not recently split from a common ancestor.

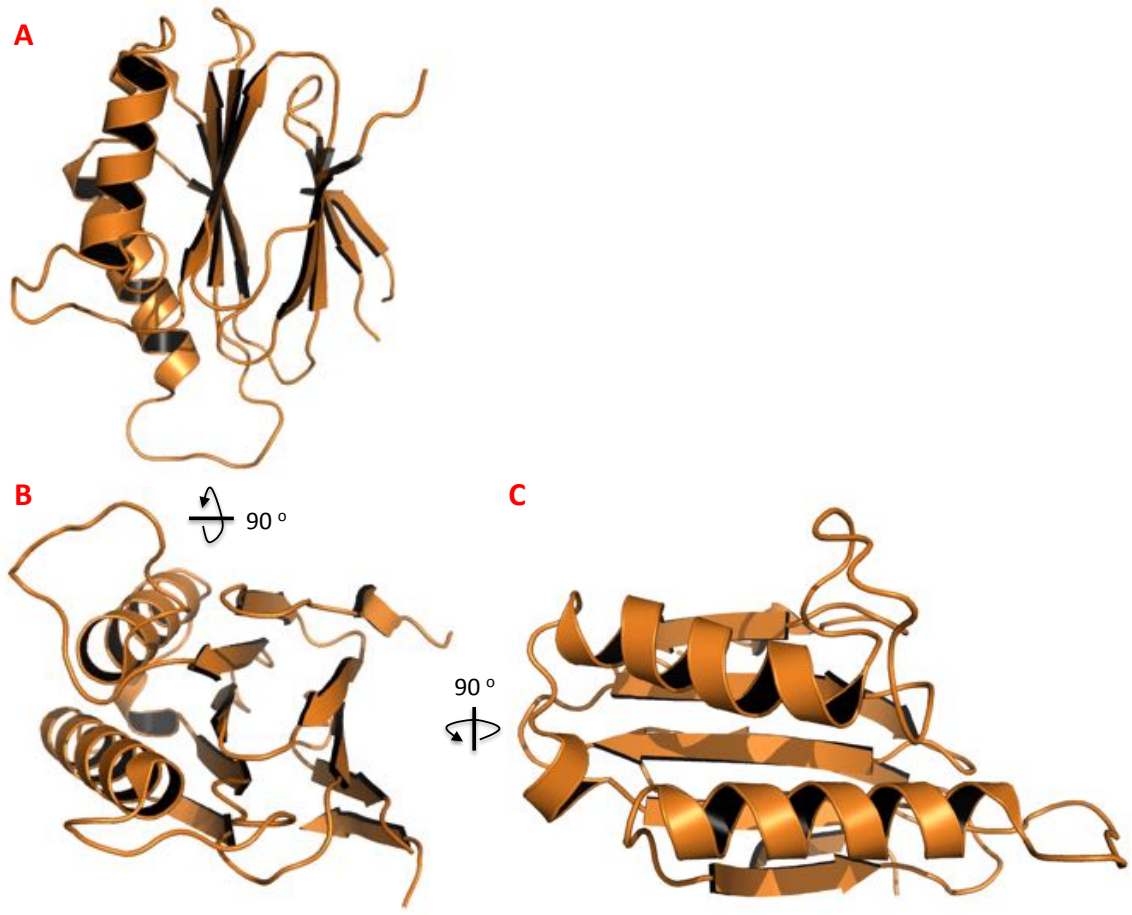


Figure 1.3.3 – 3D representation of CheD. CheD is a small 16 KDa protein, mediating MCP sensitivity in Prokaryotic cells and is located exclusively in the cytoplasm. Like C-CNF1 it is composed of a central α, β, β sandwich composed of 2 mixed β sheets. However, unlike C-CNF1 these sheets are much smaller, each composed of only 5 β strands. The arrangement of the flexible loops also differs from C-CNF1, with CheD exhibiting much shorter connective loops apart from those connecting the flanking α helices. Diagram constructed in Pymol using PDB ID: 2F9Z (Chao *et al.,* 2006).

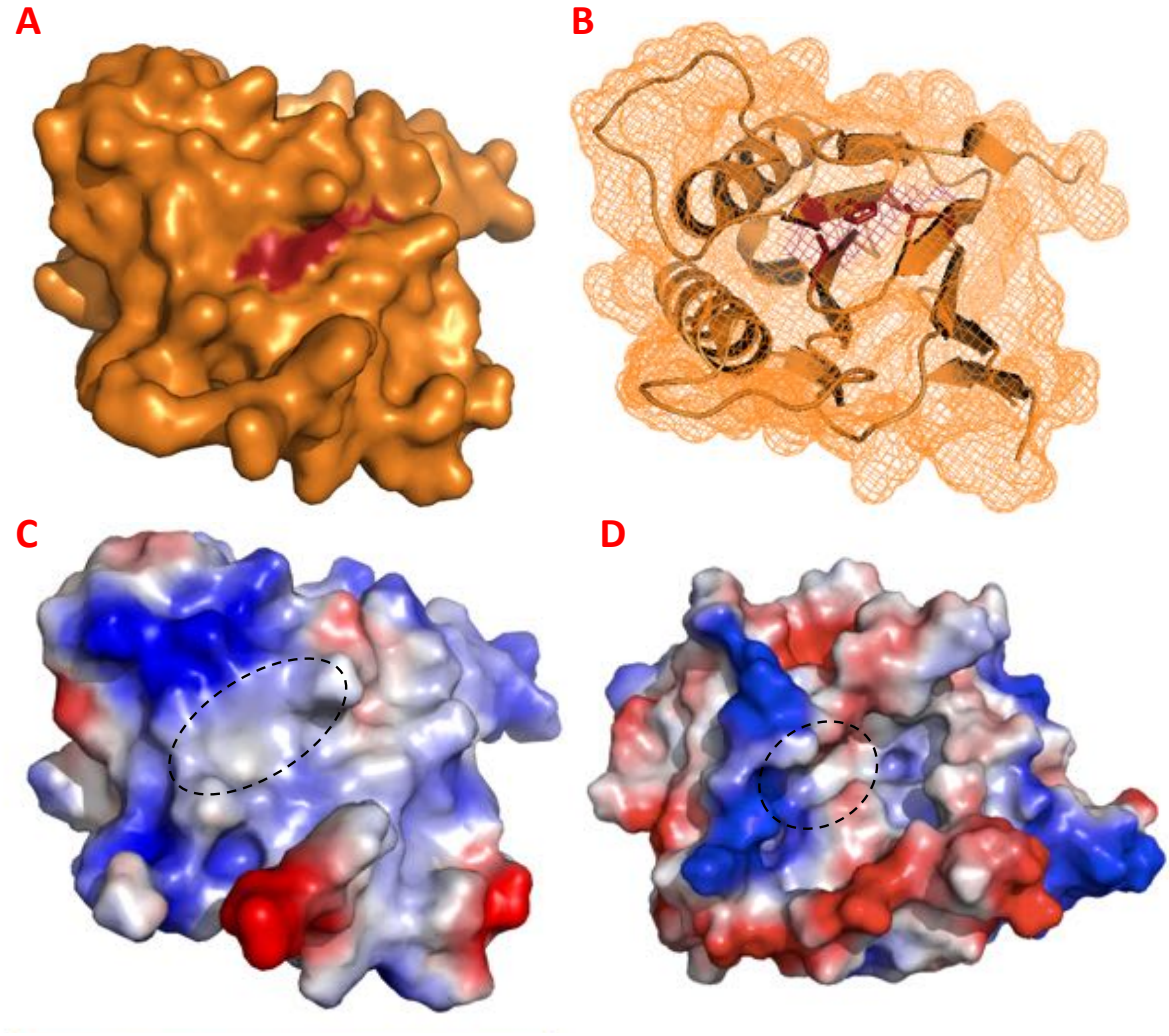


Figure 1.3.4 – Surface model of the CheD active site cleft in comparison with C-CNF1. CheD binds to both MCP receptors and CheC via the same active site cleft, through closely matched recognition sites. Panel **A** is a surface model of CheD with the catalytic dyad (CYS 27 and HIS 44) and stabilising THR 21 highlighted in red. Panel **B** is the same view but with the surface model substituted for a mesh representation, displaying the orientation of the catalytic residues relative to the active site cleft. Panels **C** and **D** are electrostatic charge diagrams of CheD and C-CNF1 respectively, their active sites are highlighted in black, with red portions of the surface corresponding to negatively charged and the blue positively charged regions. This diagram demonstrates that the active site cleft of CheD is located in a neutral valley surrounded on either side by positively charged regions, a trait shared with C-CNF1. Diagram constructed in Pymol using PDB ID: 2F9Z (Chao *et al.,* 2006).

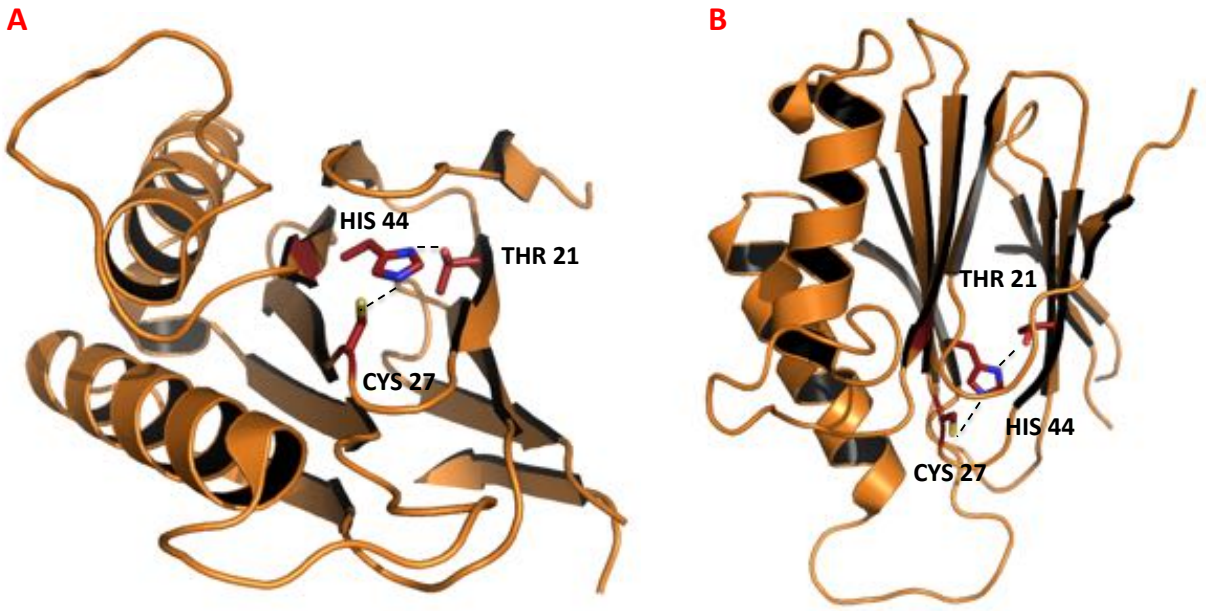


Figure 1.3.5 – CheD exhibits a similar catalytic dyad to C-CNF1. Panel **A** displays the characteristic Glutamine de-amidase active site complete with CYS - HIS dyad. In this example the dyad is located towards the top edge of the central β sandwich, which highlights the small size of CheD. Unlike C-CNF1 the catalytic dyad in CheD is located towards the N terminal with the conserved HIS supported through hydrogen bonding with a side chain hydroxyl supplied by THR 21. Panel **B** shows a top down view of the active site, which better demonstrates the direction and length of the hydrogen bonds involved. Diagram constructed in Pymol using PDB ID: 2F9Z (Chao *et al.,* 2006).

1.4.1 - Burkholderia lethal factor 1 – BLF1

BLF1 is the third and final phenotypically distinct Glutamine de-amidase enzyme to have been uncovered. It is a 23 KDa protein and the first toxin to be characterised from the pathogen *Burkholderia pseudomallei.*

1.4.2 - *B. pseudomallei* a growing threat

B. pseudomallei is an intracellular (Ray *et al.,* 2009), gram-negative bacterium and the causative agent of the disease Melioidosis (Wiersinga *et al.,* 2006). It inhabits a wide array of habitats ranging from moist soil through to surface water (Nandi *et al.,* 2010), but can also survive in distilled water for over a decade (Aldhous, 2005). It is Endemic in South East Asia and Northern Australia (Dance, 2000), where natural climates are humid and moist. In Northern Thailand 80 % of all 4 year old children test serologically positive for having been exposed to *B. pseudomallei* (Kanaphun *et al.,* 1993). Moreover this high level of exposure within a human population is further confirmed when examining the American veterans of the Vietnam War, with all 225, 000 survivors testing positive for the organism (Howe *et al.,* 1971). As a result of its startling persistence and varied natural habitats *B. pseudomallei* has been classified as a category B bio-warfare agent (Rotz *et al.,* 2002).

The genome of *B. pseudomallei* is composed of 2 chromosomes with distinctive evolutionary backgrounds. The large chromosome is 4.07 MBp and encodes for genes involved in essential functions, with the smaller 3.17 MBp chromosome encoding a variety of accessory proteins including virulence factors (Holden *et al.,* 2004).

B. pseudomallei is transmitted by cutaneous contamination, predominantly through either cuts to the feet or via inhalation and ingestion of infected aerosols or particles. However, the organism appears to favour infection via inhalation, as the rate of infection increases during the wet season alongside elevated aerosol production (Kanaphun *et al.,* 1993). It also can infect tomato plants, suggesting that the virulence factors responsible for host cell invasion and pathogenicity are broad range (Lee *et al.,* 2010). This leaves very few Eukaryotic hosts where *B. pseudomallei* is unable to survive and propagate, making it an extremely versatile pathogen.

1.4.3 - *B. pseudomallei* is the causative agent of Melioidosis

B. pseudomallei is the causative agent of the disease Melioidosis. The name Melioidosis is derived from the Greek 'melis' (distemper of asses) and 'eidos' (resemblance), the disease is often termed the 'great mimicker' (Wiersinga *et al.,* 2006). This is because the disease exhibits a wide variety of symptoms, which make diagnosis and treatment extremely challenging. Infections are commonly misdiagnosed as Tuberculosis (Vidyalakshmi *et al.,* 2008), Typhoid fever, Malaria (Currie *et al.,* 2008) or even Cancer (Reechaipichitkul, 2004). The variety of symptoms required for *B. pseudomallei* to resemble such a wide range of diseases is only possible because of the broad suite of morphologies the organism is capable of adopting. *B. pseudomallei* cells have been observed adopting 1 of 7 morphologies, each corresponding to a different pattern of virulence (Stone *et al.,* 2007).

Melioidosis, even when correctly diagnosed, is difficult to cure requiring upwards of 5 months treatment with a cocktail of antibiotics (Wuthiekanun and Peacock, 2006). This length of treatment is necessary because of drug resistance and the high levels of anti-microbial compounds secreted by the pathogen (Mima and Schweizer, 2010). Patients typically display symptoms within 1-2 weeks of infection with 20-50 % of infected patients not surviving the prescribed treatment (Wuthiekanun and Peacock, 2006). Furthermore, *B. pseudomallei* can also enter into a dormant state for decades, later presenting with symptoms in an opportunistic manner. A pertinent example is an US citizen who served in Burma and Thailand during WWII. Before returning to the US for the remainder of his life, who 62 years after initial contact with the organism contracted Melioidosis (Ngauy *et al.,* 2005).

Gaining intra-cellular access to the host is a crucial component of *B. pseudomallei* virulence. The pathogen gains entry to the host through extensive modification of its actin cytoskeleton, followed by arresting its cell cycle (Cui *et al.,* 2010). Once present in the host *B. pseudomallei* secretes a selection of virulence factors, such as Superoxide dismutase C an enzyme that protects the pathogen from host super-oxides (Lefebre *et al.,* 2001). However, the vast majority of proteins involved in pathogenesis and propagation of *B. pseudomallei* remain uncharacterised.

1.4.4 - BLF1 the first characterised lethal toxin from *B. pseudomallei*.

BLF1 was one of 14 hypothetical proteins of unknown function, which had been identified through proteome comparison between pathogenic *B. pseudomallei* and a non-pathogenic relative *B. thailandensis,* as possible virulence factors (Wongtrakoonate *et al.,* 2007).

The toxin was identified through a structural genomics project and comparison with structures of known function, using the Dali-lite server, which identified it as a structural relative of C-CNF1 (Cruz-Migoni *et al.,* 2011). Like C-CNF1, BLF1 displays site-specific Glutamine de-amidase activity. However, it does not share with C-CNF1 an affinity for GTPases. Instead BLF1 targets eukaryotic initiation factor 4a (eIF4a), which plays a crucial role in translation (figure 1.4.1) (Hautbergue *et al.,* 2012). eIF4a is a RNA helicase that melts the secondary structures present in mRNA, prior to translation at the ribosome. Knocking out eIF4a results in extensive inhibition of protein synthesis in human cells (Pause *et al.,* 1994). The inhibition of protein synthesis is further compounded by the stalled mRNA blocking access to the ribosome, which prevents it from interacting with non-initiation dependent mRNA transcripts (Hautbergue *et al.,* 2012).

1.4.5 - BLF1 structure

BLF1 is a 23 KDa single domain protein, the structure was solved via X-ray diffraction utilising a Se-MET derivative and MAD phase solution using 1.04 Å resolution diffraction data (Cruz-Migoni *et al.,* 2011). Its tertiary structure comprises the now characteristic β sandwich (figure 1.4.2). Like C-CNF1 before, the central β sandwich in BLF1 is made up of 2 mixed β sheets flanked on either side by α helices, with these central β strands following the same route as the C-CNF1 equivalent. This similarity is clearer when comparing α carbon alignment along the peptide backbone, with 170 out of 211 residues in BLF1 aligning to within 3.9 Å RMSD of a corresponding C-CNF1 position. However, despite matching closely in the central β sandwich region these two toxins diverge significantly in both the α helical and loop regions, with BLF1 displaying far shorter loops and α helices than observed in C-CNF1.

The active site cleft in BLF1 is also different from that of C-CNF1, exhibiting a clearly different electrostatic layout (figure 1.4.3). Unlike C-CNF1, the active site opening in BLF1 is located in a shallow negatively charged crater, not a neutral valley, with the catalytic residues deeply buried. Examination of the BLF1 active site reveals a CYS – HIS dyad in common with all the previously characterised Glutamine de-amidases (figure 1.4.4). CYS 94 is the essential nucleophile position, with HIS 106 co-ordinating its orientation and TYR 164 supporting it from below, with THR 88 orienting the Nε of HIS 106 in a position conserved with CheD (Cruz-Migoni *et al.,* 2011).

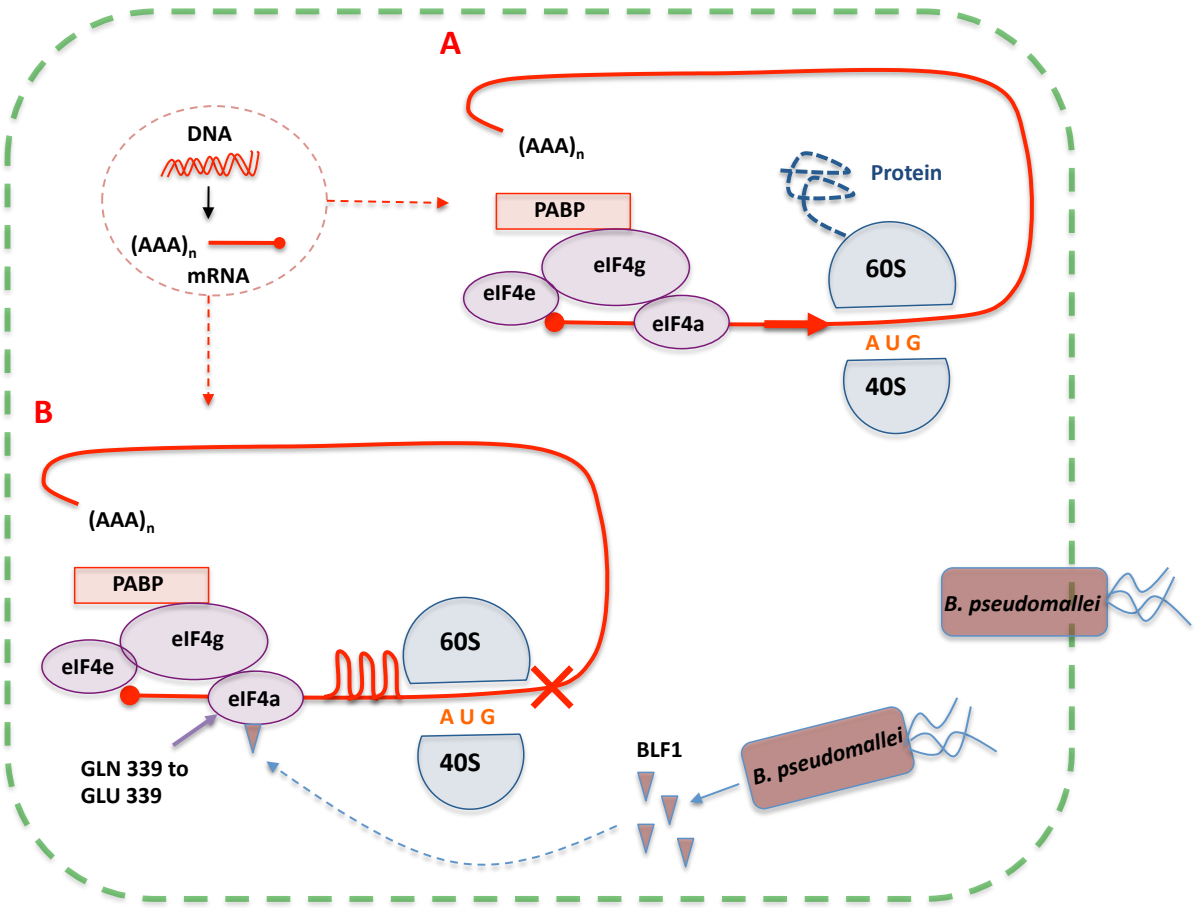


Figure 1.4.1 – De-amidation of eIF4a GLN 339 results in stalled protein synthesis in host cells. *B. pseudomallei* is an intracellular pathogen, once internalised it produces a selection of secreted virulence factors including the toxin BLF1, represented with red triangles. BLF1 influences initiation dependent translation by blocking post-transcriptional modification of mRNA, preventing it from accessing the ribosome. In BLF1 infected cells transcription occurs as usual, with viable folded mRNA exported from the nucleus. The next stage is translation of the mRNA at the ribosome. In order for the ribosome to process the mRNA it is first circularised, through interactions with the Poly-A binding protein (PABP) (red) and the Eukaryotic initiation factor complex at the mRNA cap (purple). BLF1 de-amidates Eukaryotic initiation factor 4a (eIF4a) at GLN 339, a position mediating the initiation factors interaction with mRNA. This is crucial because eIF4a is a helicase enzyme, responsible for unfolding secondary structures in the mRNA allowing it to enter the ribosome, panel **A**. De-amidation of GLN 339 abolishes this helicase activity, which stalls translation by irreversibly binding any unprocessed mRNA at the ribosome. Cell death occurs when the majority of ribosomes have been de-activated, as shown in panel **B**. Diagram adapted from (Hautbergue *et al.,* 2012).

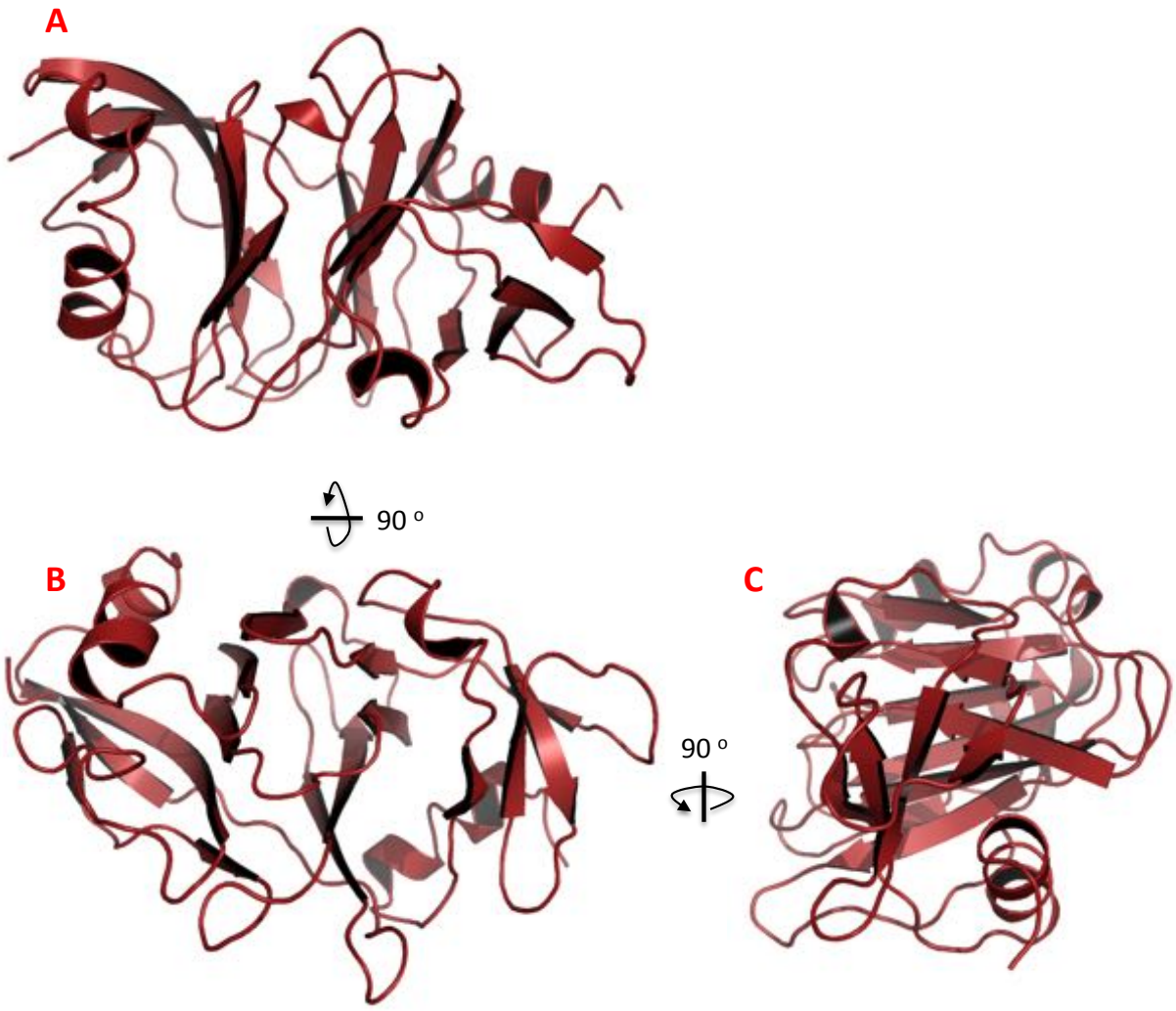


Figure 1.4.2 – 3D representation of BLF1. BLF1 is a 23 KDa toxin that de-amidates GLN 339 in the mRNA binding region of eIF4a, a helicase enzyme and component of the Eukaryotic translation initiation factor complex. eIF4a mediates the melting of secondary structures present in mRNA allowing it to access the ribosome. BLF1, like both CheD and C-CNF1, folds into a α, β, β sandwich with the central β stranded region composed of opposing mixed β sheets made up of 5 and 4 β strands (left to right). However, unlike either C-CNF1 or CheD there also is a second β-stranded region located on the right flank, best displayed in panels **A** & **B**. Diagram constructed in Pymol using PDB ID: 3TU8 (Cruz-Migoni *et al.,* 2011).

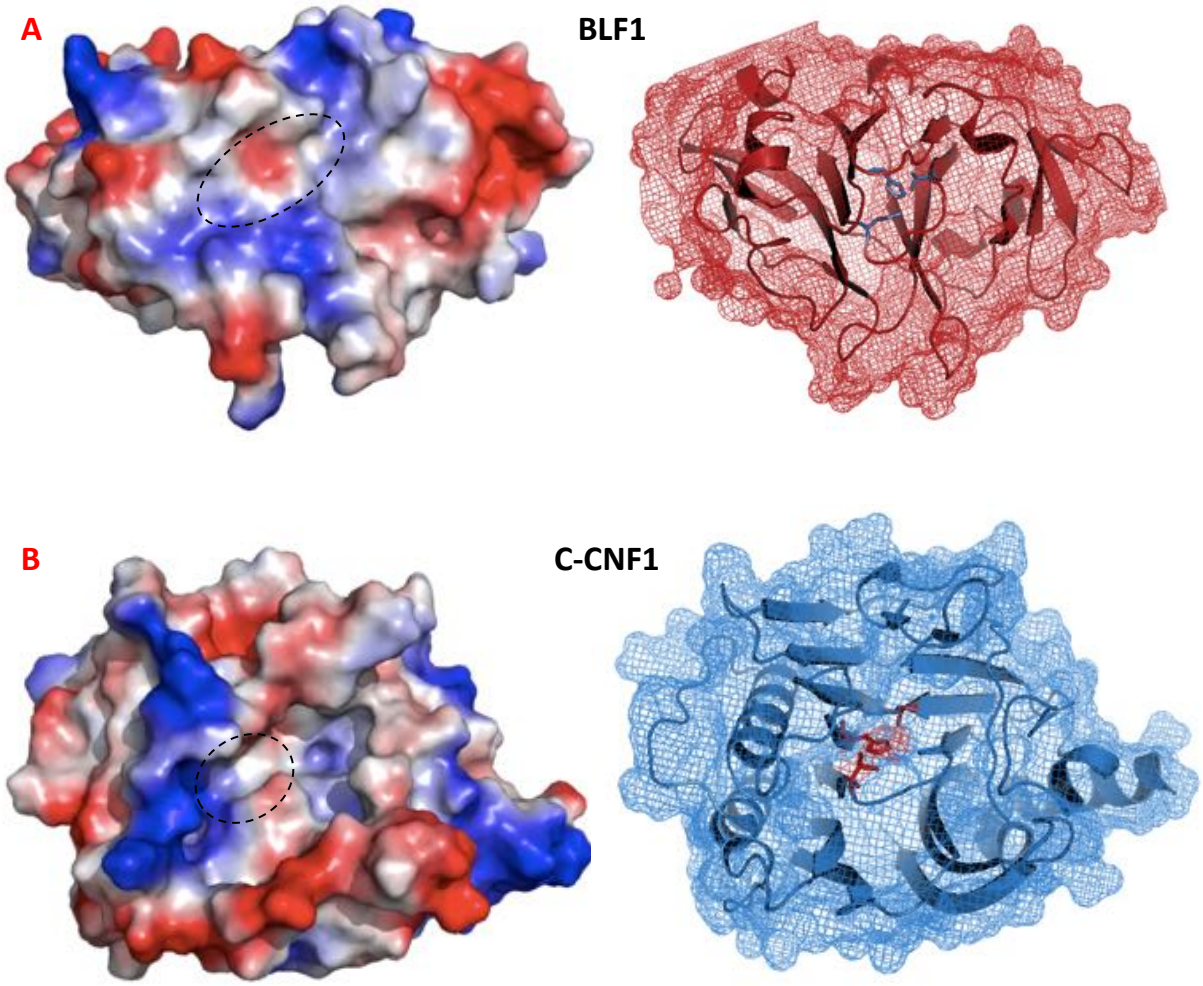


Figure 1.4.3 – Surface models comparing the active site cleft of BLF1 with C-CNF1. Panel **A** is an electrostatic charge diagram of BLF1, with the active site cleft denoted by a dashed black circle. Alongside this is the corresponding surface mesh representation of BLF1 in red with THR 88, CYS 94 and HIS 107 highlighted in blue. Panel **B** displays C-CNF1 in the same fashion, but with the surface mesh in blue and the active site residues VAL 833, CYS 866 and HIS 881 highlighted in red. The clearest deviation from C-CNF1 is the shallow active site cleft, with the catalytic CYS – HIS dyad in BLF1 buried deeper from the surface. There is also a different distribution of charged residues surrounding the active site; with C-CNF1 exhibiting a neutral valley surrounded by a positively charged ridge, whereas BLF1 is surrounded on either side by positively charged regions with the active site located in a shallow negatively charged crater. Diagram produced in Pymol using PDB ID: 1HQ0 and 3TU8 (Buetow *et al.,* 2001; Cruz-Migoni *et al.,* 2011).

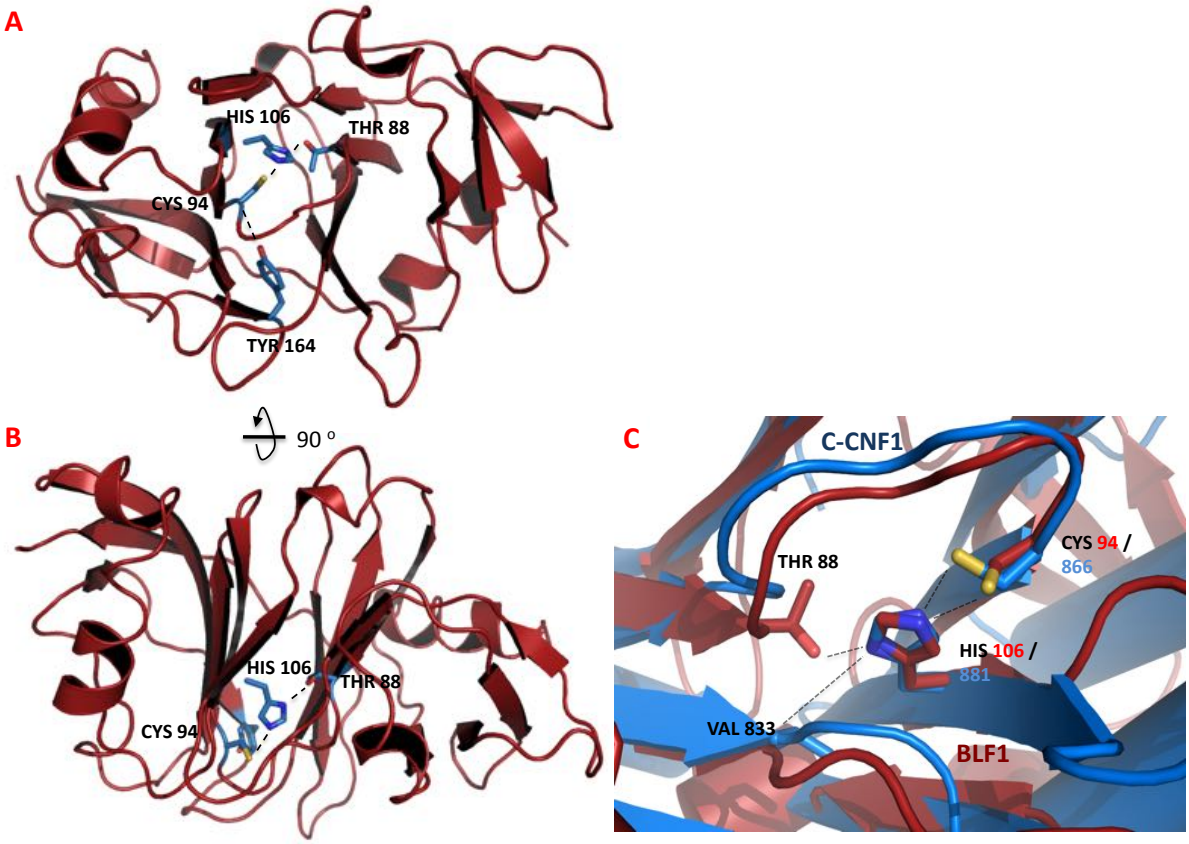


Figure 1.4.4 – BLF1 exhibits a characteristic active site dyad, which aligns strongly with C-CNF1. Panel **A** displays the active site of BLF1 from the solvent accessible face, whilst panel **B** shows the same region but from the top face. Panel **C** zooms into the active site, comparing the alignment of catalytic dyads between BLF1 (red) and C-CNF1 (blue). BLF1 exhibits a characteristic Glutamine de-amidase catalytic dyad, located centrally in the toxins primary sequence. Like C-CNF1 the essential CYS is located on a loop and supported from underneath by a TYR residue. BLF1 deviates from C-CNF1 at the HIS co-ordinating component of the active site, with THR 88 hydrogen bonding with the Nε of HIS 106. However, when BLF1 and C-CNF1 are aligned along their central β sandwich (panel **C**), the active site dyad and its relative location in the overall fold are clearly conserved. Diagram produced in Pymol using PDB ID: 1HQ0 and 3TU8 (Buetow *et al.,* 2001; Cruz-Migoni *et al.,* 2011).

1.4.6 - BLF1 plays an important role in *B. pseudomallei* virulence and future treatments

BLF1 was the first toxin isolated from *B. pseudomallei* capable of inciting cell death in vivo. When administered via intra-peritoneal injection to BalB/C mice, BLF1 is lethal within 14 days (Cruz-Migoni *et al.,* 2011). BLF1 is an extremely effective agent against immune response cells, with only 3 days incubation alongside J774 Macrophage cells required for LD_{50} cell death (Cruz-Migoni *et al.,* 2011). It has also been calculated, in vivo, that a single molecule of BLF1 can de-amidate approximately 700 eIF4a molecules per minute, stalling the activity of the same number of ribosomes. Different cell types exhibit varying numbers of ribosomes, but at a rate of 700 de-activated ribosomes per minute even the most densely active cells will reach total ribosome depletion within a mater of hours (Hautbergue *et al.,* 2012).

BLF1 has also been shown to play a role in pathogenesis as ΔBLF1 knockout mutants of *B. pseudomallei* are 100 fold less virulent than the WT, with a median lethal dose of 1.26×10^5 colony forming units. This depreciation of virulence in the ΔBLF1 strain indicates that BLF1 is essential for *B. pseudomallei* virulence. Furthermore mutating the catalytic residue CYS 94, leads to significantly reduced toxicity in J774 macrophages and no measured toxicity in BalB/C mice, confirming its importance. However, BLF1 when incubated with a 3T3 Eukaryotic cell line does not display toxicity unless the BLF1 is first internalised using Bio-porters. This suggests that BLF1 does not have the ability to independently gain entry to the host, without being exported from intracellular *B. pseudomallei* (Cruz-Migoni *et al.,* 2012). Currently BLF1 secretion is not well understood.

Despite not understanding how the pathogen delivers BLF1, its characterisation has the potential to influence the development of several novel therapies. The first is through the design of small molecule rational inhibitors, analogous to eIF4a, which could become a preventative drug for high-risk candidates. The second is the modification of BLF1 to produce an effective vaccine; the C94S mutant for example is not lethal and could be made safe for use. The final avenue relates to cancer treatment, as BLF1 could be targeted specifically to cancer cells where it would act against protein synthesis (Malina *et al.,* 2011).

1.5.1 - Elucidating a mechanism for Glutamine de-amidation through comparison with Papain.

All the currently characterised glutamine de-amidase enzymes exhibit a central catalytic CYS – HIS dyad. This core component of their catalytic machinery has drawn comparison with another family of enzymes the cysteine proteases, most notably Papain. Papain is a protease,

an enzyme that cleaves peptide bonds. This cleavage occurs through a nucleophilic attack on the protein backbones scissile C-N bond via an essential CYS residue (Brocklehurst *et al.,* 1988). Therefore, Papain shares a similar target bond and likely chemical mechanism with the Glutamine de-amidases. The Papain superfamily has been shown to play a key role in both the immune response in lyzosomes (Amamoto *et al.,* 1984), additionally also in muscle degradation (Gopalan *et al.,* 1986). De-regulation of this family of proteases has also been linked to the development of cancer (Lah *et al.,* 1989) and muscular dystrophy (Gopalan *et al.,* 1986). Papain was one of the earliest proteins to be studied and as a result is very well characterised, making it an ideal protein to model glutamine de-amidase activity upon.

1.5.2 - Papain's structure

The first papain structure was solved using non-recombinant protein isolated from the latex layer of papaya fruit, to 2.8 Å resolution (Drenth *et al.,* 1968). However, examination of the structure at an even higher resolution of 1.65 Å (figure 1.5.1) reveals that it is composed of two distinct domains separated by a cleft (Kamphuis *et al.,* 1984). The structure of papain does not closely resemble any Glutamine de-amidase, with the only structural similarities arising at the active site where both families share a catalytic CYS – HIS dyad.

Papain contains two catalytic residues CYS 25 and HIS 159, which are located facing one another from across the domain cleft. The sequential distance between this dyad is large, with CYS 25 present on a α helix at the N-terminal domain and HIS 159 located much further downstream on a β strand in the C-terminal domain. Both residues are surface accessible and like the Glutamine de-amidases are thought to form a thiol – imadizolate ion pair through a hydrogen bond. Figure 1.5.2, displays a schematic of the catalytic dyad in Papain and its accessory residues GLN 19 and ASN 175 interacting with a peptide substrate. The side chain amide nitrogen of GLN 19 forms a hydrogen bond with the carbonyl moiety of the peptide. This peptide carbonyl also interacts with the main chain amide of CYS 25 to produce an oxyanion hole. This interaction is involved in stabilising reaction intermediates (Garavito *et al.,* 1977), with Q19A and Q19S mutants displaying a ~600 fold reduction in activity (Ménard *et al.,* 1991). ASN 175 forms a hydrogen bond with HIS 159 between its Cγ - carbonyl and the HIS residues Nε. This interaction permits the rotation of the Imidazole ring along the C_β - C_γ bond, allowing a favourable orientation for stabilisation of the thiol – imadizolate ion pair (Harrison *et al.,* 1997). An important step in peptide cleavage is the exchange of protons between the thiol – imadizolate ion pair, with this exchange believed to be highly dependent on the conformation of HIS 159 in relation to CYS 25 (Rullmann *et al.,* 1989).

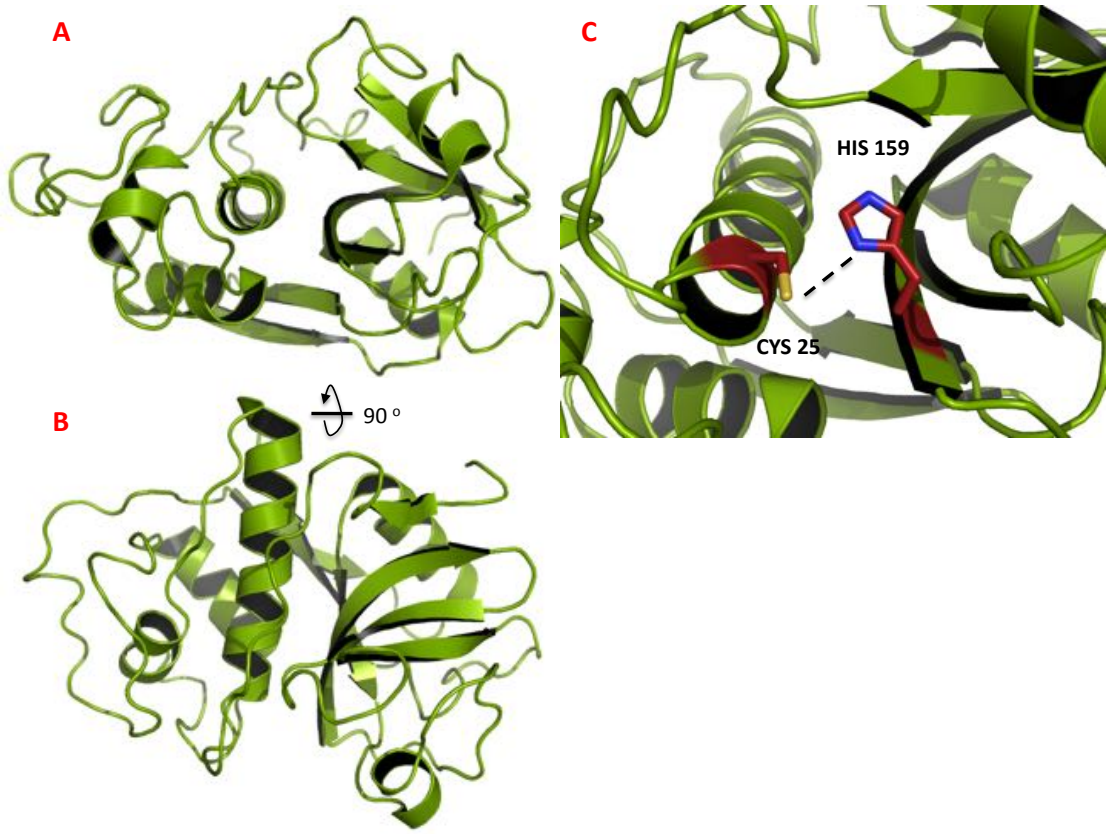


Figure 1.5.1 – 3D representation of Papain and its active site dyad. Panels **A** and **B** show the overall tertiary structure of Papain from two perspectives, with panel **C** zooming into the active site where the catalytic dyad is highlighted in red. Papain is composed of a single polypeptide chain divided into two clear domains, separated by a central cleft. The N-terminal domain to the right hand side of panels **A** and **B** and is composed predominantly of α helices. Whereas, the C-terminal domain on the left hand side is more heavily populated by β strands. The catalytic dyad of Papain comprises a CYS 25 – HIS 159 dyad, which is separated by the domain cleft. CYS 25 is located on an α helix and HIS 159 on a β sheet, both of which are solvent accessible. A hydrogen bond interaction forms between the terminal thiol of CYS 25 and the Nδ position in HIS 159. Overall Papain (with the exception of the active site dyad) shares little structural or sequential similarity to the Glutamine de-amidases. Diagram produced in Pymol using PDB ID: 9PAP (Kamphuis *et al.,* 1984).

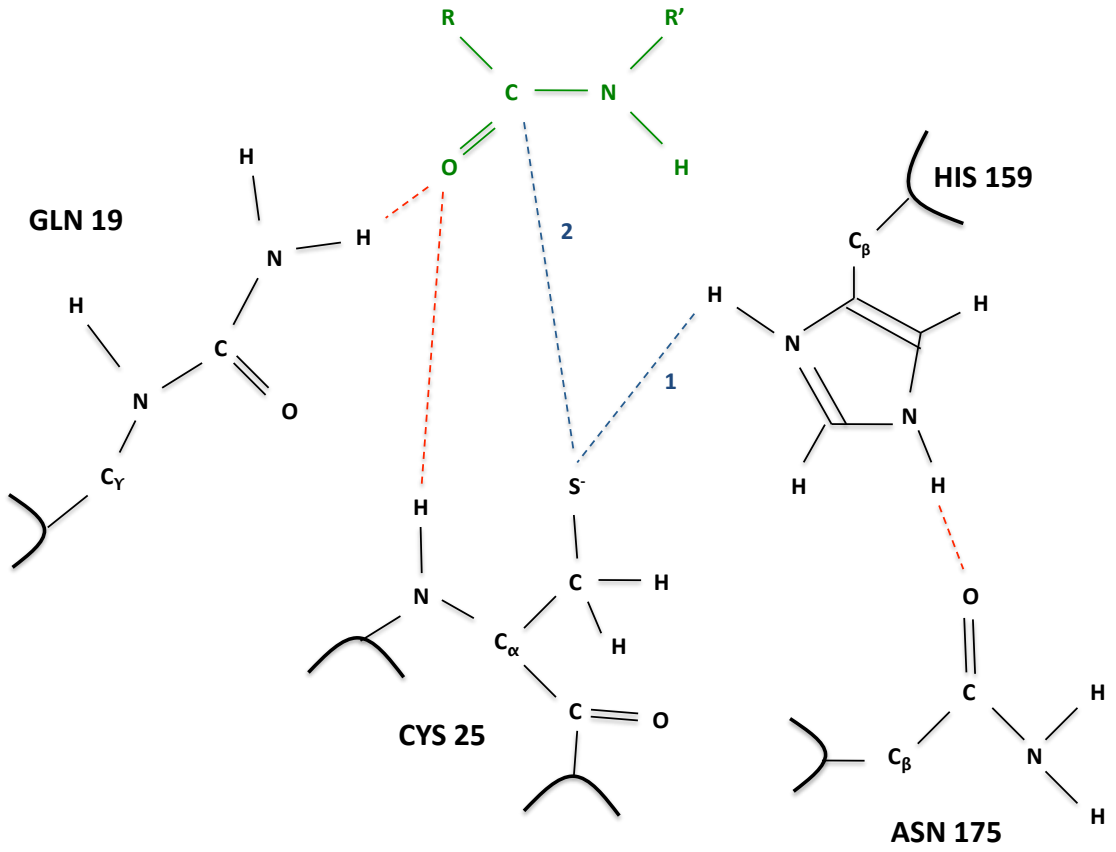


Figure 1.5.2 – Schematic representation of the Papain active site interacting with a peptide. At the top of the diagram is a single peptide unit (green) interacting with Papain. The catalytic bonds are highlighted in blue with supportive interactions coloured red. Papain has 4 active site residues responsible for first binding to peptides and then cleaving the scissile C-N peptide bond. CYS 25 is essential supplying the nucleophilic sulphur atom, which when oxidised attacks the scissile bonds carbon atom (2). But prior to this reaction HIS 159 orients and de-protonates CYS 25, through a hydrogen bond (1) between the Nδ of HIS 159 and the terminal thiol of CYS 25, producing an active Sulphur ion. Meanwhile, ASN 175 as an accessory residue that supports HIS 159 through a hydrogen bond between its side chain carbonyl and the catalytic HIS residues spare Nε. This interaction permits rotation around the Cβ – Cγ bond in HIS 159, facilitating eventual disassociation from CYS 25. On the opposite side of the active site GLN 19 together with CYS 25 form an oxyanion hole, with both residues hydrogen bonding to the peptides carbonyl. This oxyanion hole is thought to be important in stabilising the tetrahedral transition state and acyl intermediates of the reaction. Diagram adapted from a figure published in (Harrison *et al.,* 1997).

In the Glutamine de-amidase enzymes, residues VAL 833 and THR 88 in C-CNF1 and BLF1 respectively are thought to fill analogous roles to ASN 175 in Papain. However, it is presently unclear which positions in the Glutamine de-amidase enzymes would pair with the catalytic CYS to form an oxyanion hole as GLN 19 does in Papain.

1.5.3 - The catalytic mechanism of Papain

The protease reaction catalysed by Papain is driven by the de-protonation of CYS 25 by HIS 159 oxidises the terminal thiol to create a powerful nucleophile. This de-protonation is driven by differences in the pKa observed between the two domains. HIS 159 is located on the C-terminal domain and held at an ionising pKa of 8.5, whilst CYS 25 is located on the N-terminal domain and held at an opposing pKa of 4 (Daggett *et al.,* 1991). This pKa discrepancy between the two domains encourages the oxidation of CYS 25, forming the reactive Sulphur ion. The nucleophilic sulphur ion then attacks the peptide backbones carbonyl carbon, to cleave the scissile peptide bond. This attack is theorised to involve HIS 159 either ionising the substrate peptide prior to, or in tandem with, the nucleophilic attack of CYS 25 to form a stable tetrahedral transition state (Arad *et al.,* 1990). The proton provided by HIS 159 leads to the release of the amino terminal of the polypeptide, with the remaining carbonyl group bonded to the CYS residue as an acyl – enzyme intermediate. This acyl group is then released through a hydrolysis reaction, which reduces the terminal sulphur returning it to a thiol moiety ready for the cycle to repeat. The catalytic CYS – HIS dyad observed in the Glutamine de-amidase enzymes is currently hypothesised to be analogous to those seen in Papain. However, there is currently no evidence for this, besides the dyads structural arrangement. Having examined the catalytic mechanism of Papain a hypothetical mechanism for the Glutamine de-amidases is described in figure 1.5.3.

1.6 - Aims and objectives

Presently there are only three characterised Glutamine de-amidase enzymes. As they all share a catalytic dyad and broad tertiary fold in common, it is highly probable that more exist. This thesis, starting with the following Bio-informatics chapter, is going to examine the methods and challenges involved in isolating and characterising novel members from this super-family.

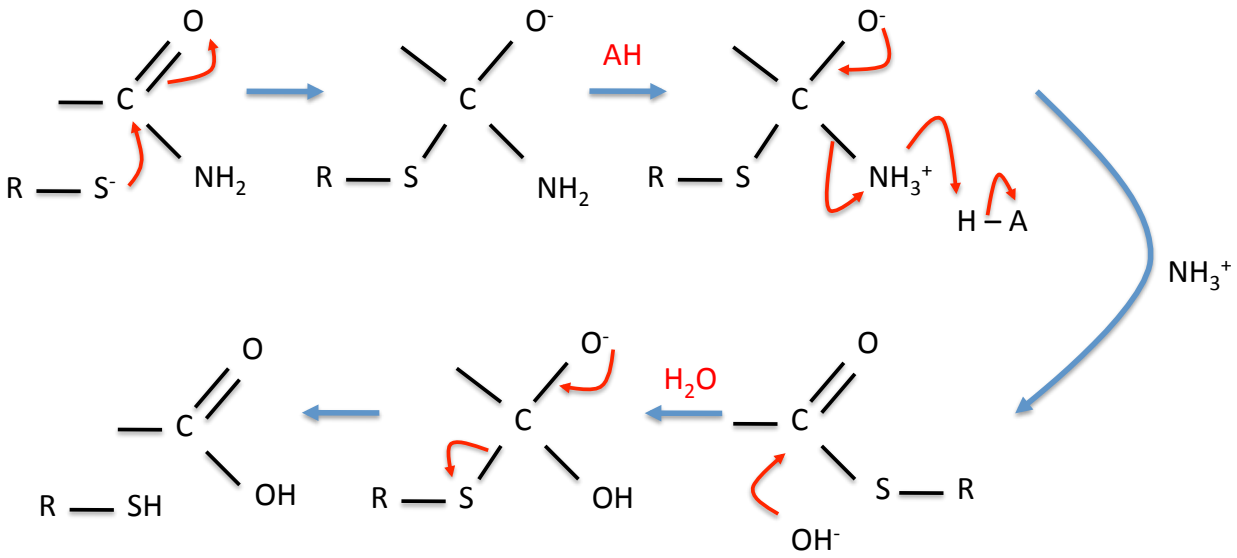


Figure 1.5.3 – Hypothetical Glutamine de-amidase mechanism. The arrows in blue denote individual stages with the arrows in red indicating the movement of electrons. The above mechanism is modeled on the Papain family of Cysteine proteases, with which the Glutamine de-amidase enzymes share a catalytic dyad. Papain cleaves peptide bonds by attacking C-N peptide bonds through a nucleophilic Sulphur ion, later forming a tetrahedral transition state and finally an acyl - enzyme intermediate. This hypothetical mechanism is therefore heavily dependant upon Glutamine de-amidases retaining Papain's tetrahedral reaction mechanism, but instead targeting a side chain C-N bond. Glutamine de-amidases are expected to initiate their nucleophilic attack through an oxidised CYS residue, interacting with the target GLN side chains δ-carboamide leading to a tetrahedral intermediate. This nucleophilic attack is coupled with the introduction of a proton (AH), supplied by an essential HIS residue, allowing the formation of a carboxyl moiety at the previously ionised oxygen resulting in an Ammonium leaving group. This leaves an acyl – enzyme complex that is broken through hydrolysis, reducing CYS 25's terminal Sulphur and releasing the substrate. The end result is that the substrate Glutamine side chain has been modified, thus mutating this position into a Glutamic acid.

The specific aims of this study were to identify and characterise through X-ray crystallography and a variety of biophysical techniques a selection of novel Glutamine de-amidase enzymes. However, up till this point none of the currently characterised Glutamine de-amidase toxins have been examined binding with either their substrates or suitable analogues. Therefore, a secondary aim of this study is to examine in more detail the active sites of these enzymes and how they may catalyse de-amidation, through co-crystallisation with either their natural substrates or analogous oligo-peptides.

Chapter 2: Bioinformatics and glutamine de-amidase target selection

The following chapter is going to introduce bioinformatics as a concept, before moving onto the programmes used to identify a selection of glutamine de-amidase targets to be examined as part of a small scale structural genomics project (figure 2.1.1).

2.1 – Introduction to Bioinformatics

Bioinformatics is the science of collecting and analysing complex biological data such as genetic codes, protein sequences and folding properties. It involves the production of computer programmes that aim to provide meaningful biological information from easily obtainable sequence data. Bioinformatics has been made possible by the parallel explosion of both computing power and low cost sequencing. There are now full genome sequences available for a vast selection of model microorganisms, with niche organisms added frequently. This wave of DNA sequence information has, largely been made possible by the seminal work of Fred Sanger (Sanger, 1977).

The Sanger approach to sequencing can broadly be divided into two classes, which are the shotgun and targeted sequencing methodologies (Sanger, 1988). Both techniques rely heavily upon the production of amplified template material, through either high copy number plasmids or the Polymerase chain reaction (PCR). The template region of the genome is then identified through thermal 'cycle sequencing' similar in principle to PCR, but with one of four differing fluorescent Deoxyribonucleotide triphosphates (dNTP) incorporated at each position. The amplified fluorescent template is then separated by gel electrophoresis and passed through a capillary. These fluorescent bases mimic standard dNTPs but upon excitation, using lasers, emit a distinct wavelength of visible light that can be recorded (Swerdlow and Gesteland., 1990). This trace of fluorescence allows the sequence of the template DNA to be read, by linking the peaks observed with the corresponding nucleotide.

However, compared to DNA sequences, polypeptide chains are far more challenging to identify experimentally. At present the easiest method is to digest the protein into smaller peptides (Steen and Mann., 2004), then identify the fragments via electro-spray Mass spectroscopy (Fenn *et al.,* 1989). Therefore, a far easier method for determining protein sequences is to translate the gene which encodes them using the standard 3 base codons (Matthaei and Nirenberg., 1962).

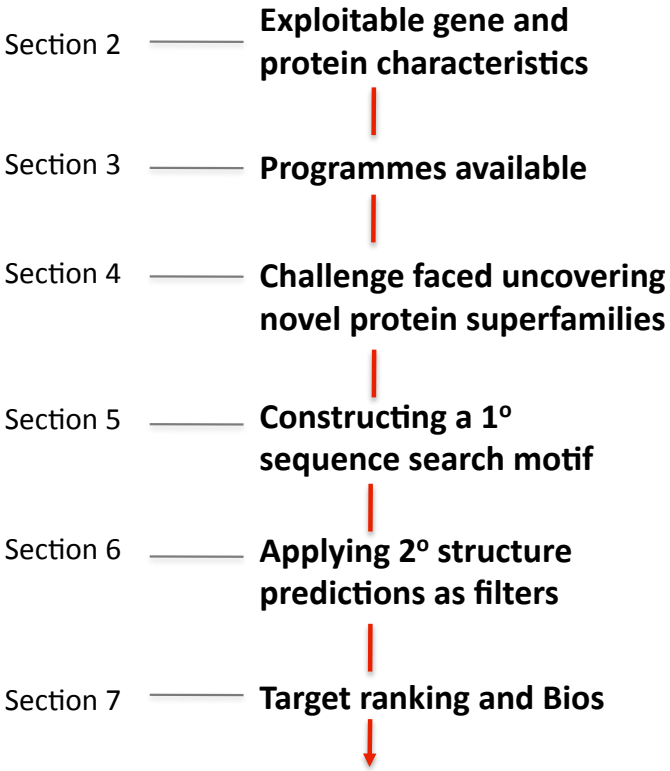


Figure 2.1.1 – Constructing a search strategy. This chapter will introduce the concepts of Bioinformatics and what can be expected, particularly from the perspective of a structural genomics project. Before moving onto describe the range of software available. The final portion of the chapter will introduce each of the shortlisted glutamine de-amidase targets, whilst ranking them in order of interest.

2.2 – Exploitable gene and protein characteristics

Gene sequences for the most part are directly related to protein sequences across most Prokaryotes, which means these sequences can be used to predict not only the location of genes encoding proteins within a genome, but also the amino acid composition of the resulting protein. Therefore, from a well-annotated genome, it is possible to assign first operons and then promoter regions. The relative promoter strength, along with any information regarding the characterisation of other genes within a given operon can provide a wealth of information regarding a candidate proteins function and importance. It is also possible from comparing sequences, to identify proteins that share a common evolutionary ancestor and will likely exhibit similar folding patterns and functional characteristics.

Knowledge of the amino-acid sequence can lead directly to functional identification. This is because residue composition particularly at the active site is normally conserved to a varying degree across all but the most distantly related homologues. Therefore, functionally related proteins that do not share significant global sequence similarity can usually be identified through short sequence motifs, which account for the active site. However, codon usage is different in all organisms therefore protein sequences can also be used to reconstruct genes to become more suitable for over-expression in a given microorganism. Finally, primary sequence information can also be used to attempt prediction of its secondary structure characteristics.

2.3 – Bioinformatics software encountered in this thesis

Bioinformatics is a common first stage in many structural biology projects, used both to identify potential targets but also to search for homologues that can be utilised as search ensembles in Molecular replacement. However, this thesis is going to examine a case where the mainstay methods of target acquisition were not capable of returning high confidence targets. Necessitating the use of a multi-faceted search strategy, employing multiple programmes, to uncover a selection of candidate targets.

2.3.1 – Gene and protein sequence homology search tool: BLAST

BLAST is the most popular sequence homology search engine, it is hosted on the NCBI website and aggregates a wide variety of the most popular sequence databases. BLAST compares an input sequence with its potential homologues via a substitution matrix, filtering out unlikely matches based on sequence identity. It is composed of two separate algorithms, run one after

the other. The first is called Gapped-BLAST and allows for significant gaps between regions of sequence similarity, to account for insertions. The second is PSI-BLAST that takes any identified homologues from the Gapped-BLAST output and uses them to create a more accurate position-specific substitution matrix (Altschul *et al.,* 1997).

BLAST is particularly useful because it incorporates over 30 searching options (Jones and Swindells., 2002), along with robust statistics quantifying the likelihood of the observed similarities occurring at random. These statistics are important because the success of a BLAST search centres on correct assumptions being made during the construction of the substitution matrix. However, BLAST only manipulates sequence information it makes no concessions for either the function or fold when the gene product has been characterised. Therefore, it is used predominantly to identify close homologues for comparative studies.

2.3.2 – Primary sequence functional motif search: PATTINPROT

PATTINPROT, like BLAST, is a web-based search engine (Combet *et al.,* 2000). PATTINPROT takes short peptide motifs written in PROSITE syntax, searching specified sequence databases to find any proteins that share the query pattern. When searching for proteins based on shared function, this search motif often corresponds to the arrangement of conserved amino acids located in or around the active site. Therefore, it is an essential tool when searching for either distantly or functionally related proteins, which share little sequence homology.

2.3.3 – Tertiary structure homology search tool: Dali-Lite

Dali-Lite is a tool for comparing 3D models with the vast selection of solved structures on the PDB (Holm *et al.,* 2008). Proteins that share functional characteristics, but share little sequence similarity often exhibit similar tertiary folds. This is because structural characteristics tend to be more stable than sequence conservation throughout evolution (Baker and Sali., 2001). As a result it is often possible to piece together distantly related super-families through structure comparison, which would not have been matched through sequence homology searches. Dali-Lite works by scanning the PDB with a variety of filters, the heavy computing load is reduced by combining homologues with sequence identities over 90 % into batches that are only searched once. Therefore, Dali-Lite is an essential tool for analysing proteins of unknown function, as structural identification can often lead to clues regarding a proteins function. This is the basis of structural genomics, which often relies on the produced structures exhibiting similarity with characterised proteins, to help in assigning a function.

2.3.4 – Protein secondary structure prediction: PSI-PRED

PSI-PRED predicts the secondary structure components of a query protein, based on its amino acid sequence (Jones., 1999). It incorporates the sequence similarity of related homologues as an integral facet of the prediction using two feed forward neural networks. These networks analyse the output of a PSI-PRED job, constructing a position specific matrix to identify conserved positions. Therefore, this matrix gives an estimated probability of how likely it is that a given residue will be located in that position. Using this information portions of the sequence can then be identified that exhibit similar residue types consistently. The secondary structure is then assigned based on values estimating how accessible that position or region will be to the surrounding solvent; allowing for assignment of either an α-helix or a β-strand for each stretch of residues (Benner and Gerloff., 1991). PSI-PRED is one of the most consistent secondary structure prediction tools available, however this form of prediction is by no means biologically accurate. Therefore, It is typically used to profile prospective targets based on residue distribution, offering no information regarding the queries tertiary fold.

2.3.5 – Threading: PHYRE$_2$

PHYRE$_2$ is a template homology tool for the detection of homologues with known 3D structures. PHYRE$_2$ aggregates PSI-BLAST scoring matrices along with multiple secondary structure prediction algorithms to identify a consistent putative folding pattern. It then predicts the tertiary structure of the query, based on mapping the fold to the nearest homologues determined by a combination of sequence homology and secondary structure predictions (Kelley and Sternberg, 2009). It is particularly useful when isolating potential targets with poor sequence conservation, which exhibit strong structural preservation.

2.4 – Uncovering novel glutamine de-amidases

There are currently three phenotypically unique glutamine de-amidase enzymes that have been characterised, all of them with 3D structures. They are: C-CNF1, BLF1 and CheD (chapter 1). The aim of this project is to uncover additional glutamine de-amidase enzymes from novel species, to expand this interesting super-family. All three enzymes conserve a central β-sandwich flanked by α-helices, with their respective active sites comprised of a catalytic CYS-HIS dyad.

However, typical search strategies involving BLAST and $PHYRE_2$ in isolation do not identify many novel targets, just analogues from closely related species. What is especially indicative of the challenge faced in identifying new glutamine de-amidase enzymes; is that neither BLF1 nor C-CNF1 returns one another in $PHYRE_2$, despite displaying strong structural conservation. Therefore, because both sequence and structural template homology tools are unable to identify the currently characterised enzymes, it is unlikely that they will resolve any sequentially novel distant homologues.

$PHYRE_2$ fails to identify the characterised glutamine de-amidase proteins because they share little primary sequence similarity. BLF1 and C-CNF1 for example share 7% sequence conservation, which is barely above the probable 5% similarity expected between two unrelated proteins. Therefore, one route that was explored to identify novel glutamine de-amidase proteins was to produce a short sequence motif, based on the conserved residues found across the 3 distantly related examples available. Sequence alignments based on the superposition of these 3D molecules along their active sites, are shown for BLF1 and C-CNF1 (figure 2.4.2) and C-CNF1 and CheD (figure 2.4.3). These alignments show that the small portion of sequence identity conserved across all three enzymes, is clustered in or around the active site (figure 2.4.4).

2.5 – Constructing a primary sequence search motif

Shown below are the conserved positions identified near to the active site across both BLF1 and C-CNF1 (figure 2.5.1):

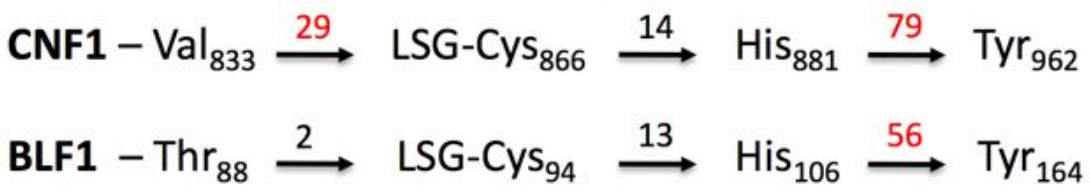


Figure 2.5.1 – Sequence conservation across the toxin glutamine de-amidase enzymes. There are no other positions present in both CNF1 and BLF1 that combine these two enzymes considerable structural alignment with residue conservation. However, within this conserved region there are significant sequence insertions, which limit the resolving power of any resultant primary sequence motif. The red numbers correspond to the gaps between functional residues, highlighting those that if included would weaken the search motif.

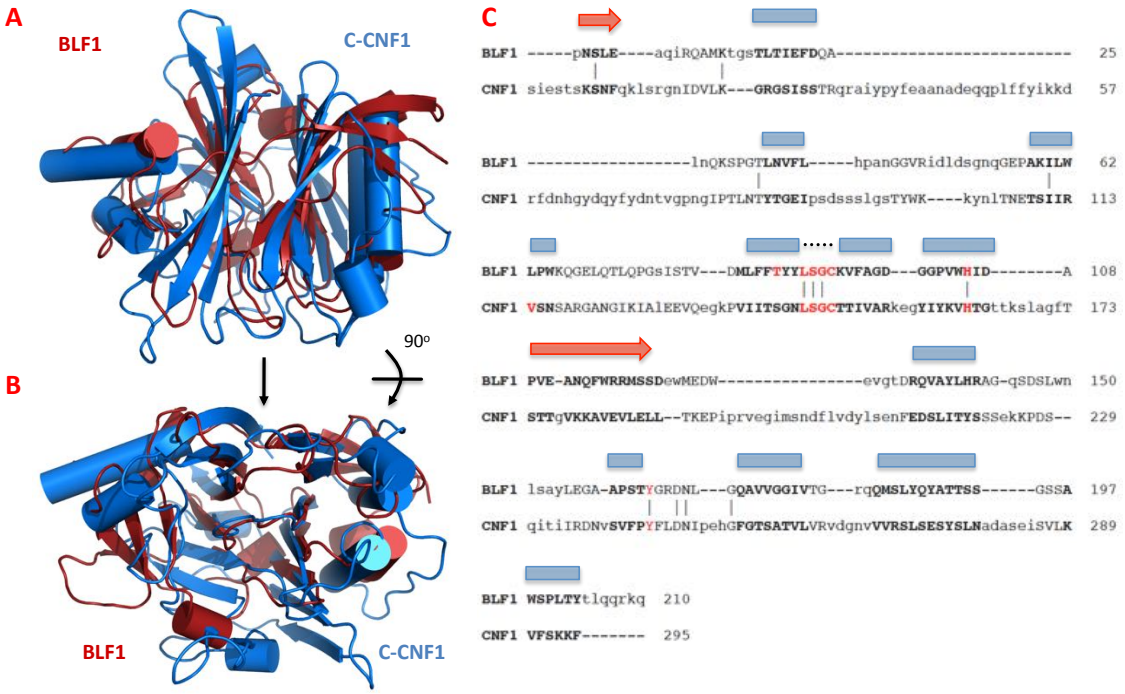


Figure 2.4.2 – Structure based sequence alignment of BLF1 with C-CNF1. Panels **A** and **B** display a superposition of C-CNF1 in blue with BLF1 in red. The central β-sandwich is very closely conserved along the protein backbone, however the flanking α-helices are not aligned. Panel **C** is a structure-based sequence alignment with secondary structure indicated above the primary sequence, with red arrows for α-helices and blue rectangles for β-strands. This alignment indicates that BLF1 and C-CNF1 share little sequence similarity apart from at the active site, highlighted in red. Where there is a strong L-S-G-C motif with HIS and TYR residues conserved downstream. Diagram constructed using PDB ID: 1HQ0 (Beutow *et al.,* 2001), PDB ID: 3TU8 (Migoni-Cruz *et al.,* 2011) and DaliLite (Holm and Rosenström, 2010).

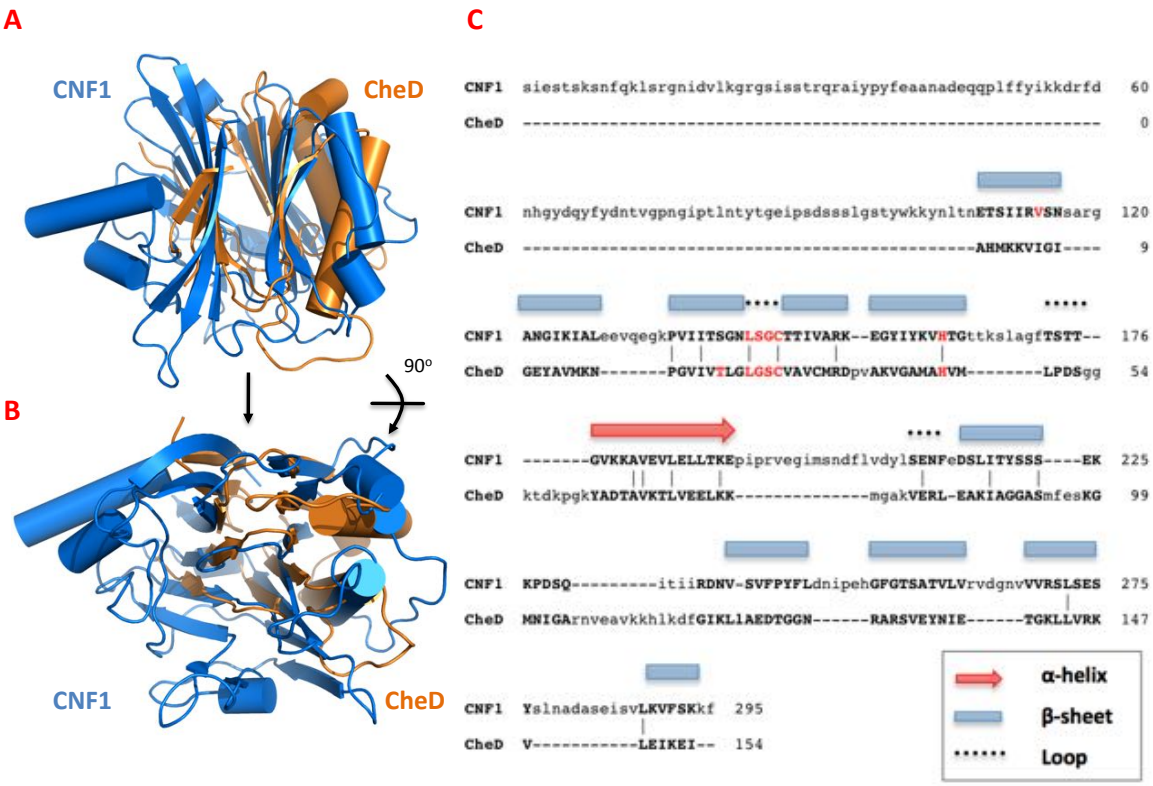


Figure 2.4.3 – Structure based sequence alignment of CheD with C-CNF1. Panels **A** and **B** display a superposition of C-CNF1 in blue with CheD in orange. The central β-sandwich is conserved. However, the flanking α-helices on the right hand side of C-CNF1 are not present in CheD. Panel **C** is a structure-based sequence alignment with conserved secondary structure features indicated above the primary sequence, with red arrows for α-helices and blue rectangles for β-strands. This alignment indicates that there is little primary sequence similarity between these enzymes apart from at the active site (red). However, close inspection of the conserved residues reveals that the central LSGC motif has been exchanged for an LGSC in CheD. Diagram constructed using PDB ID: 1HQ0 (Beutow *et al.,* 2001), PDB ID: 2F9Z (Chao *et al.,* 2006) and DaliLite (Holm and Rosenström, 2010).

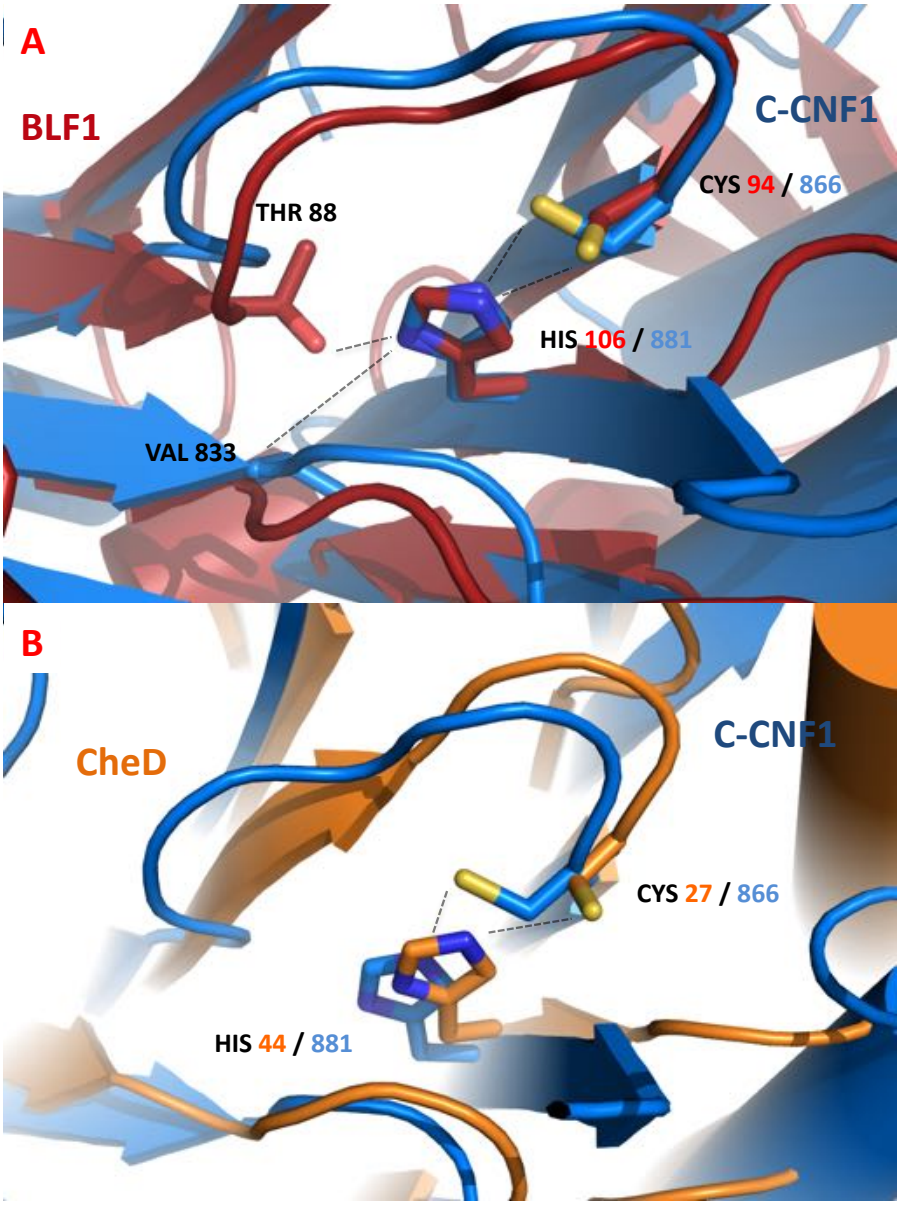


Figure 2.4.4 – Active site comparison between the currently characterised glutamine de-amidase enzymes. Panel **A** displays C-CNF1 in blue and BLF1 in red aligned along the conserved LSGC motif, the active site residues have been expanded with hydrogen bonds displayed by dotted lines. These two structures align closely in the active site, with the catalytically essential CYS –HIS dyad overlaying closely. The key difference between these two structures, is where the other residue interacting with the HIS position is located. In C-CNF1 VAL 833 occupies this position and is located higher and further away from the HIS, when compared to THR 88 which fulfils this role in BLF1. Panel **B** displays a superposition of CheD in orange with C-CNF1 in blue. CheD does not match with C-CNF1 as closely, but the catalytic dyad is broadly analogous. Diagram constructed using PDB ID: 1HQ0 (Beutow *et al.,* 2001), PDB ID: 3TU8 (Migoni-Cruz *et al.,* 2011), PDB ID: 2F9Z (Chao *et al.,* 2006).

With several of the conserved locations having been discounted, what is left behind is a short motif encompassing the conserved loop containing the nucleophilic CYS, along with the other member of the catalytic dyad, a HIS residue located on an adjacent β-strand.

Initial search motif: **LSGC – X (10,16) – H**.

This motif is strong, with PATTINPROT returning 2500 hits from the non-redundant database using these terms. However, most of these hits belong to a family of membrane anchors that also incorporate the LSGC motif (which the above search motif is anchored around) at their N-terminal region. The key issue is that the LSGC motif is not fully conserved. Homologues of BLF1 have been identified with an LAGC motif and CheD adopts a LGSC pattern. Therefore, the conserved locations across all three examples would be better represented by an **L-[AGSTV](2)-C** motif. Where the diverging positions, located on a short conserved loop, are limited to smaller residues.

When this new pattern is incorporated with the conserved HIS position, the search motif becomes: **L-[AGSTV](2)-C-X(10,16)-H**, which returns >65,000 hits. However, this is a very large number of candidates to examine, in stark contrast to the previous pattern, necessitating the production of a more discriminating motif.

Whilst the primary sequence identities exhibited by the characterised glutamine de-amidases appear almost random, the structural motif surrounding the active site is extremely well conserved. This allows assumptions to be made regarding amino acid distribution, particularly relating to solvent accessibility and probable secondary structures. The sequence alignments for all the glutamine de-amidases are shown below along with a consensus (figure 2.5.2):

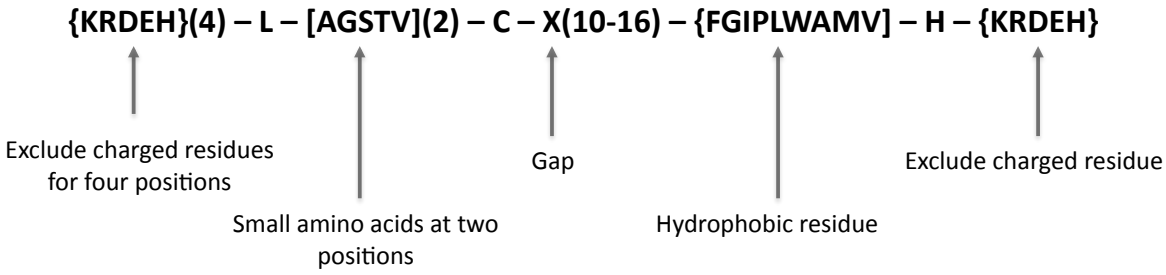
```
C-CNF1-  TSGN-LSGC-TTIVAR----KV-H-TG
  BLF1-  FTYY-LSGC-KVFAGD----VW-H-ID
  CheD-  VTLG-LGSC-VAVCMR----MA-H-VM
Consensus- BBBB-LBBC-XBHBHC----XH-H-BX
```

H - Hydrophobic
B - Buried
C – Charged
X – Mixed

Figure 2.5.2 – Residue conservation surrounding the active site across the glutamine de-amidase super-family. Structure based sequence alignments of the active site regions of all three glutamine de-amidase enzymes, show several regions of conserved residue type.

For example, the catalytic CYS is located directly after a β-strand in all 3 examples and the residue distribution for the 4 positions upstream of this CYS are consistently uncharged. However, there is no obvious patch of residue distribution to be taken advantage of directly after the catalytic loop. Until the HIS member of the catalytic dyad, which across all these enzymes is flanked on either side by first a hydrophobic residue and then a buried position. This information can readily be used to strengthen the motif, providing specificity towards the expected structure without discriminating against novel examples by imposing rigid sequence requirements.

The distribution of structurally buried residues can be harnessed to adapt the previous motif into:



The motif above results in >10,000 hits, which is a significantly larger pool of potential targets than the initial motif but far more manageable than the accurate but weak intermediate motif.

2.6 – Filtering out primary sequence targets

The primary sequence motif provides a list of >10,000 possible targets from the non-redundant database, an intimidating number. However, a large proportion of these hits are of known function, or from eukaryotic organisms that would potentially require more advanced expression systems than were locally available. Therefore, passing forwards only the hits from unknown or hypothetical proteins produced by prokaryotic organisms. Whilst also discriminating based on the location of the search motif within the candidate sequence, avoiding hits near either the N or C – terminal domain. The initially large target list, was reduced to just over 300 hits.

However, 300 candidate glutamine de-amidases were still too many for a small-scale structural genomics project. Therefore, the search strategy required filters to highlight 5-10 particularly interesting examples. This was achieved through a combination of secondary structure prediction (PSI-PRED) and tertiary fold template homology ($PHYRE_2$). $PHYRE_2$ proving

particularly useful, as it incorporates a batch mode where up to 20 sequences can be run as a single job. All 300 target sequences were tested using $PHYRE_2$, with candidates closely related to proteins of known function discarded. These filters produce a target list of 15 proteins, of unknown function. These proteins were then ranked based on interest, availability of the native organism and the likelihood of it being a functional glutamine de-amidase. Producing a shortlist of 5 targets to explore via a small-scale structural genomics project.

2.7 – Selected targets and priority assignment

The five targets explored in this section are the closest matches found using the search strategy described previously. This section aims to convince that these proteins are suitable candidates, whilst briefly offering a background into the organisms that produce them.

2.7.1 – *Hahella chejuensis*: HCH_03101

Hahella. chejuensis is a gram-negative facultative anaerobe from the γ-proteobacteria family, which requires NaCl concentrations over 2% (w/v) for growth (Lee *et al.,* 2001), making it a moderate halophile. Its genome is 7.2 MBp large and well equipped for survival in an ocean environment, it also encodes for several genes expected to convey pathogenic activity, via genomic islands containing both type III excretion systems and putative toxins (Jeong *et al.,* 2005). 16S ribosomal DNA analysis, indicates that the *Hahella* genus is distantly related to the *Pseudomonas* family. *H. chejuensis* is also implicated in the production of a red polysaccharide pigment, implicated in the destruction of harmful algal blooms, through the lysis of the constituent *C. polykrikaides* dinoflagellate cells (Kwon *et al.,* 2010).

The candidate is annotated as a hypothetical protein, labeled HCH_03101. It is 234 residues long and shares significant sequence similarity with BLF1 surrounding the toxins active site residues:

```
HCH_03101   48   RFNFVLGDGANSRTVQWLPWEEGGVTKLKLNRDS-ADYWFVTAALSGCSVGFDRNRGEVFHLN  109
                 |     |          ||||  |    |     |  |  | |  |||| | |    | | |
BLF1        46   RIDLDSGNQGEPAKILWLPWKQGELQTLQPGSISTVDMLFFTYYLSGCKV-FAGDGGPVWHID  107
```

Global sequence conservation is under 5% between HCH_03101 and C-CNF1, growing to 9 % shared with BLF1, when compared with BLAST. However, all the identified conservation is located within the active site region. HCH_03101 fits the primary sequence motif well,

exhibiting a conserved central LSGC motif along with nearby THR and HIS residues in conserved positions of the sequence.

The secondary structure prediction (figure 2.7.1) exhibits a largely β-stranded fold, with the LSGC motif correctly assigned to a loop and the coordinating residues both located on β-strands. There are however large stretches of un-assigned sequence, located between amino acids 110-135 and 140-172. It is unlikely that both these regions are loops that long, but BLAST searches on both regions yield no similarity to domains of known function.

PHYRE[2] predicts the tertiary structure using BLF1 as the template, consequently the tertiary fold exhibits a characteristic central β-sandwich. However, the aforementioned sequence conservation at the active site with BLF1 suggests HCH_03101 may be a distant homologue of the toxin. Nevertheless, with substantial deviations in sequence encountered elsewhere and what appears to be a large insertion between residues 140-175, it could equally be expected to exhibit a phenotypically distinct function. Therefore, HCH 03101 represents a high priority target for expanding the Glutamine de-amidase super-family.

2.7.2 – *Pseudomonas putida*: PPUT_1063

Pseudomonas putida is a gram negative, rod shaped bacterium that inhabits soil environments. It was the first microorganism to be patented and is of particular interest for its commercial uses in the bioremediation of oil (Wackett and Gibson., 1988). An undergraduate project student identified PPUT 1063, by using a less discriminating search strategy integrated into an automatic python script. Their strategy differed from the one presented, in that it did not discriminate based on the distribution of charged residues surrounding the conserved active site. As a result two charged residues, within what was previously held to be a buried region of the protein, precede the conserved LSGC motif.

The candidate protein PPUT_1063 is 239 residues long and does not conserve any significant patches of sequence similarity when aligned with either BLF1 or C-CNF1 outside of the conserved LSGC motif. However, unlike the previous candidate HCH_03101, PPUT_1063 lacks a conserved THR or VAL residue that would be anticipated to interact with the identified putative HIS of the catalytic dyad. Despite this reservation, PPUT 1063 is a promising target displaying a PSI-PRED secondary structure prediction (figure 2.7.2) closely resembling the currently characterised glutamine de-amidases.

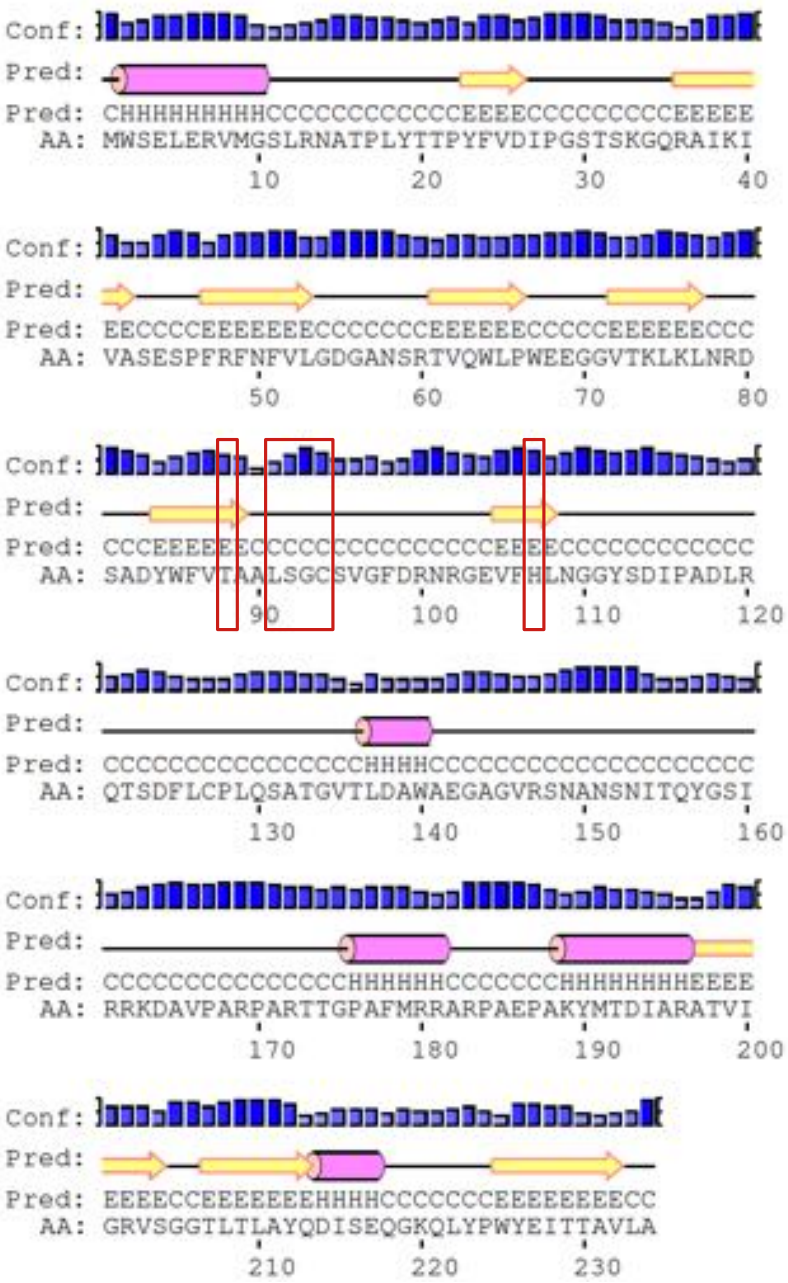


Figure 2.7.1 –PSI-PRED secondary structure prediction of HCH_03101. The PSI-PRED secondary structure prediction for HCH_03101 shows that the LSGC motif is centrally located within the protein. With the residues highlighted in red corresponding to the putative catalytic positions, which are all assigned to the correct secondary structure element. However, the distinguishing feature of this prediction is the long stretch of unassigned residues between positions 140-175, with the sequence offering no clues regarding the regions function or fold.

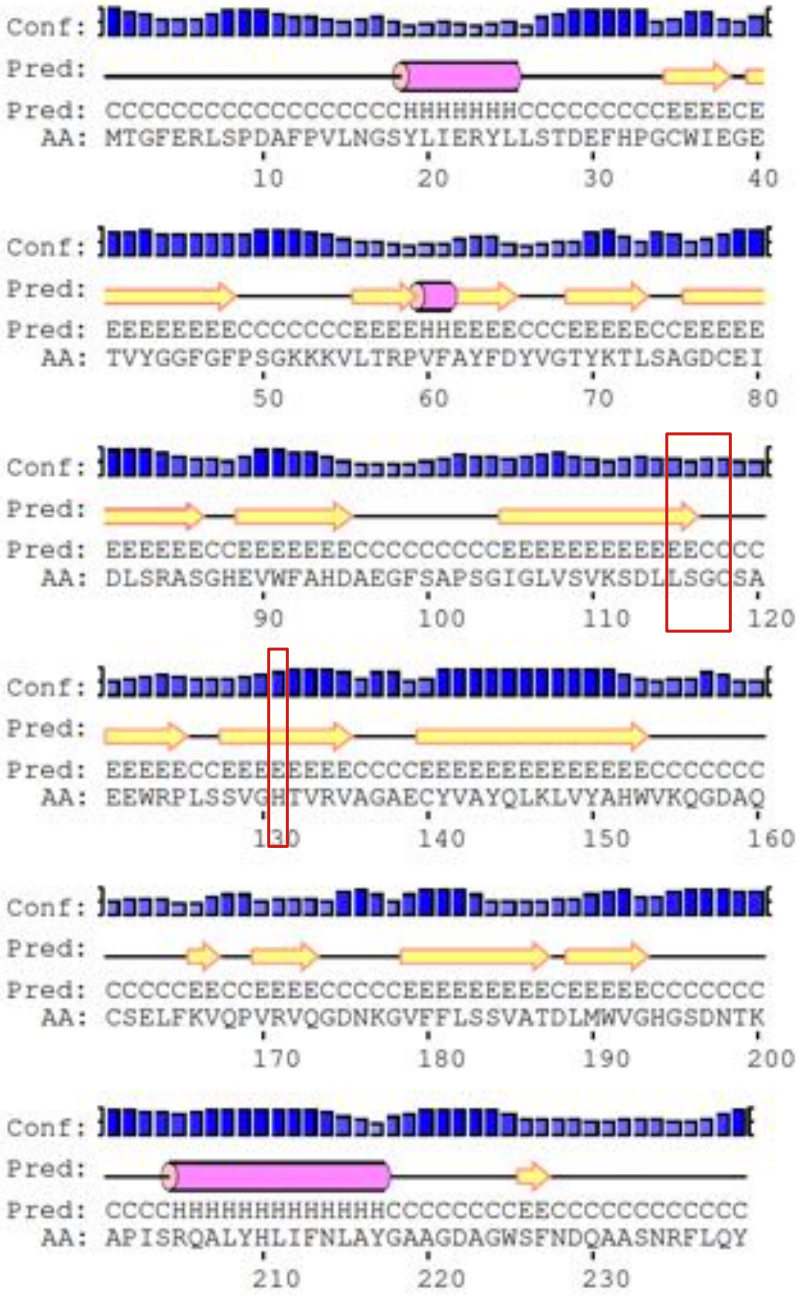


Figure 2.7.2 –PSI-PRED secondary structure prediction of PPUT_1063. The PSI-PRED secondary structure prediction shows that PPUT_1063 is likely to exhibit a predominantly β-stranded fold. The conserved residues believed to make up the active site, are highlighted in red. PPUT_1063 conserves the LSGC motif, seen in both BLF1 and C-CNF1, which is located towards the centre of the sequence well away from either terminus. The putative catalytic dyad residues CYS 118 and HIS 131 are also located on the anticipated secondary structure component.

With the catalytic residues correctly assigned to the anticipated secondary structure components and a predominantly β-stranded secondary structure overall. However, the PHYRE[2] tertiary structure prediction fails to produce a sensible model, due to a lack of suitable homology models to base the model on. Therefore, PPUT_1063 potentially represents either a distantly related member from a currently known protein super-family, or is from a novel class of uncharacterised proteins.

2.7.3 – *Serratia odorifera*: dermonecrotic toxin (DNT)

Serratia. odorifera is a gram negative, rod shaped facultative anaerobic member of the enterobacteria family (Grimont *et al.,* 1978). It is an opportunistic pathogen leading predominantly to urinary or respiratory tract infections (Julie *et al.,* 2009). Unlike, the previously described targets this protein has been assigned to the CNF1 super-family as a dermonecrotic toxin. However, this *S. odourifera* DNT is a much smaller protein at 138 residues long, than the previously characterised C-CNF1 at 194 residues. With Its size more closely resembling CheD than either of the toxin examples. This target conforms to the primary sequence motif with the exception of an additional upstream HIS residue at position 69, located within a putative buried location.

DNT shows significant sequence conservation with the active site of C-CNF1:

```
DNT    39   NVIEIANGNCGVIGIRFHLGQLKSNNPLLIHGGALSGCTIAFAIKDDCFYAFHCGQS  95
              |    |    |  ||    |        |   |   | |||||    | |     |   | |
CNF1   110  SIIRVSNSARGANGIKIALEEVQEGKPVIITSGNLSGCTTIVARKEGYIYKVHTGTT  166
```

Across the whole protein it shares 15% sequence conservation with C-CNF1 and 5% with BLF1. Secondary structure predictions show that *S. odourifera* DNT, conforms to a predominantly β-stranded fold (figure 2.7.3). With the LSGC motif correctly assigned to a loop with the conserved HIS 91 located on an adjacent β-strand. Tertiary structure predictions select C-CNF1 as the closest homologue, as a result the catalytic dyad is arranged as expected. However, due to its shorter primary sequence DNT cannot encode for the full central β-sandwich motif conserved across all the currently characterised glutamine de-amidase toxins and will not include the flanking α-helices on one side. These differences make *S. odourifera* DNT an interesting proposition for structural studies, despite strong sequence similarity in the active site likely indicating that it is both a functional and structural homologue of CNF1.

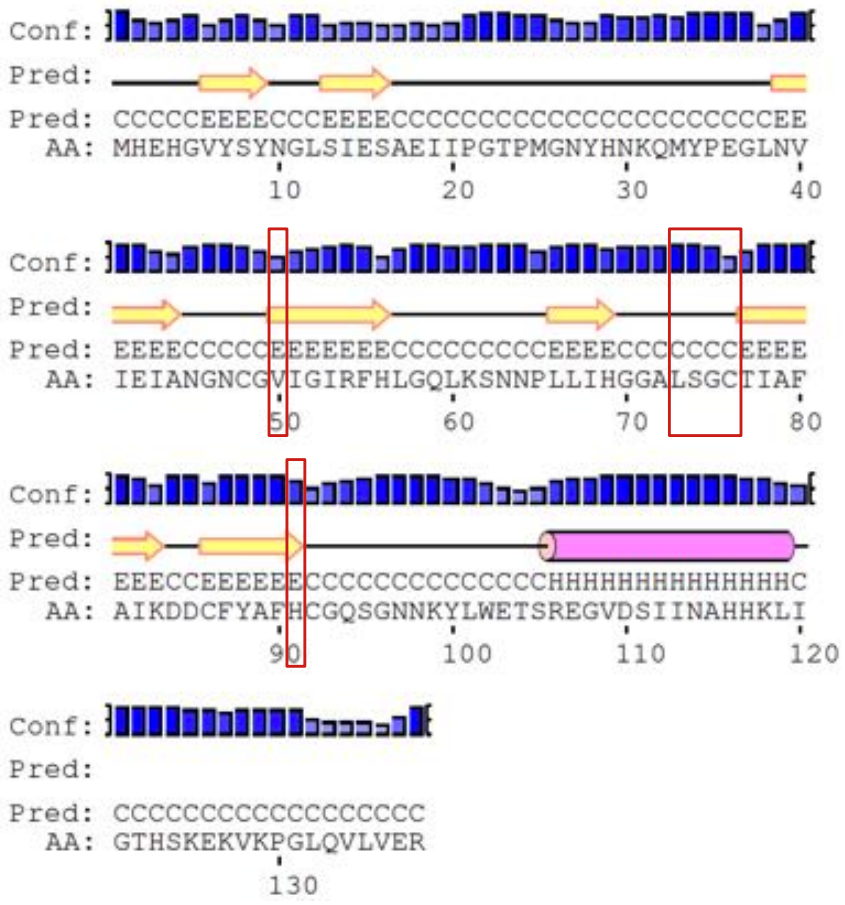


Figure 2.7.3 – PSI-PRED secondary structure prediction for *S. odorifera* DNT. The PSI-PRED secondary structure prediction for *S. odourifera* DNT is shown above, with the putative catalytic residues highlighted in red. The secondary structure is predicted to be predominantly β-stranded. However, the NTD α-helix observed in both BLF1 and C-CNF1 is not present. The sequence is over 80 residues shorter than the previously characterised glutamine de-amidase toxins, more closely resembling the non-toxic CheD. Consequently *S. odorifera* DNT could to be result of a gene truncation from a full-length glutamine de-amidase.

2.7.4 – *Vibrio splendidus*: VSII 1134

Vibrio. splendidus is a gram negative, rod shaped member of the gamma-proteobacteria family, a trait that appears to be shared across all the candidate proteins identified. Its natural habitat is seawater and ocean sediment, where it requires NaCl concentrations higher than 0.5M to grow (Baumann *et al.,* 1980). *V. splendidus* is a pathogenic bacteria capable of infecting marine organisms with a disease called vibrosis, which can then be passed onto humans via ingestion presenting with symptoms resembling severe food poisoning (Jensen *et al.,* 2003). Its genome is composed of two chromosomes the first encodes is 3.5 Mbp large and encodes for predominantly essential genes, the second is smaller at 1.6 Mbp and encodes accessory proteins.

VSII 1134 is a 212 amino acid hypothetical protein that shares 24% sequence similarity with the active site region of C-CNF1 shown below:

```
1134   69   LSSSAKALQCIHIPVSQFEQLKPESISKVTHYDSANFLVTTQLTGCTFAIRPGKGGGLEF  128
             | ||     | |        ||  |                | |||  |   | |
CNF1   114  VSNSARGANGIKIALEEVQEGKPVIIT-----------SGNLSGCT-TIVARKEGYIYK  160


1134   129  LHVQPNRDFDGAKIQQAIKKEFQV  152
             |        |       ||   |
CNF1   161  VHTGTTKSLAGFTSTTGVKKAVEV  184
```

However, on a global scale VSII_1134 and CNF1 only share 9% sequence similarity suggesting that this candidate may be a distant homologue of CNF1. Sequentially the major deviation from the previously characterised glutamine de-amidases is that the central LSGC motif has been substituted for a LTGC pattern.

The secondary structure prediction (figure 2.7.4) is promising with significant stretches of β-sheet and two distinct α-helices. The coordinating members of the catalytic triad are correctly arranged on β-strands but the LTGC motif is only partially contained on a short loop, however these encroaching β-strands are predicted with low confidence. The tertiary structure is poorly defined with no high confidence homology template identified. However, when the sequence is folded using CNF1 as a template despite little global sequence conservation the target does fold into a characteristic glutamine de-amidase structure, with the catalytic residues arranged as a triad. VSII_1134 represents the most likely toxin type glutamine de-amidase to add to the super-family from a human pathogen.

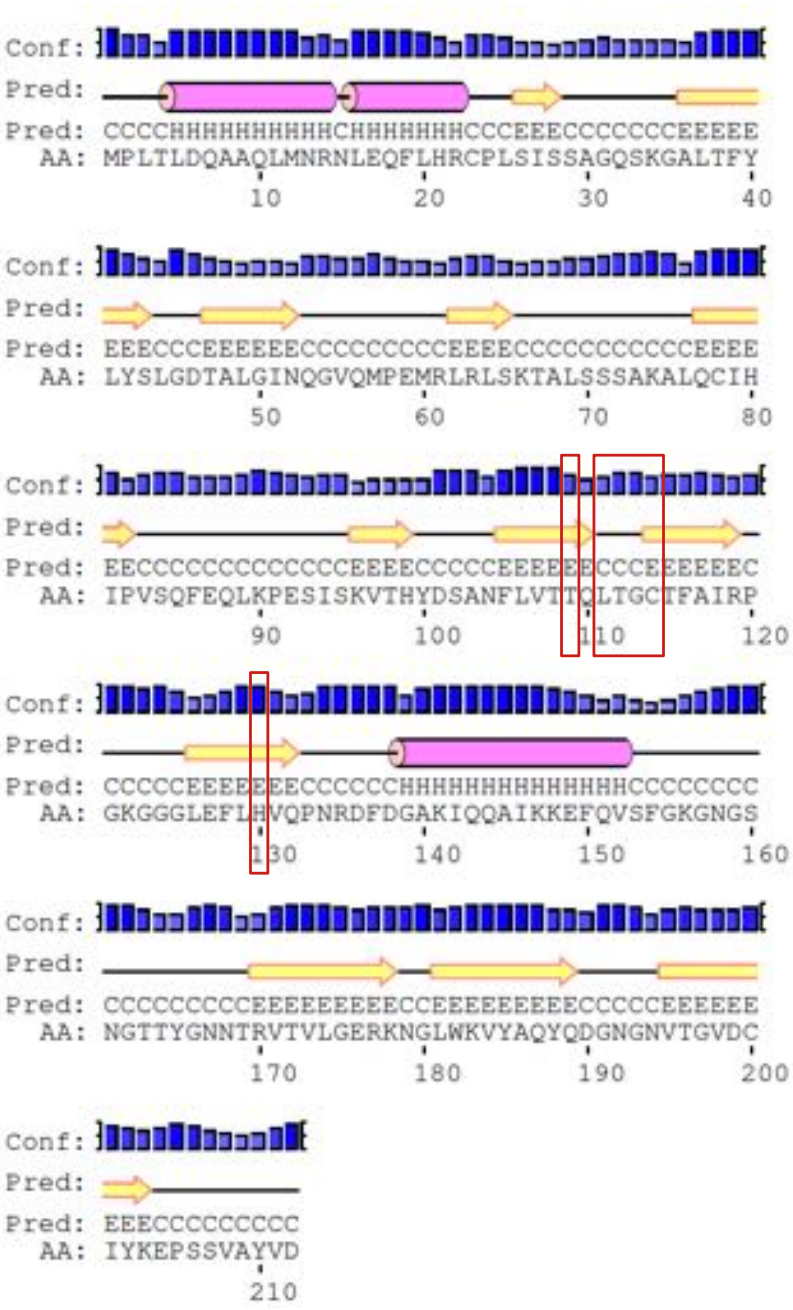


Figure 2.7.4 – VSII_1134 PSI-PRED secondary structure prediction The secondary structure prediction for VSII_1134 is shown above, with the catalytic triad highlighted in red. VSII_1134 is expected to fold into a predominantly β-stranded structure with two α-helical regions. The central motif is expressed as a LTGC motif rather than a LSGC, with the local buried residue distribution fitting the search strategy.

2.7.5 – *Pseudomonas stutzeri*: PSTAA_2862

Pseudomonas. stutzeri like *P. putida* before, is a gram negative, rod shaped soil dwelling bacterium capable of bioremediation. However, unlike *P. putida* it is also an opportunistic pathogen, which has previously been isolated from human spinal fluids (Lalucat *et al.,* 2006). PSTAA_2862 is a low confidence target displaying significant deviations, from the anticipated search filters. It also exhibits large regions of high confidence α-helix, with the HIS component of the catalytic dyad located on one (figure 2.7.5). In spite of these predicted deviations the search motif fits well with no unexpected charged residues in putative buried positions. PSTAA_2862 shares no significant sequence similarity with any previously characterised proteins and as a result all template homology prediction methods yield disordered tertiary structures. This target represents a low chance of expanding the super-family and is the lowest priority target moving onwards.

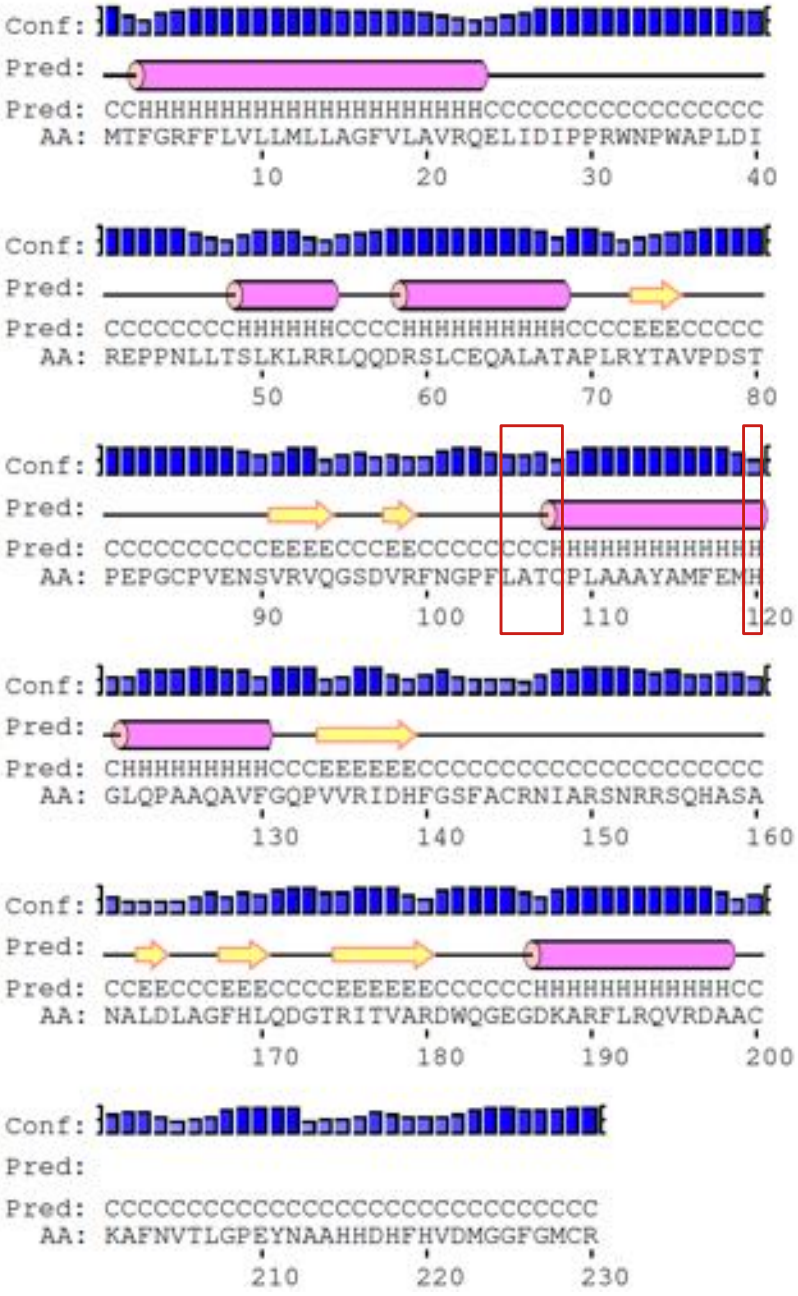


Figure 2.7.5 – PSTAA_2862 PSI-PRED secondary structure prediction. The secondary structure prediction for PSTAA_2862 predicted using PSI-PRED, shows that the typical LSGC pattern is not conserved; the catalytic CYS is instead preceded by an LAT motif highlighted in red. However, despite conserving the primary sequence motif and the buried residue distribution, the predicted secondary structure is predominantly α-helical.

Chapter 3: X-ray crystallography theory

All the structures represented in this thesis have been determined with data obtained from X-ray diffraction experiments. Using X-rays to resolve atomic detail began in 1895, when Wilhelm Röntgen first isolated the wavelengths of electro-magnetic radiation now characterised as X-rays. Max von Laue in 1912 showed that X-rays could be diffracted at measurable levels through crystalline samples (Eckert, 2012). However, the first X-ray structures solved were of small molecule salts such as NaCl and KCl by Lawrence and William Bragg in 1912 (Bragg, 1913). It took over 50 years for the first protein structure to be solved in the form of Myoglobin (Kendrew *et al.,* 1958). The field has since flourished with the number of structures deposited in the Protein data bank (PDB) currently exceeding 100,000, of which 90,000 were solved using X-ray crystallography. This chapter is going to detail first the theory and methods involved in obtaining crystalline materials, followed by the collection and interpretation of X-ray diffraction data (figure 3.1.1).

3:1 Protein crystallisation

3.1.1 Bragg's law

The production of crystalline samples is necessary when collecting diffraction data using X-rays. This is because X-rays with their high energy, short wavelengths (λ), pass through most low-density materials without interacting with them. With this in mind Lawrence Bragg's greatest contribution to X-ray crystallography was his insight into how incident X-rays constructively interfere with one-another provided that the path difference between rays reflected from successive planes are equal to an integral of their λ (figure 3.1.2) (Bragg and Bragg, 1913). This law can be defined by the following equation:

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

Equation 3.1.1 – Bragg's law. d_{hkl} is the distance (Å) between separate planes within a crystal with indices h, k and l; θ is the angle at which the incident X-ray strikes the plane; *n* is an integer and λ is the wavelength of the incoming X-ray.

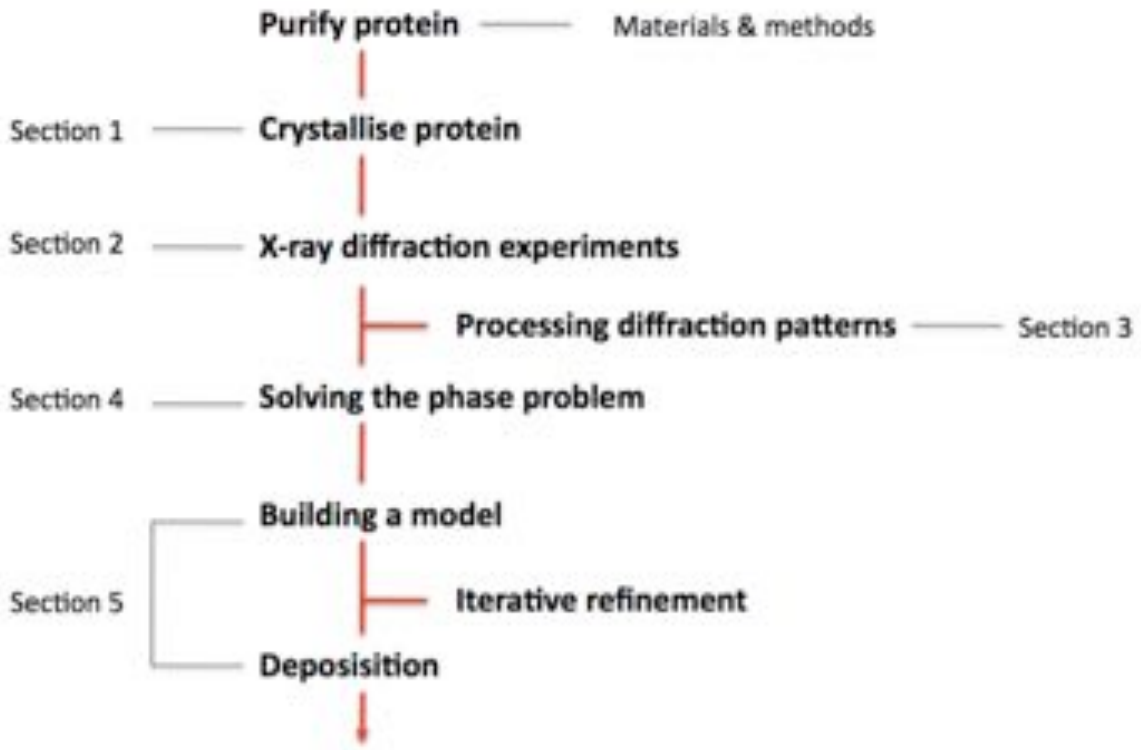


Figure 3.1.1 – Flow diagram outlining the methodology of an X-ray crystallography experiment. X-ray crystallography experiments predominantly follow a standard protocol. First the desired recombinant protein is purified; this pure sample is then placed in crystal trials to establish a set of parameters that will allow the formation of single crystals. These crystals are then tested with the best conditions optimised across smaller variations in hanging drop trials. In most cases any crystals produced were tested at home to identify well diffracting samples and conditions, with the top examples sent on to a synchrotron to collect the best quality data set possible. This data comprises thousands of 2D images, but the eventual protein model will be built into a 3D electron density map. Therefore, the 2D diffraction patterns require processing to produce a merged dataset from which a model can be built. The final experimental hurdle is to identify the phases of the reflected X-rays. However, the initial experimental phases contain significant errors, improved upon with information provided by a model in a process termed refinement. Refinement is an iterative process, undertaken until the model accounts for the majority of the information provided by the data. The final stage of the process is to validate and deposit the structure in the PDB.

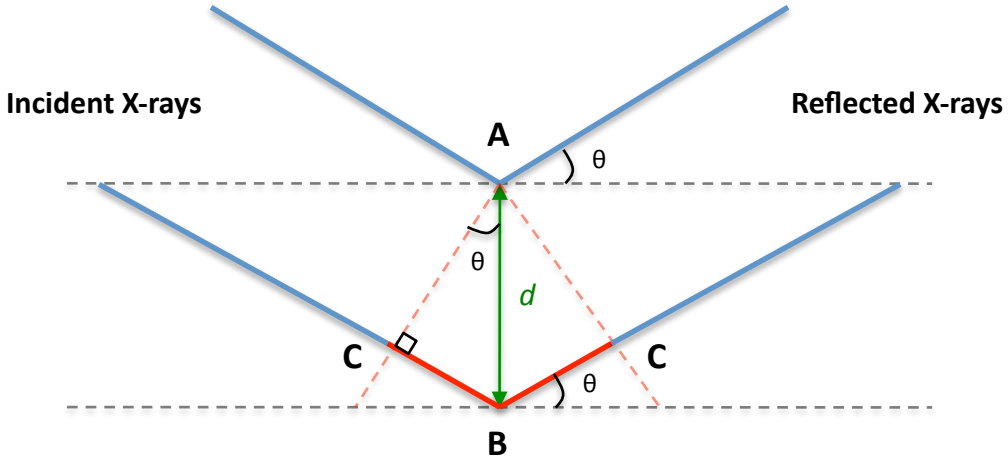


Figure 3.1.2 – Bragg's Law - Conditions for the production of diffracted X-rays. When a crystal is exposed to X-rays, thousands of crystallographic planes interact with the incident X-ray beam. In order for these scattered X-rays to constructively interfere with one another, thus allowing measurement, the path difference between the scattered beams has to be an integer of their wavelength. This can be summarised as: $n\lambda = 2d\sin\theta$. The principle of path difference is best described geometrically. Using the diagram above the incident X-ray interacts with successive planes of a crystal, separated by distance *d.* This leads to a path difference between the two reflections of 2CB. However, if a right angle triangle is constructed (red line) with the hypotenuse corresponding to the *d* and with the angle θ positioned at point A, then the path difference can also be expressed as 2d sinθ.

When Bragg's law is not satisfied the path difference of the reflected X-rays are no longer in phase and will interfere with one another destructively, preventing measurement.

3.1.2 Protein crystal composition

The most prominent interactions commonly found within the crystal lattice are hydrogen bonds, formed between residue side chains or intermediate water molecules and van der Waals forces. Protein crystals are generally smaller and more fragile than those encountered with small molecules. This is because the crystal contacts made between the constituent protein molecules are small in both number and size.

3.1.3 Protein crystal growth

A common crystallisation experiment is the vapour diffusion method. Vapour diffusion works by drawing water out of a solution containing the target protein mixed with buffering and precipitant components, into a much larger reservoir of just buffer and precipitant. Reducing the water concentration raises both the protein and precipitant concentrations and when the rise in protein concentration reaches super-saturation, the protein molecules in certain conditions are pushed towards a crystalline state.

3.1.4 Vapour diffusion

There are two classic examples of a vapour diffusion experiment. The first is the sitting drop (figure 3.1.3A), which involves placing the droplet on top of a plastic well located alongside a precipitant reservoir. These trials are typically laid down robotically 96 drops at a time in pre-set trials, designed to elucidate suitable crystallisation conditions, termed sparse matrix screening. The second type of trial is the hanging drop (figure 3.1.3B), which adheres the droplet onto a siliconised glass slip holding it above the precipitant reservoir. Hanging drop trials are predominantly used to optimise established growth conditions to improve both their size and diffraction properties. Typically vapour diffusion experiments require the protein to super-saturate and concentrate into the spontaneous nucleation phase where crystals will form. From this point onwards protein concentration will decrease while the relative precipitant concentration increases, this imbalance is what drives crystal growth until the protein in solution has been depleted (figure 3.1.4).

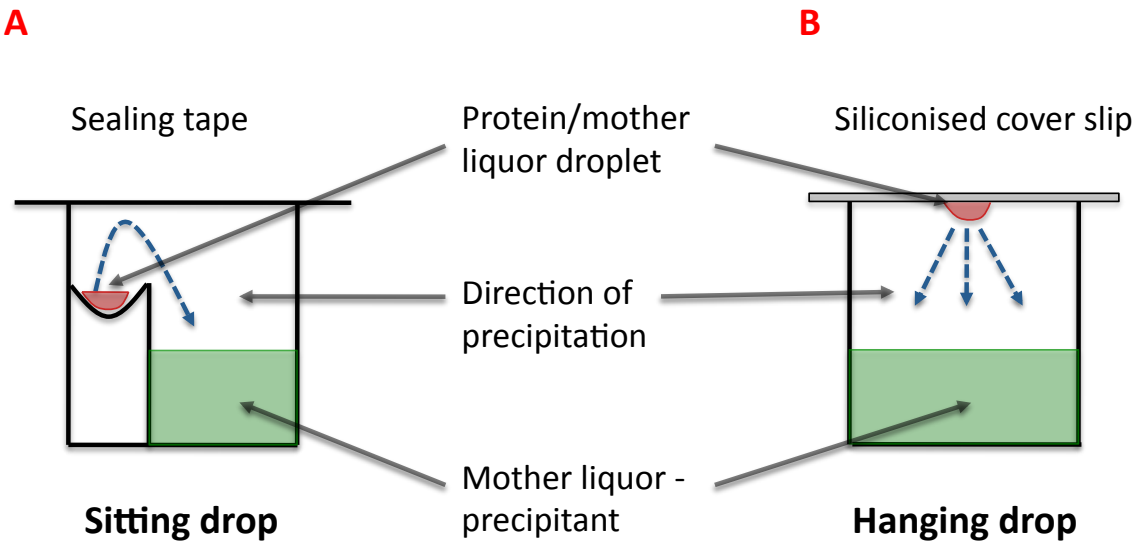


Figure 3.1.3 – Vapour diffusion crystallisation trials. There are two common formats of vapour diffusion. The first are sitting drop experiments with pre cast plastic trays incorporating both a well and reservoir as shown in panel **A**. These trays commonly accommodate drop sizes approaching 200 – 500 nl, administered by robotics and sealed with a layer of adhesive tape. They are predominantly used as part of a sparse matrix screening strategy. However, there are several drawbacks to sitting drop experiments, relating to crystal extraction and the irreversible binding of some proteins to the plastic trays. The second form of vapour diffusion involves hanging drops shown in panel **B**. These are normally laid down by hand taking considerably longer to produce; as a result they tend to be reserved for optimisation of known conditions. The major advantage of hanging drop trials is that the crystals are easier to extract. They are also useful in scaling up the size of crystals with much larger drops and different ratios of protein to mother liquor possible.

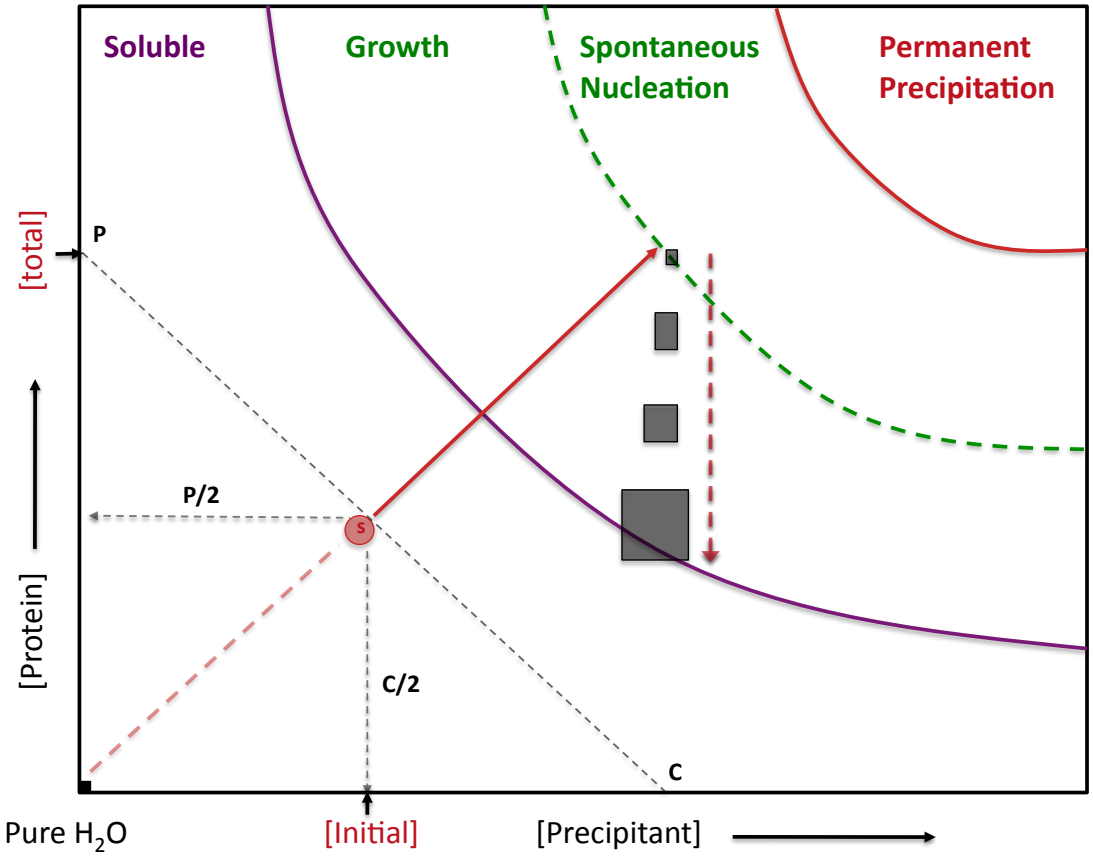


Figure 3.1.4 – Phase diagram for a successful vapour diffusion experiment. This diagram shows a crystallisation trial with an initial droplet containing 1:1 ratios of protein to mother liquor. The initial concentrations of protein and precipitant are represented by black dotted lines labeled P/2 and C/2 respectively. The driving force of vapour diffusion is the precipitant, with the reservoir dragging water away from the droplet bringing it into equilibrium. Therefore, it is important that the P and C values (at equilibrium) correspond to an intersection in the in-soluble phase where spontaneous nucleation is possible. Once nucleation has occurred the crystal will grow, with growth driven by the protein: precipitant imbalance in the droplet as more protein is incorporated into the crystal. Crystal growth generally halts when the crystalline protein content equals that of the initial droplet at P/2.

However, the process described above is rarely encountered in real trials (figure 3.1.5). Instead, a common occurrence is clear droplets, where the protein never reaches the super-saturation point. This lack of nucleation is normally due to insufficient precipitant concentrations. Another alternative is the formation of micro-crystals. Typically large numbers of small crystals occur when there is a higher precipitant to protein ratio at the point of nucleation. This large number of tiny crystals prevents the formation of larger crystals because the total protein content in the drop has been depleted. Nevertheless, these small crystals are useful as they can be introduced to fresh droplets, adding nucleation sites and reducing the precipitant concentration required for super-saturation. Unfortunately, crystal growth prior to the establishment of a suitable set of conditions is highly unpredictable. In order to ascertain the correct conditions, hundreds of trial conditions are tested prior to optimising around a smaller selection of promising hits.

3.1.5 Protein crystal growth defects

Crystals predominantly grow from a single nucleation point with each new unit preferentially binding onto a ledge, step or gap in the existing lattice. As crystal growth progresses the supply of homogenous target protein decreases, this opens the door for contaminants or non-homogenous isoforms of the target protein to bind onto the pre-existing lattice. However, there is also a small chance that larger debris contained within the crystallisation condition will act as a nucleation site, becoming incorporated into the crystal (figure 3.1.6). These defects lead to highly mosaic crystals, which can limit the resolution and introduce anisotropy.

3.1.6 Cryogenic crystallography

When exposed to X-rays at room temperature the radiation damage incurred, due to free radical formation, normally prevents collection of a whole dataset from a single crystal. To circumvent this crystals are commonly frozen in liquid nitrogen to cryogenic temperatures, immobilizing the damaging free radicals and prolonging the crystals exposure life, by a factor of 70 (Nave and Garman, 2005). Prior to freezing, the crystals are mounted on a loop along with a small volume of mother liquor. However, due to the high solvent content in protein crystals special care has to be taken when freezing them to avoid Ice crystals. External ice crystals commonly cause the crystal to shatter, but hidden within the lattice internal ice crystals can introduce or amplify mosaicity.

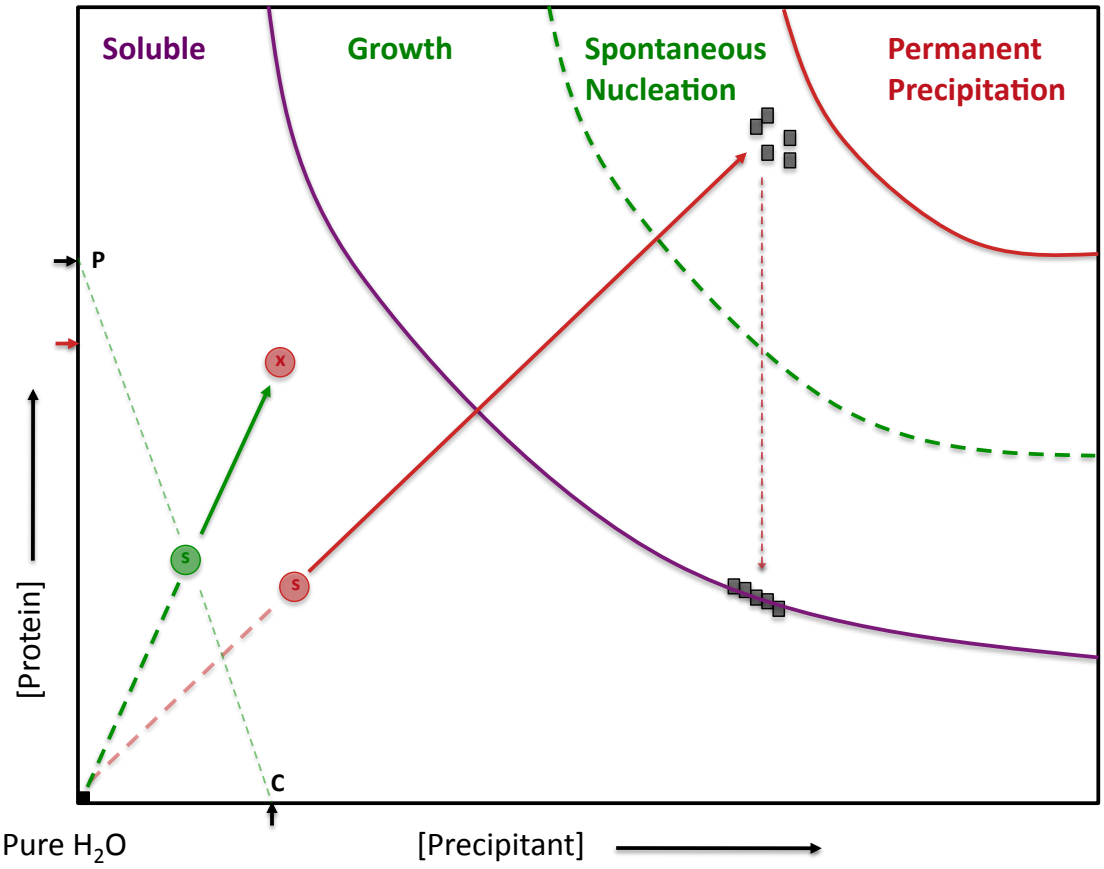


Figure 3.1.5 – Phase diagram detailing failed vapour diffusion experiments. The most common example of a failed crystal trial is non-crystalline precipitate, which occurs when either the precipitant or protein concentrations are so high that the phase diagram intersects in the thermodynamically unstable region, highlighted in red. Analysing sparse matrix screens often reveal conditions producing clear drops, or small micro-crystals unsuitable for use in diffraction experiments. Clear drops are usually symptomatic of a lack of precipitant, with the reaction unable to reach the in-soluble phase. Whereas, the occurrence of micro-crystals is due to nucleation occurring further into the in-soluble phase, with lots of tiny crystals forming, which depletes the protein content of the drop.

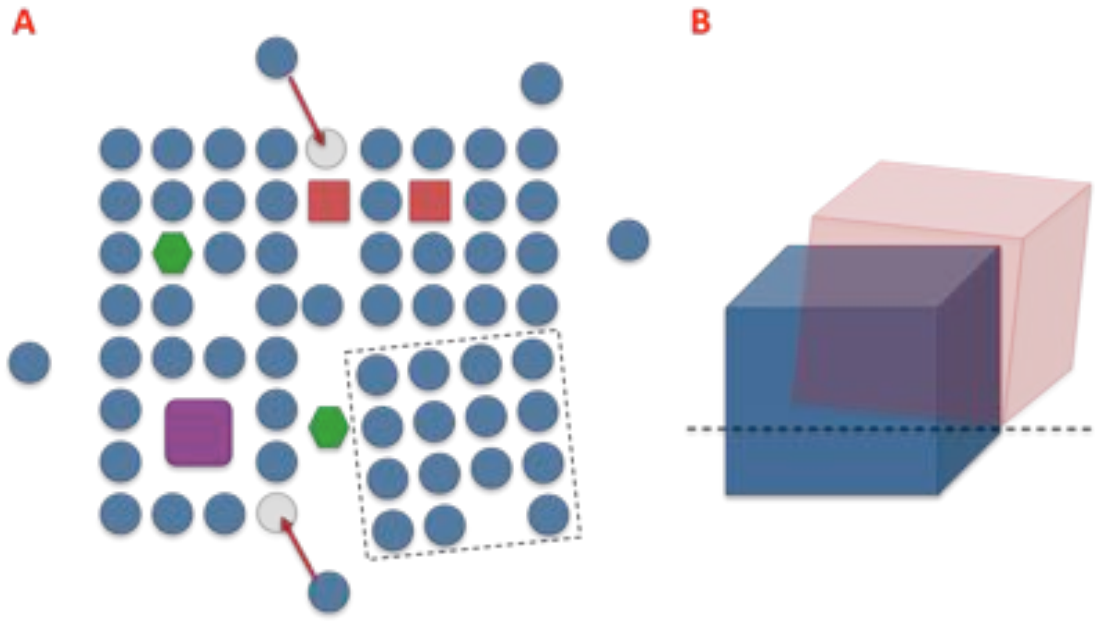


Figure 3.1.6 – Schematic of a high mosaicity crystal with poor lattice formation and twinning. Panel **A** represents a highly mosaic crystal. The homogenous target protein is represented in blue, with trace impurities in red and green. The large purple block is debris capable of forming a preferential nucleation site. Crystal growth rarely occurs along planar surfaces, instead bonding on steps, ledges and gaps as indicated by the red arrows. This allows for the integration of impurities and the introduction of gaps in the lattice, where the integral water is far more flexible than a lattice component. Another major defect found mostly in large crystals is off centre domains highlighted in the black dashed box. These factors on their own reduce the quality of diffraction and together will totally preclude it. Panel **B** shows how two independent crystal lattices can combine and appear homogenous, whilst actually being out of alignment, this is termed macroscopic twinning.

To avoid ice crystals great care is taken to match the mother liquor with a cryo-protectant composed of the same buffers and precipitants, along with 15-30 % (v/v) cryo-protectant. Protectants include Ethylene glycol, Glycerol or even low molecular weight PEG solutions and are soaked into the crystal for approximately 10 seconds. These chemicals displace surface water molecules, without disrupting the crystal lattice, refracting X-rays or forming crystalline structures at cryogenic temperatures. Prior to mounting crystals, cryo-protectants are tested by looping a small amount without a crystal and testing for ice formation on a home X-ray source. There are two ways of quenching the protected samples the easiest is to dip them directly into a static store of liquid nitrogen, in a single motion to rapidly cross the warm N_2 gas above. The second method utilises a cryo-stream freezing the crystals with a continuous stream of 100 K nitrogen gas. The frozen crystals are then stored in liquid nitrogen until required.

3.1.7 Crystal twinning

Twinning is a common issue in protein crystals, often yielding diffraction patterns that are difficult to interpret. There are three classes of twinning, two of which are readily identifiable with a final subtle type that is far harder to identify.

The first category is macroscopic twinning, which occurs when separate crystals grow from independent nucleation events coming together as a cluster. This is usually detectable when observing the crystals prior to looping them, however with flat plates or thin rod shaped crystals multiple lattices can be hard to identify. Figure 3.1.7 displays the effects of macroscopic twinning on diffraction patterns. Macroscopic twinning is generally the easiest form to identify, due to low-resolution spots that resemble a figure of eight and circular patterns in the high-resolution region (figure 3.1.7C).

The second possibility is non-merohedral twinning, where two crystal lattices that are related along a shared face but with misaligned lattices form together. This is shown in figure 3.1.8A where the long faces of each lattice are the same but one has been turned 90 degrees along a single axis. Figure 3.1.8C shows how the combination of the two lattices yields high intensity spots at symmetry related positions. Crystals with this defect are typically hard to identify visually, but the persistent high intensity spots are recognisable during data processing.

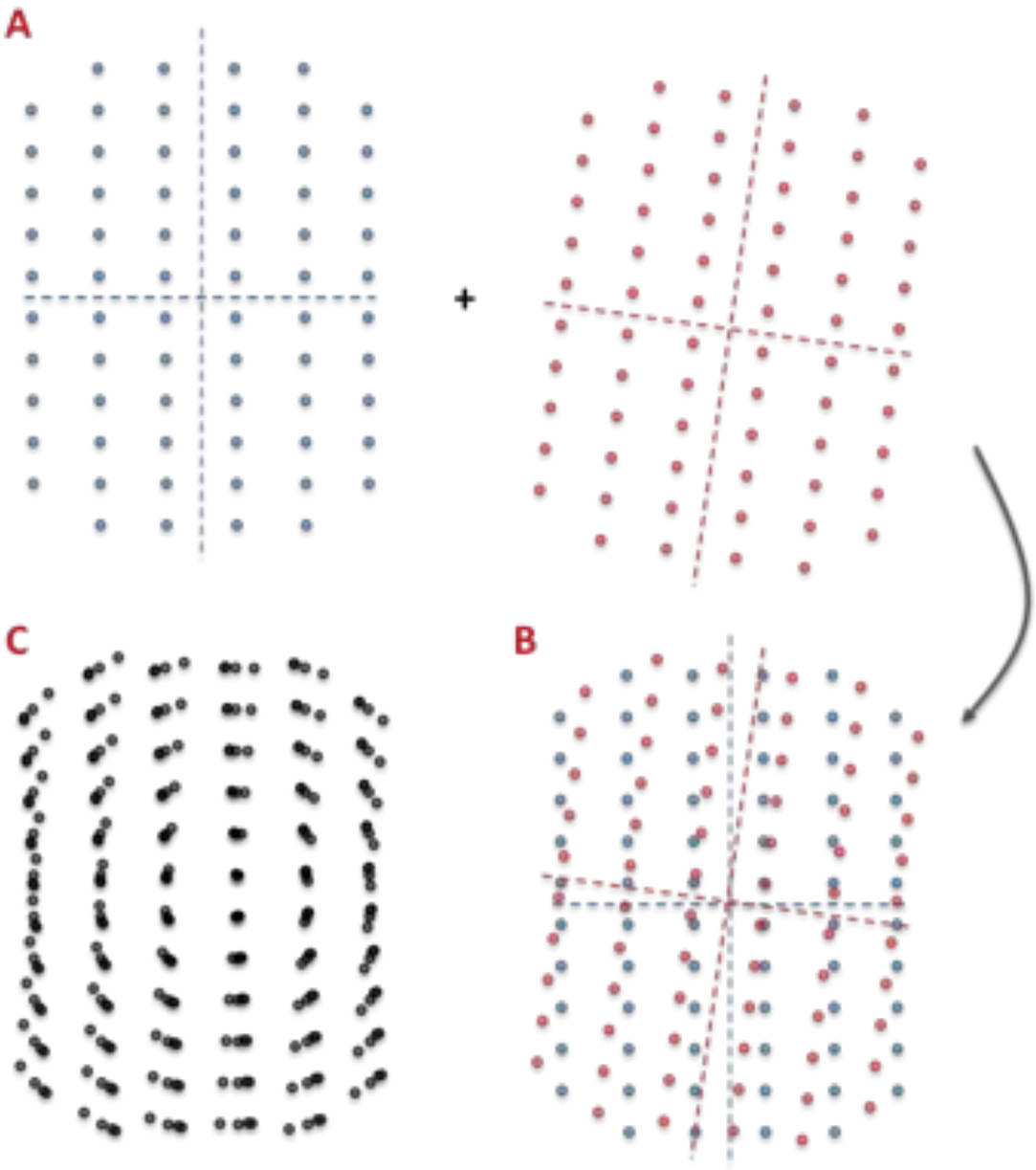


Figure 3.1.7 – The effects of macroscopic twinning and mosaicity on diffraction patterns. Macroscopic twinning occurs when two crystal clusters combine on a plane without a common point of origin, resulting in misaligned lattices. Panel **A** shows independent diffraction patterns for two lattices prior to combining them in panel **B**. In this case the difference between the two lattices is simply a rotation along a central axis, which is easy to identify at the point of collecting the data. In large crystals constructed from multiple smaller domains it is possible to acquire a pattern like that shown in panel **C**, where multiple misaligned lattices have formed. The reflections in this case typically look like a figure of eight and are not suitable for structure solution.

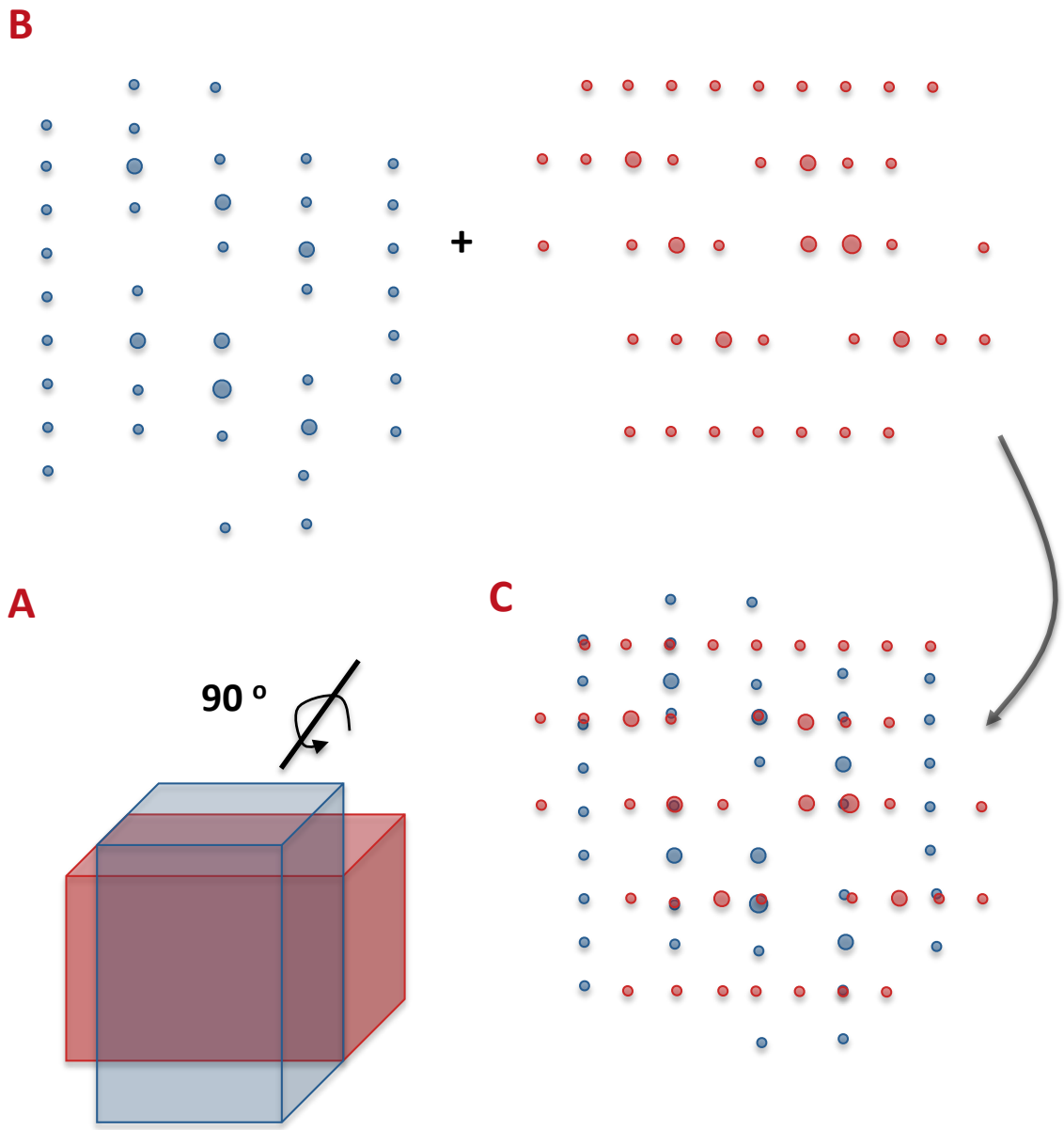


Figure 3.1.8 – Non-merohedral twinning. The two lattices depicted in panel **A** are related along one face but oriented apart by 90°. Panel **B** shows each lattices individual diffraction pattern with panel **C** combining them as if they had both been exposed. In this case processing software will struggle to assign the space group, whilst still taking account of all the strongly diffracting spots. However, occasionally it is possible to isolate a higher intensity set of reflections that correspond to a single lattice from the combined diffraction patterns, allowing non-merohedral datasets to be processed.

The third and most difficult class to identify is the merohedral twin. The previously discussed mosaic issues are predominantly the result of small misalignments, but crystal domains can also grow together in differing orientations (figure 3.1.9). As a consequence of this arrangement the diffraction patterns are indistinguishable from the norm, apart from high intensity spots in regions attributable to symmetry related positions in the lattice.

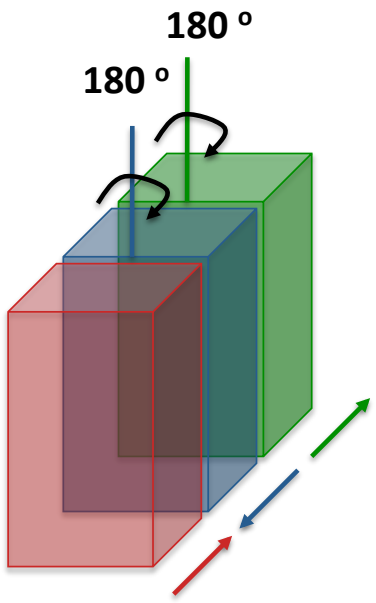


Figure 3.1.9 – Merohedral orientation of crystal lattices within a cluster. The three crystallographic domains above are the same lattice type but oriented along a symmetry axis in opposite directions, without a common related plane.

3.2 – X-ray diffraction experiments

3.2.1 X-ray sources

X-rays are electromagnetic emissions with wavelengths ranging between 0.01 - 10 nm. The emission spectrum of X-rays can be broadly divided into two groups; hard X-rays with low wavelengths in the 0.2 – 0.1 nm, 5+ keV range and soft X-rays with larger wavelengths and lower corresponding energies. Macromolecular experiments typically utilise hard X-rays because the wavelength matches the length between two covalently bonded carbon atoms, allowing for the resolution of individual atoms within a protein.

3.2.2 X-ray Production

X-rays are generated using vacuum tubes, through the bombardment of a metal anode with high-energy electrons directed by an electrical current. The most common form of lab X-ray source is a rotating anode (figure 3.2.1). X-ray emission from a metal anode is typically characterised by a spectrum (figure 3.2.2A) exhibiting multiple peaks. These differing energies of X-ray emerge because the incident electrons can interact with the anode in a variety of ways (figure 3.2.2B). The resulting emission from this variable bombardment can generate both heat and Bremsstrahlung radiation, or it can displace an inner shell electron from the anode. This displacement leads to a higher shell electron filling the gap in the inner shell and emitting X-ray radiation (figure 3.2.2C). Of the released radiation the most prominent peak is the Kα, this is because the energy required to produce Kα radiation is much smaller than Kβ (figure 3.2.2D). Therefore, the emission most commonly selected for diffraction experiments is the high intensity Kα peak.

3.2.4 Laboratory based X-ray diffraction apparatus

The aim of collecting X-ray diffraction data is to produce a set of consistently measured and indexed intensities for as many unique reflections as possible. Figure 3.2.3 displays a typical X-ray diffraction setup grouped into four main categories, X-ray generation in red with focusing in blue, detection in purple and crystal manipulation in green.

3.2.5 Producing monochromatic X-rays with collimators and focusing mirrors

The first challenge for a diffraction experiment is that the X-rays emitted from a copper anode tube cover a spectrum of differing wavelengths. In order to discern structural information from a dataset during processing, a single species of monochromatic X-rays is required. The highest intensity portion of the spectrum is the Kα radiation and the simplest way to separate them is to discriminate by wavelength. Figure 2.2.2A shows the broad range of Kβ and Bremsstrahlung radiation that requires filtering away, the classic method for achieving this is through a crystal monochromator (figure 3.2.4). In most cases a silicon crystal monochromator is employed, which only diffracts the Kα wavelengths. This smaller range of wavelengths is then filtered once again by passage through a precisely oriented collimator, made out of Nickel, to produce a narrow beam of X-rays.

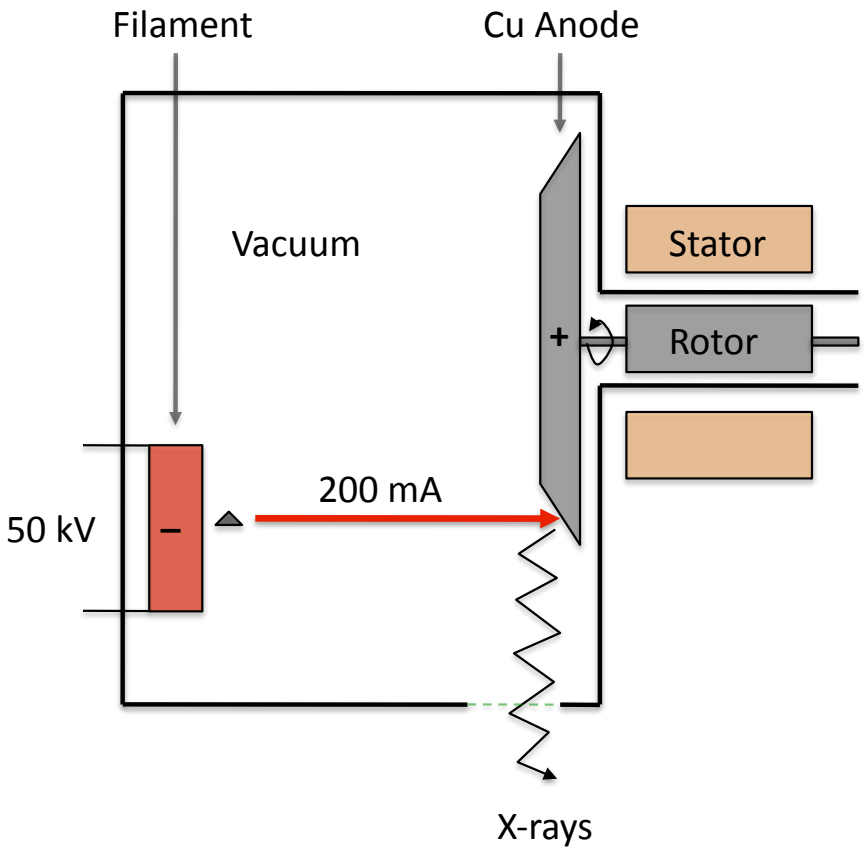


Figure 3.2.1 – Rotating anode X-ray source. Rotating anode vacuum tubes are the most common laboratory X-ray source. These tubes produce X-rays by bombarding an anode with a steady stream of electrons emitted from a cathode; these electrons are accelerated towards the anode by passing a large electrical current between the two electrodes. A magnetic motor rotates the anode, preventing the same spot from being continuously bombarded. Consequently rotating anode sources are capable of generating more flux and higher energy X-rays than fixed anode counterparts; this is because they dissipate the heat generated far more efficiently.

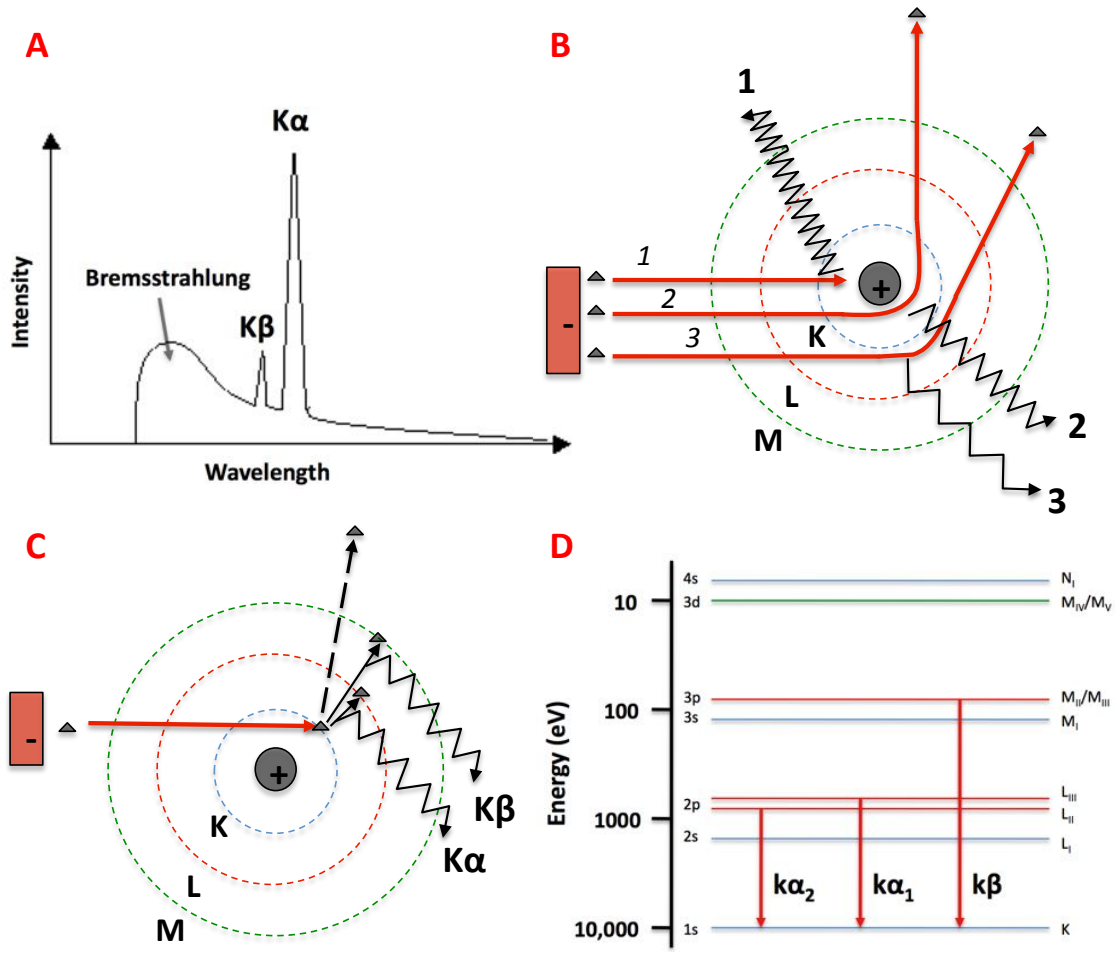


Figure 3.2.2 – X-ray generation at the atomic level and the characteristic emission spectra. Panel **A** displays the typical radiation emission spectra from a copper anode X-ray tube; with its three characteristic peaks. The first from the left is the Bremsstrahlung radiation created when the incident electrons interact with the copper atoms nuclei and are deflected away, as shown in panel **B**. Copper can only lose electrons from its M or L electron shells, with any gap having to be filled from the K shell, shown in panel **C**, leading to the characteristic Kα and Kβ radiation prominent in spectra A. In all X-ray emission spectra the lower energy M shell radiation Kβ is less prevalent than Kα, this is due to the energy requirement to fill this vacancy, illustrated in panel **D**.

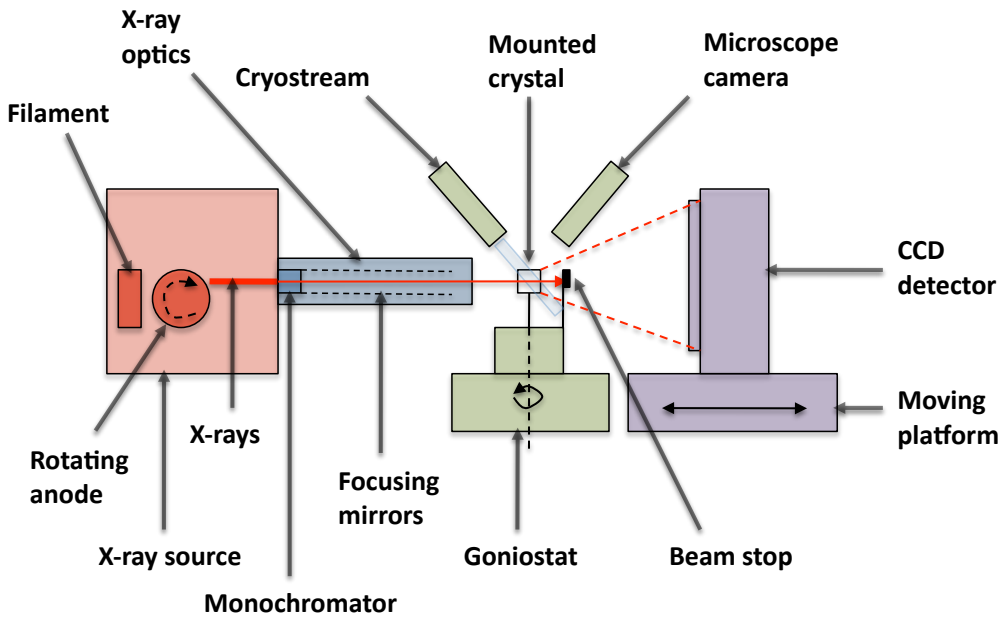


Figure 3.2.3 – Laboratory X-ray diffraction apparatus. An X-ray diffractometer has four key modules. The first is X-ray generation via a rotating anode tube. The optics portion then produces and focuses a monochromatic beam, covering a small enough area that the radiation can be directed at the sample. The third fraction of the equipment manipulates the crystal via a goniostat that orients the crystal and a cryo-stream responsible for maintaining cryogenic conditions. The final stage of the experiment is to detect and record the diffracted X-rays using a CCD detector on a moving platform, with a small lead beam stop preventing the powerful non-scattered incident beam from damaging the sensitive detector.

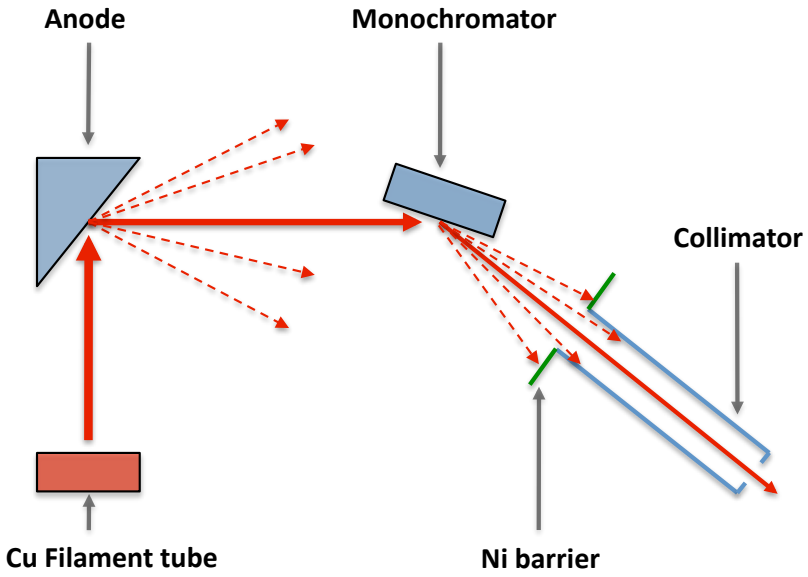


Figure 3.2.4 – Wavelength filtering and beam focusing tools. Panel **A** displays how Kα X-rays suitable for diffraction experiments are filtered from the rest of the emitted radiation, using a crystal monochromator and fine tube collimator.

Therefore, to achieve smaller beam sizes the flux of the initial wide beam is sacrificed, which is why micro-focus beams are only possible when using high flux synchrotron light sources.

3.2.6 Manipulating the crystal sample

In order to determine the structure of a protein it is necessary to collect a full asymmetric unit of reciprocal space, accounting for the unique portion of the diffraction pattern. This requires the crystal be placed in a range of different orientations and rotated within the X-ray beam. The apparatus responsible for this is called a goniostat (figure 3.2.5).

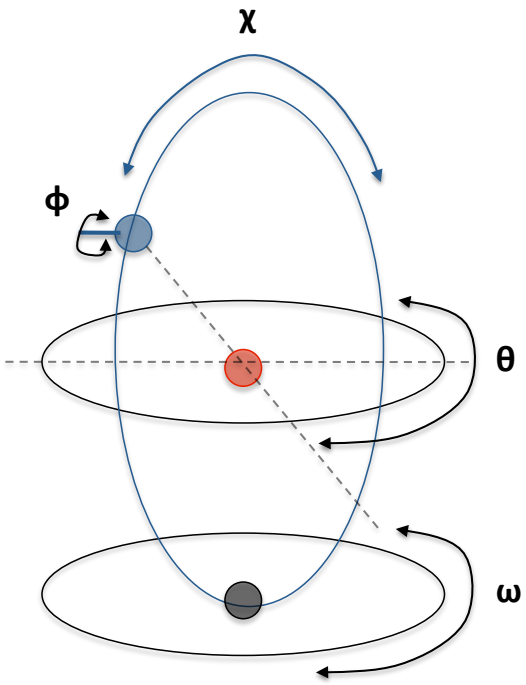


Figure 3.2.5 – Schematic representation of a goniostat. The primary rotation axis on which the data is collected is ϕ, the motor that drives this motion is both accurate and stable to reduce errors. The other rotation axes are responsible for orienting the crystal into the X-ray beam. The bottom axis ω is not routinely required with the θ and χ axes responsible for orienting the crystal.

A convenient method of determining which reflections will be measurable at a given orientation of the crystal is an Ewald construction (figure 3.2.6). This type of construct is useful for determining the Miller indices of the exposed crystal and informs the rotation range required to collect a complete dataset.

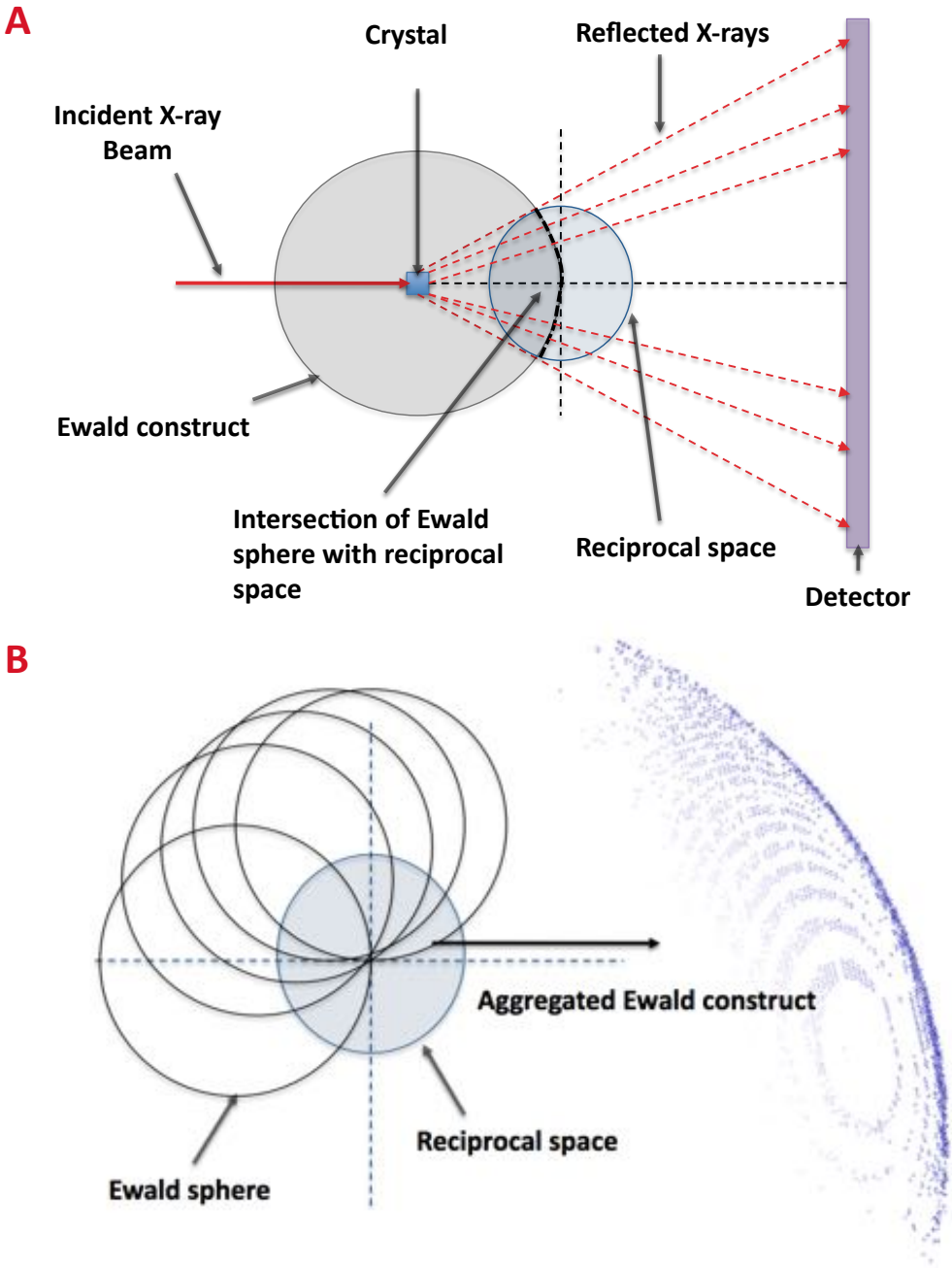


Figure 3.2.6 – Interpreting 2D diffraction patterns using an Ewald sphere construct. An Ewald construct describes the relationship between the angle and path of the scattered X-rays with the unit cell of the crystal, by reconstructing which reflections will be present in a given lattice orientation. Panel **A** shows how using the detector distance as a guide for maximum resolution a 2D diffraction pattern can be used to place the spots into a spherical construct. With panel **B** showing how the Miller indices are examined by integrating multiple images encompassing the whole asymmetric unit into a single construct shown to the right.

3.2.7 Detecting the diffracted X-rays

The most common in house detector type is the charge-coupled device (CCD) (figure 3.2.7). The reflected X-rays interact with a fine phosphorous screen, which generates visible light species that can travel down a fibre optic taper leading directly onto the CCD chip. The CCD then detects these photons through the generation of free electrons in the semi-conductor. This portion of the detector is fully digital resulting in much shorter scanning and storage times than analogue image plate and photographic film technologies.

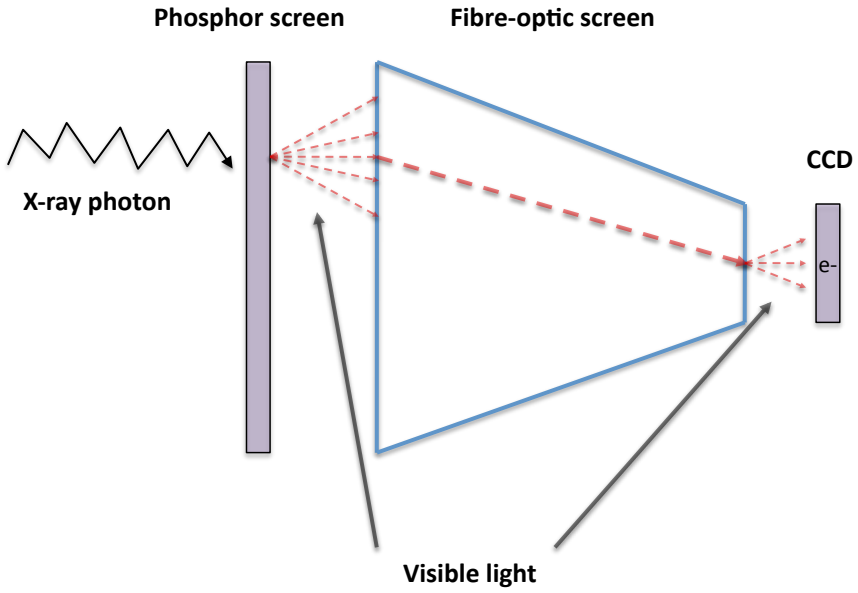


Figure 3.2.7 – Schematic of a charge-coupled device coupled with a semiconductor chip. CCD detectors first capture the diffracted X-rays and then convert them into a format of light that a semi-conductor chip can accurately measure. Diffracted X-rays bombard a phosphorus screen producing visible light, which is passed along a fibre optic taper to the CCD that measures the light and produces a digital image of the diffraction pattern directly from it.

However, Synchrotrons are currently equipped with a newer technology called pixel array detectors (PAD), which incorporate silicon diodes directly bonded to a complimentary semiconductor with the reflected X-rays creating a measurable electrical charge. This removes the need for separate X-ray capturing and detection components, reducing the noise and time required to expose the detector then store the image. Resulting in shutter less operation and reduced noise, which is well suited to phasing and high-throughput diffraction experiments.

3.2.8 Synchrotron light sources

All the data represented in this thesis were collected at the Diamond light source Oxford, UK. A synchrotron is a particle accelerator, with the electrons orbiting the booster and storage rings traveling at relativistic speeds measured by the total energy input in electron volts (eV). The electrons are generated by a cathode filament at 90 KeV and accelerated by a series of magnets to 100 MeV by a linear accelerator. They are then exposed to magnets and centrifugal forces in the booster ring that amplify the energy of the beam to 3 GeV, which is maintained in the storage loop by radio frequency generators. This storage ring is not a perfect circle but instead is constructed from straight sections punctuated by 46 bending magnets to maintain the beams circular orbit. Experimental stations are positioned either alongside the bending magnets or at insertion devices located on the straight segments. Macro-molecular diffraction experiments rely on X-rays generated by insertion devices called undulators, which vary the distance between parallel magnet arrays to select for the desired wavelength of X-ray. Diamond light source is an example of a 3rd generation synchrotron, defined by its constant re-charging of the storage ring with fresh photons (figure 3.2.8).

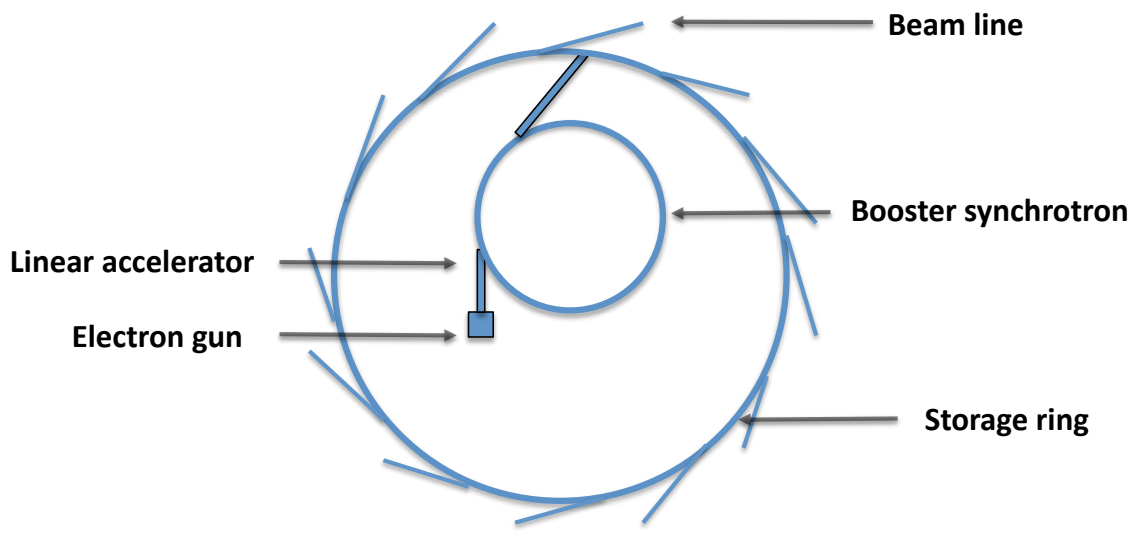


Figure 3.2.8 - Top down view of a synchrotron. Synchrotrons are composed of 4 key portions the electron gun that supplies the photons which are then accelerated and maintained at extremely high energies in the booster and storage rings through the use of powerful magnets. The final stage is the experimental beam line fed by an insertion device.

3.3.1 Wave theory and how X-ray diffraction can yield electron density

The aim of the diffraction experiments is to construct a map of electron density using information derived from the diffraction patterns. This section is going to briefly introduce the key equations and theory involved in moving from 2D spots onto 3D density.

The diffraction patterns collected and their corresponding spots represent complicated cosine / sine waves for each reflected X-ray.

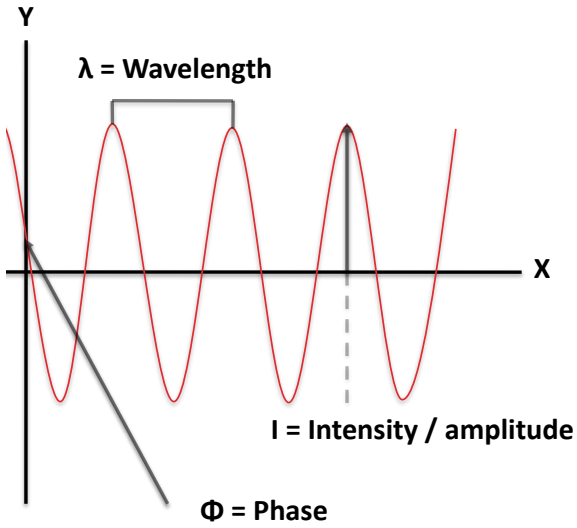


Figure 3.3.1 - A 2D wave can be described using three terms. The phase describes the point on the Y-axis the wave originated from. The amplitude, how far up the Y-axis the peaks and troughs of the wave extend to. The wavelength being a measure of the frequency of the wave.

Two-dimensional waves can also be described via an equation:

$$f(x) = F\ cos2\pi(hx + \alpha)$$

Amplitude

Wave height at given point on X axis

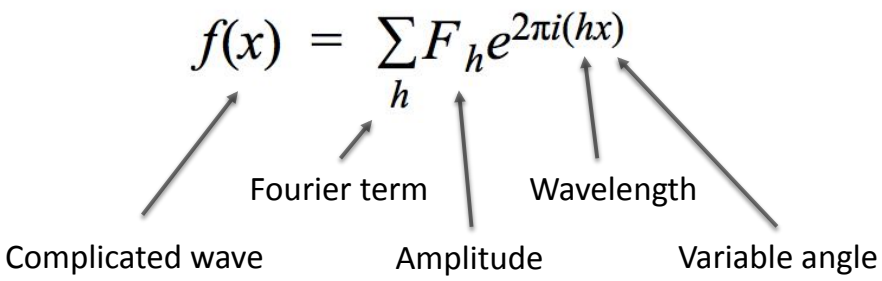
Wavelength

Phase

Variable angle

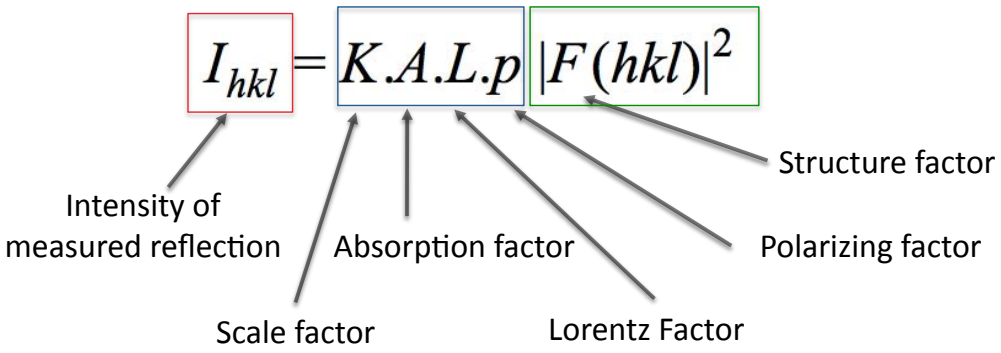
Equation 3.3.1 – Simple 2D wave equation. This equation describes the path of a 2D wave with a given point on the x-axis described by a variable angle along with constant constraints such as the wavelength, amplitude and phase.

A diffracted X-ray can be treated as a complicated waveform, which is described by taking the terms of equation 3.3.1 and applying a Fourier transform. This is done by integrating multiple waves as described by the previous equation into the following equation:

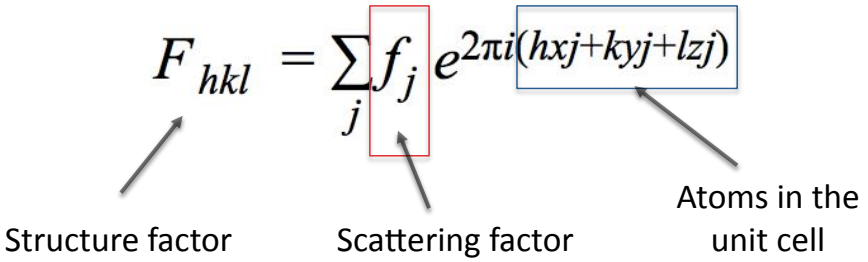


Equation 3.3.2 – Complicated 2D wave equation.

Experimental diffraction data yields information relating to both the crystals unit cell and the measured intensity of the reflected X-rays scattered from a given location in the crystal lattice. This intensity value is directly proportional to the amplitude of a wave and can be described with the following equation:

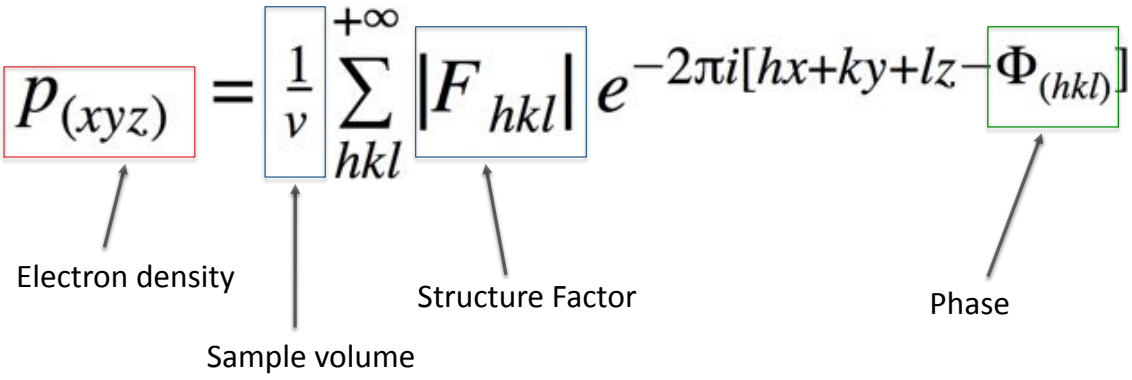


Equation 3.3.3 – Measured intensities equation. This equation relates the collected intensities for each reflection with the eventual structure factor required to derive electron density. The red box represents the intensity measurement produced during the scaling portion of processing. This intensity measurement is a description of the frequency of reflected X-rays measured from a single point in the lattice. The blue box contains all the factors derived empirically from data obtained during the indexing and integration portions of processing, which relate to the unit cell and Miller indices of the crystal. The final green box represents the structure factor for each unique reflection, which is expressed as a complex waveform (equation 3.3.4) and is directly related to the amplitude of the reflected X-rays.



Equation 3.3.4 – Structure factor equation. The structure factor is a complicated function because it accounts for the variable scattering potentials of atoms from different species. This scattering is factored into the above equation in the red box, with f_j corresponding to a scattering value and j contributing the amplitude value to the individual atom responsible. However, this scattering is not just influenced by a single atom at a time, but by all the atoms present within the unit cell. Therefore, the location of atom j is expressed through a coordinate system in the blue box in both real and reciprocal space.

The end result of a diffraction experiment is the production of an electron density map. In order to produce this map there are three prerequisite experimental values, the unit cell dimensions and for each reflected X-ray a structure factor and phase measurement.



Equation 3.3.5 – Electron density equation. The red box is the end result of an electron density calculation given in $Å^3$ for a given 3D point in real space, as defined by X, Y and Z axes. However, what this is describing in real terms is the frequency of reflected X-rays, which is proportional to the electrons present at a given location. The blue boxes contain components that can be experimentally derived from standard diffraction patterns through data processing (section 3.3.2). The green box holds the phase component, which is not determined from the diffraction data and has to be derived from separate experiments (section 3.3.4).

3.3.2 Processing diffraction patterns into usable intensity values

The purpose of processing is to uncover the structure factor for each spot on a diffraction pattern. This structure factor can be calculated directly from the measured intensities (equation 3.3.3). However, the only data available to elucidate the intensity of the reflected rays from are; the collected diffraction images, the distance parameters for the detector and goniostat, along with the incident X-rays wavelength. There are three stages when processing data from X-ray crystallography experiments, indexing, integration and scaling (figure 3.3.2). The measured intensity value for each diffraction spot is calculated during integration. The other components of the structure factor equation such as the scale, absorption, Lorenz and polarizing factors are all calculated during the scaling portion of processing, using unit cell information determined during the first stage indexing.

3.3.3 Indexing

Indexing is the process of taking predetermined goniostat and detector parameters and applying them to the observed diffraction spots. It is used to determine the unit cell dimensions, Miller indices and crystal space group. Indexing does this by taking each diffraction image and identifying the spots, before assigning them their Miller indices.

3.3.5 Integration

Integration is where the intensity of each identified reflection is calculated. It takes the indexed spot locations and Miller indices along with how many degrees of rotation each reflection persists for, termed the mosaic spread, and assigns an intensity value. This value is calculated by counting the number of pixels on the detector influenced by the reflection and is proportional to the intensity of scattered X-rays. The measured Intensity is related to the atomic arrangement and number of unit cells present in the crystal, therefore larger crystals tend to diffract more strongly. The primary challenge is that intensity calculations have to account for any background radiation or noise surrounding the diffraction spots. As a result reflections with low intensity are hard to identify from the background, necessitating the prediction of spots from unit cell and space group parameters.

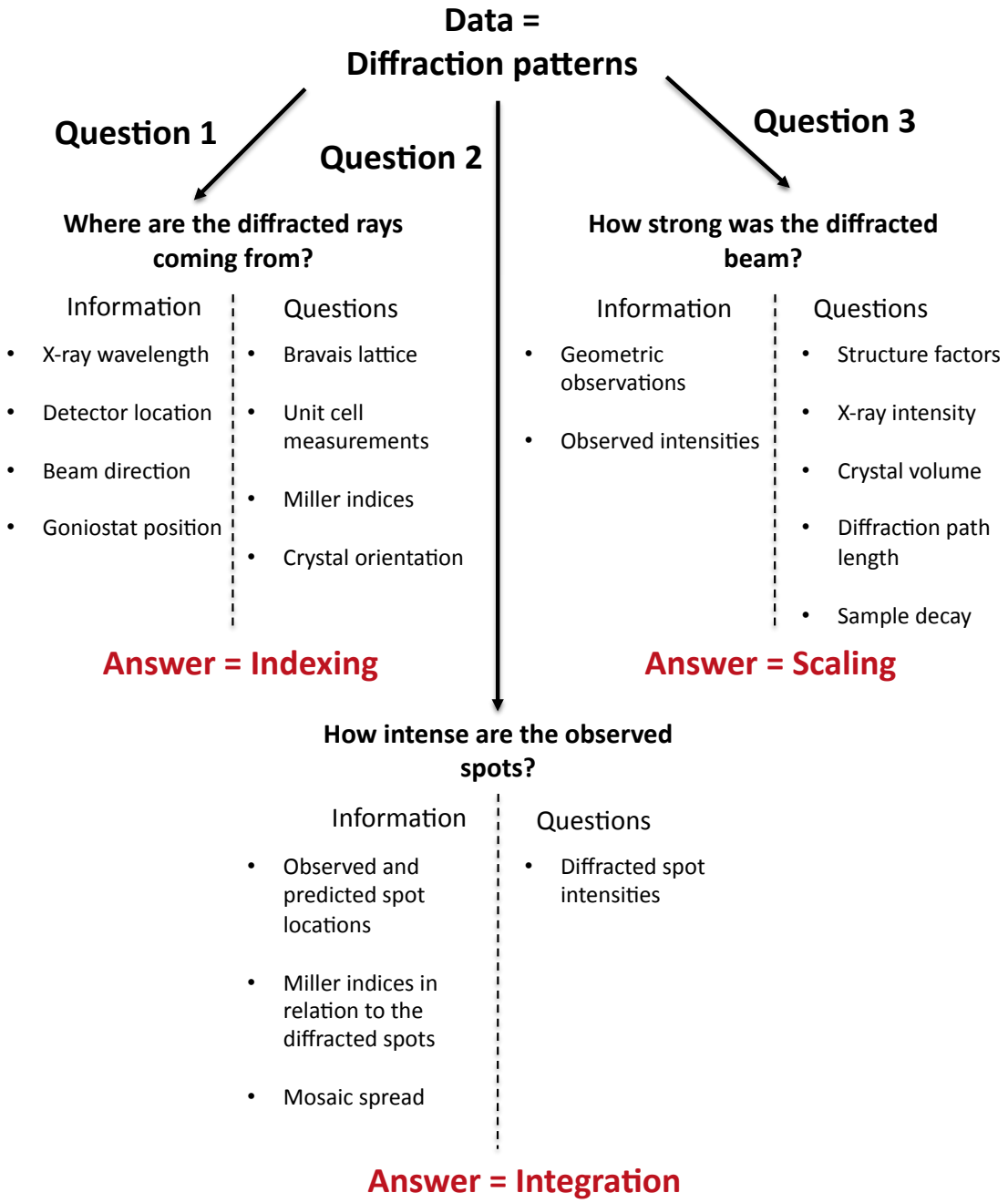


Figure 3.3.2 – Flow diagram detailing information available during data processing. The diagram above details the information available at each stage of processing. Indexing provides the unit cell dimensions and Miller indices of the crystal. Integration quantifies intensity values based on the size and darkness of diffraction spots on a pattern, in relation to how many degrees they persist and how often they repeat in reciprocal space. Scaling then models empirically unknown values related to the unit cell dimensions and Miller indices. When the outputs of scaling and integration are combined it is possible to calculate a structure factor.

3.3.6 Scaling

Scaling takes the measured intensity and combines it with knowledge of the unit cell and its symmetry operators to calculate a structure factor for each reflection. The structure factor equation (equation 3.3.3) incorporates a variety of correction factors, empirically determined and taken into account along with a unified scale value for the whole dataset. The absorption factor models the reflecting X-rays passage through the crystal. The Lorentz factor corrects the angular velocities of the reflected X-rays from an Ewald construct (section 3.2.6). The polarizing factor corrects for the polarizing effect of the incident X-ray beam and is a measure of how intense the initial X-ray was. All of the above values are modeling using data brought forward from the indexing and integration stages, refined against the scaled intensities. This refinement aims to normalize these values until they return intensities for symmetry related spots that are the same.

3.3.7 Quality control

The most convenient opportunity to audit the quality of the data is directly after scaling, where a standard set of quality control indicators are calculated. These relate to one of three processing functions, the merging of the indexed reflections, signal to noise and completeness.

Merging statistics are prominent when assessing data quality because many of the reflections are measured multiple times. The first figure of interest is the number of unique reflections in comparison to the total number of reflections, which offers a gauge of data redundancy. The simplest representation of this is the linear merging R-value or R_{merge}:

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

Average Intensity for each reflection

Reflection

Redundant observations

Equation 3.3.6 – Merging statistic R_{merge}. The R_{merge} relates the number of redundant observations (N) of a single reflection (h), to the measured intensity of the reflection and the average intensity (ī).

R_{merge} values tend to rise along with resolution, as high-resolution spots normally diffract weakly introducing larger errors. However, the R_{merge} does not take into account the inherent benefits of redundancy, with each individual reflections contribution a function of the redundancy. As a result large datasets, such as those from phasing experiments, exhibit larger R_{merge} values than those from smaller native datasets.

To correct this effect a more accurate representation of merging quality is R_{pim}, which incorporates a precision factor that acknowledges that Intensity measurements are more accurate when more observations have been taken.

$$R_{pim} = \frac{\sum_h \left(\frac{1}{N-1}\right)^{1/2} \sum_{i=1}^{N} |I_{(ih)} - \bar{I}_{(h)}|}{\sum_h \sum_{i=1}^{N} I_{(ih)}}$$

Additional precision indicating term

Equation 3.3.7 – Merging statistic R_{pim}. The terms are the same as for the linear merging value, but an additional $(1/N-1)^{1/2}$ term factors redundancy of the same scattering component in the reciprocal lattice into the calculation. As a result Rpim declines when there is a higher degree of redundancy.

The second quality determinant, signal to noise, is expressed by the term I/σ(I):

$$I/\sigma(I) = \frac{1}{N} \sum_h^N \frac{|I(h)|}{|\sigma I(h)|}$$

Equation 3.3.8 – Signal to noise. The importance of this value differs depending on the intended purpose of the data. A typical high-resolution dataset for model building involves software that weights the data so signal to noise is not a crucial quality determinant. However, phasing datasets are dependant on strong signal to noise; as the differences being measured are small values between much larger numbers.

The final quality determinant is the completeness statistic. Completeness equates the total unique reflections that could be observed at a given resolution range to the anticipated value.

3.3.8 Constructing the electron density map

Unit cell information and structure factors are the most prominent values derived from a diffraction experiment, however they do not encompass all the terms that are required to solve the electron density equation (equation 3.3.5). The remaining unknown term is the phase of the reflected X-rays and phase information cannot be recovered from native diffraction patterns. This is important, as the phase is not a trivial value. It contributes more to the appearance of an electron density map than the measured amplitudes (figure 3.3.3).

Figure 3.3.3 – Relative importance of phase and amplitude values in image reconstruction. Panel **A** and **B** are representative of Bill Clinton and Hillary Clinton phases and amplitudes respectively. Panel **C** however is a combination of Bill's amplitudes with Hillary's phases. The predominant image taken away from panel **C** is that of Hillary, this is because the phases of the reflected X-ray beams provide more information in the electron density equation than structure factors. (Diagram modified from a lecture series presented by Lindsay Sawyer)

3.4 Phasing experiments

The phase of a wave shares no formal relationship with its amplitude. Therefore, all phasing experiments involve measuring the difference between native reflections and a subset of reflections behaving differently. There are three mainstream methods for estimating the phase, which are covered in historical order.

3.4.1 Isomorphous replacement

Isomorphous replacement (IR) was the earliest method of dealing with the phase problem (Perutz, 1954). IR relies upon incorporation of a heavy metal compound within the crystal lattice through either displacement of, or interaction with an amino acid side chain. The technique is extremely flexible as any native crystal can be soaked and incorporation of heavy metals is not uncommon. However, there are considerable issues with crystal isomorphism, because the crystals space-group and unit cell dimensions have to be retained to accurately differentiate between the native and derivative samples. Heavy atoms are used because they contain significantly more electrons, which contribute more to the diffraction pattern and are easily identified and measured. It was estimated by Watson Crick in 1956 that 1 incorporated atom of Mercury in a protein of 1000 atoms would elicit a 25% change in measured intensity. As a result most proteins only require a single heavy atom to bind consistently within the asymmetric unit to phase the entire molecule.

There are two IR formats employed to solve the phase problem. The easiest to prepare for is single Isomorphous replacement (SIR), where a single heavy atom type is incorporated (figure 3.4.1). However, an alternative to SIR is multiple Isomorphous replacement (MIR), which employs multiple independent heavy atom derivatives to simplify phase calculation (figure 3.4.3).

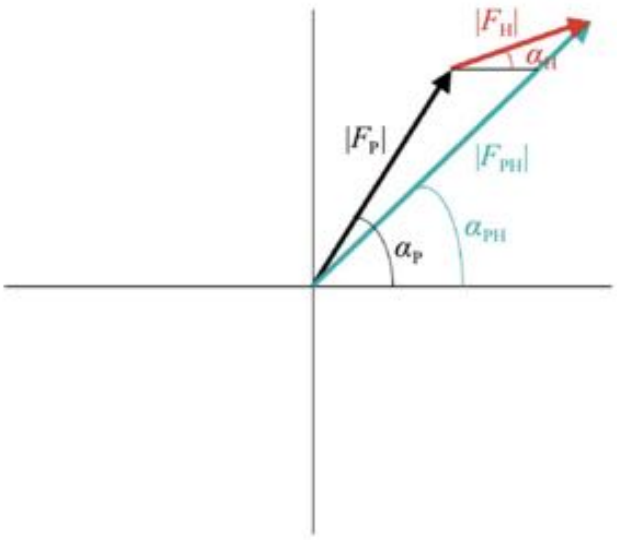


Figure 3.4.1 - Argand plot from a typical SIR experiment. Heavy atoms exhibit different intensity values than native reflections. Black and teal lines represent structure factors calculated from these intensities for the native and derivative reflections respectively. The isomorphous difference between the two structure factors is represented by the red $|F_h|$ term. Using the isomorphous difference it is possible to determine a phase difference termed α_h, accounting for the heavy atom. $|F_h|$ is then used to determine the heavy atom(s) location in the reciprocal lattice and in turn real space. This XYZ interpretation permits reverse calculation of the derivative phase, through refinement of the heavy atom(s) location in real space. This derivative phase is then used to approximate the native phase through application of the cosine rule described in equation 3.4.1. (Taylor, 2010).

$$\alpha_p = \alpha_h \pm cos^{-1}\left[\frac{(F^2_{ph} - F^2_p - F^2_h)}{2F_pF_h}\right]$$

Equation 3.4.1 – IR cosine rule. This equation relates the calculated derivative phase to the distance between measured amplitudes from the native dataset.

Unfortunately, when phasing with a single derivative the cosine term provides two disparate solutions for the native phase leading to a high level of ambiguity.

MIR attempts to break this phase selection ambiguity, by manipulating Harker constructs. If two independent derivatives are present each with unique amplitude values then the resulting circular constructs will likely only intersect at one location, corresponding to the correct phase (figure 3.4.3).

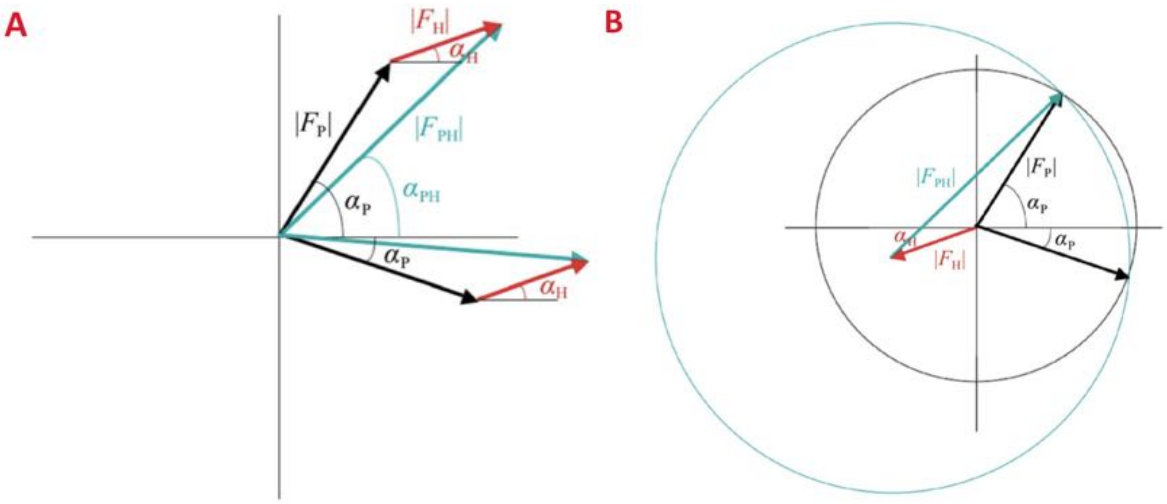


Figure 3.4.2 – Argand diagram and Harker construct for SIR phase determination. Panel **A** displays SIR via an Argand plot, there are two possible native phase (α_p) values related by the derivative phase (α_H). This ambiguity is better represented in a Harker construction, Panel **B**. Where the calculated Isomorphous difference $|F_h|$ acts as the origin of the $|F_{ph}|$ term, with the $|F_p|$ term expressed as on Argand plot. The α_p is calculated via two circles drawn with the origin of the $|F_p|$ and $|F_{ph}|$ lines at the centre and the amplitude value dictating the radius. The points where the two circles intersect indicate potential α_p candidates. (Taylor, 2010).

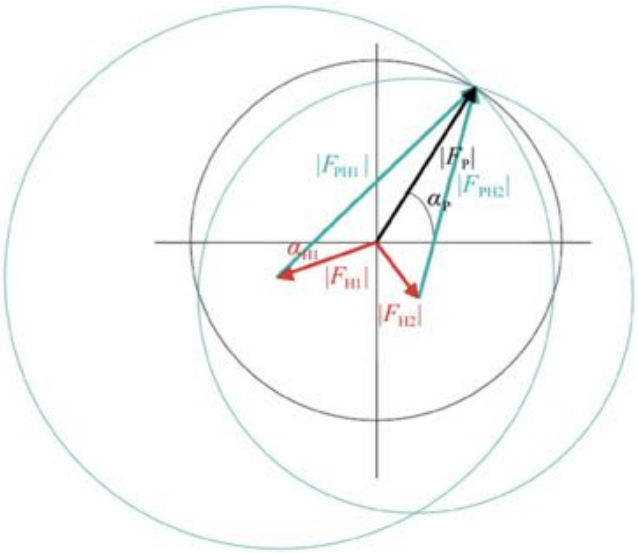


Figure 3.4.3 – Harker construct for MIR phase determination. SIR does not provide a definitive phase value. However, if data is collected for more than one derivative then several Isomorphous differences can be measured. In this example an additional derivative $|F_{ph2}|$ is included and where the two green derivative circles intersect the black native amplitude plot is the point from which to calculate the native phase. (Taylor, 2010).

3.4.3 Anomalous scattering

Structure factors (equation 3.3.4) incorporate an atomic scattering factor comprising three components. The first is a normal scattering term f_0 measuring an atoms scattering potential. The second and third terms f' and f'' are wavelength dependent, accounting for scattering anomalies that occur at the atoms absorption edge. These anomalies occur because at an atoms absorption edge the incident X-rays do not provide the energy required to promote an inner shell electron. Therefore, anomalous scattering is the only occasion where Friedel's law breaks down. Friedel's law ($F_{hkl} = F_{-h-k-l}$) states that within a unit cell equivalent atoms from each asymmetric unit should scatter in the same way. If two asymmetric units aren't diffracting in an identical fashion as a result of a small number of target atoms, it is possible to measure where this anomalous signal is originating from in the reciprocal lattice. The difference between two measured intensities, or Bijvoet pairs, for a given point is termed the Bijvoet difference (figure 3.4.4A). This difference value is used in an identical fashion to the isomorphous difference encountered in IR experiments (Hendrickson, 1979). The Bijvoet difference can also be combined with SIR to offer a second value with which to determine the correct native phase, in a similar fashion to MIR, called SIRAS.

Multiple wavelength anomalous dispersion (MAD) is the most widely used experimental phasing technique (Hendrickson, 1985). It offers several benefits over Isomorphous replacement, as it does not require multiple crystals for native and derivative data collection. Instead crystals are grown from a protein derivative, which incorporates Selenium or an alternative heavy atom, in the place of the Sulphur atoms present in Methionine residues. The major advantage is that both the native and anomalous data can be collected from a single crystal, alleviating non-isomorphism. MAD experiments work by exposing the sample to three differing wavelengths of X-ray. The wavelengths chosen correspond to the derivative atoms absorption edge. The first two wavelengths determine the anomalous scattering and are collected at the absorption peaks calculated for both f' and f", the final data set is at a remote wavelength designed to act as a native set of reflections providing a stark contrast to the anomalous data. Collecting the data at multiple wavelengths allows calculation of the phase using a Harker construction (figure 3.4.5).

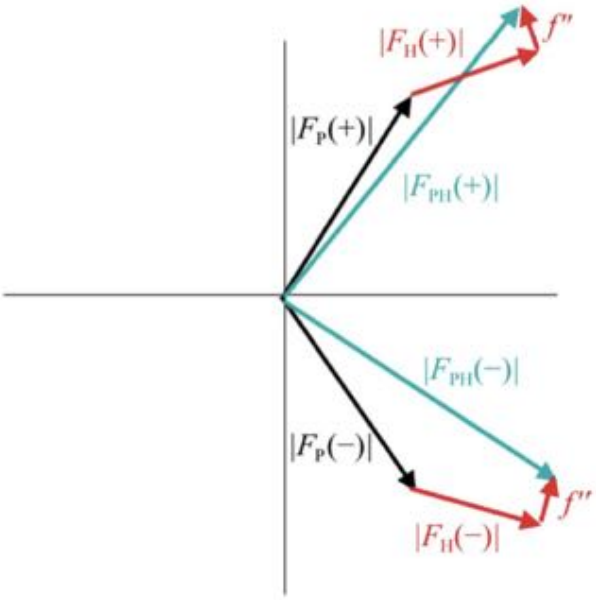


Figure 3.4.4 – Argand plot for calculating anomalous scattering factors f' and f". An Argand plot is used to show how the anomalous scattering factor *f"* is related to the anomalous differences measured when Friedel's law breaks down. The measured native and anomalous amplitudes are termed by $|F_p|$ and $|F_{ph}|$ respectively. However with Friedel's law broken there is a gap between $|F_{ph}|$ and $|F_h|$ which is accounted for by the anomalous scattering factor *f"*. (Taylor, 2010).

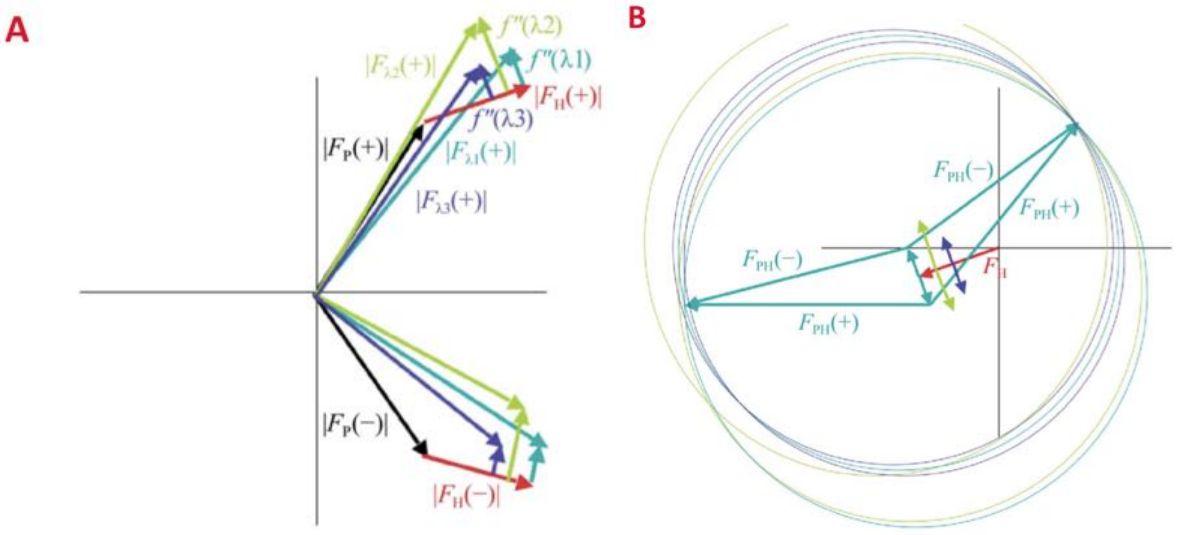


Figure 3.4.5 – Multiple wavelength anomalous dispersion (MAD) phase determination. Panel **A** displays each wavelength of X-ray, yielding different structure factor and f" values. With panel **B** showing the same values but plotted on a Harker construct. The point where the native, f' and f" peak structure factors come together can be used to plot a phase in relation to the native structure factor on its own. (Taylor, 2010).

The second and increasingly common format is single wavelength anomalous dispersion (SAD). A SAD experiment involves collecting an entire dataset on the *f"* absorption peak of the anomalous atom, usually Selenium but with modern PAD detectors increasingly Sulphur. Because SAD experiments do not vary the wavelength of the incident X-rays it can only measure a single Bijvoet difference. This Bijvoet difference is substituted for a |Fh| value and the experiment follows in a similar fashion to SIR (figure 3.4.6). However, just as with SIR there is a phase ambiguity to deal with. As a result phases calculated using SAD often require improvement and the technique is commonly combined with density modification (section 3.4.5). SAD experiments are especially challenging because the Bijvoet difference is only 1-2 % different from the native intensity. With most datasets displaying an R_{merge} statistic significantly larger than the Bijvoet difference, anomalous scattering it is often interpreted as an error. Therefore, highly redundant data is required to provide statistically significant anomalous values.

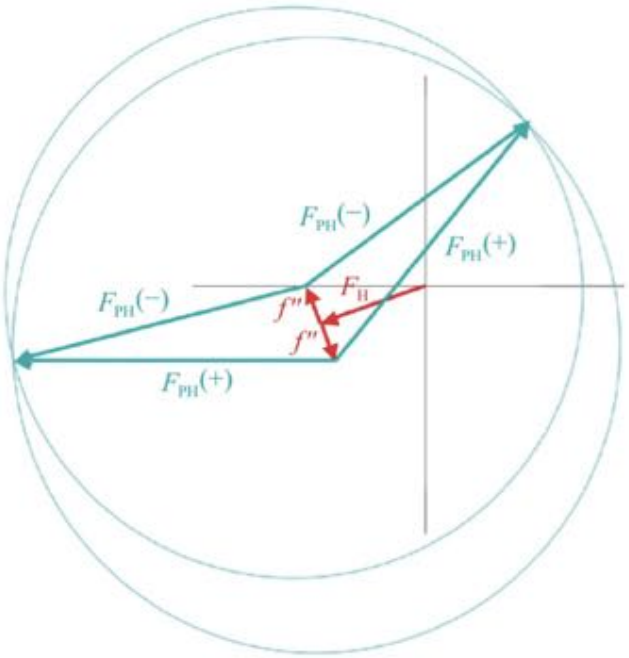


Figure 3.4.6 – SAD phase derivation and comparison with SIR. SAD experiments only provide a measurement of a single Bijvoet difference. When the $|F_{ph}|$ and $|F_h|$ values are plotted on a Harker construct as a symmetrical function of the *f"* there are two possible phase values. This is a very similar situation to the phase ambiguity encountered in SIR, but in SAD experiments the two potential phase regions are spread much further apart. (Taylor, 2010).

3.4.4 Molecular replacement

Molecular replacement (MR) solves the phase problem by estimating a set of phases obtained from a homologous model (Rossmann and Blow, 1962). It requires a complete data set and a replacement model, homologous to the molecule present in the unit cell. In most cases a model with 60% sequence identity and 2-3 Å RMSD between the model and molecules protein backbone is sufficient. With increasingly sophisticated software it is possible to modify a model with a closely predicted 3^{o} fold to exhibit the 1^{o} sequence of the molecule of interest. However, large insertions or deviations in the input model require deleting, as correctness is more important than completeness.

Before its phases can be applied to the experimental data the model needs to be reoriented to match the target molecule in the unit cell. The rotational and translational position of the target molecule is determined by comparing Patterson functions. A Patterson function is a vector map that represents the atomic distances associated with a crystallised molecule, calculated independently of the phases by squaring the calculated structure factors. As a result heavier atoms with larger scattering potential will lead to stronger peaks. If there are several atoms that are expected to scatter more prominently that others it is possible to calculate a vector between them, which are then used to fingerprint the protein. All proteins can be assigned an individual Patterson fingerprint and moreover, proteins with similar structures will bare a resemblance to one another. Therefore, MR provides phase information by reorienting a model into the provided data.

There are two groups of Patterson vectors; smaller distances are assumed to be intra-molecular vectors between two atoms within the asymmetric unit termed self-vectors, with larger distances typically accounting for inter-molecular distances termed cross-vectors. Self-vectors are used to determine the rotation component of the orientation, with cross-vectors utilised later to translate the model within the experimental unit cell. Rotation and translation movements of the model require the calculation of self-Patterson vectors. The unknown protein termed F_{obs} is derived from the measured intensities and the search model F_{calc} is backwards calculated from the model by reconstituting its electron density and estimating its structure factors. The resulting vector maps contain both self and cross vectors, for rotation the search model is limited to just self-vectors through careful space-group selection. It is then compared with the experimental vector map through small angle rotations along the origin. Once the input model has been rotated it is then placed into the unit cell of the experimental data and new F_{calcs} are produced. However, the translation required to complete re-orientation

is best attempted using cross-vectors. Therefore the self-vectors identified from the previous step are subtracted from both maps. The model is then translated with the highest scoring position applied to the rotated model. This process of rotation and translation is summarised in figure 3.4.7.

The newly rotated and translated search model should provide a set of co-ordinates similar to the target protein. The phases are then calculated using a Fourier Transform and combined with the experimental structure factors to produce an electron density map. This means that the model can introduce significant bias to the map. Therefore, a molecular replacement model that inspires confidence should always incorporate information that the input model did not provide.

3.4.5 – Phase / Density improvement techniques

Phasing methods such as SAD and SIR produce an inaccurate phase value, but even phase values far from correct can be taken to produce crude electron density maps. These maps are often of sufficient quality to identify a solvent barrier between the protein molecules, but inadequate to attempt building an initial model. However the inaccurate phases leading to this poor density information can be improved upon, without having to resort to inaccurate model building. The process of improving the phases, by manipulating the experimental data is termed density modification. There are two commonly used density modification techniques termed solvent flattening and histogram matching.

Solvent flattening involves defining where the globular protein component of the asymmetric unit starts and only incorporating reflections from this portion of the data when calculating the phase. This is possible because the density of scattering atoms is much higher in the protein portion of a crystal, with an average electron density of 0.43 e^{-} $Å^{-3}$. This density value is easily separated from the solvent portion with an average electron density of 0.33 e^{-} $Å^{-3}$. All reflections contributed by the 40-50% solvent region are then discounted, resulting in a lower signal to noise ratio for calculating the phase.

On the other hand, histogram matching involves adapting the broad electron density peaks typically seen in poorly phased maps for a sharper peak in the same position. The exchanged peak more closely resembles what would be observed with correct phases. This method however is heavily reliant on the accuracy of the measured amplitudes being correct, as it is possible to introduce substantial bias.

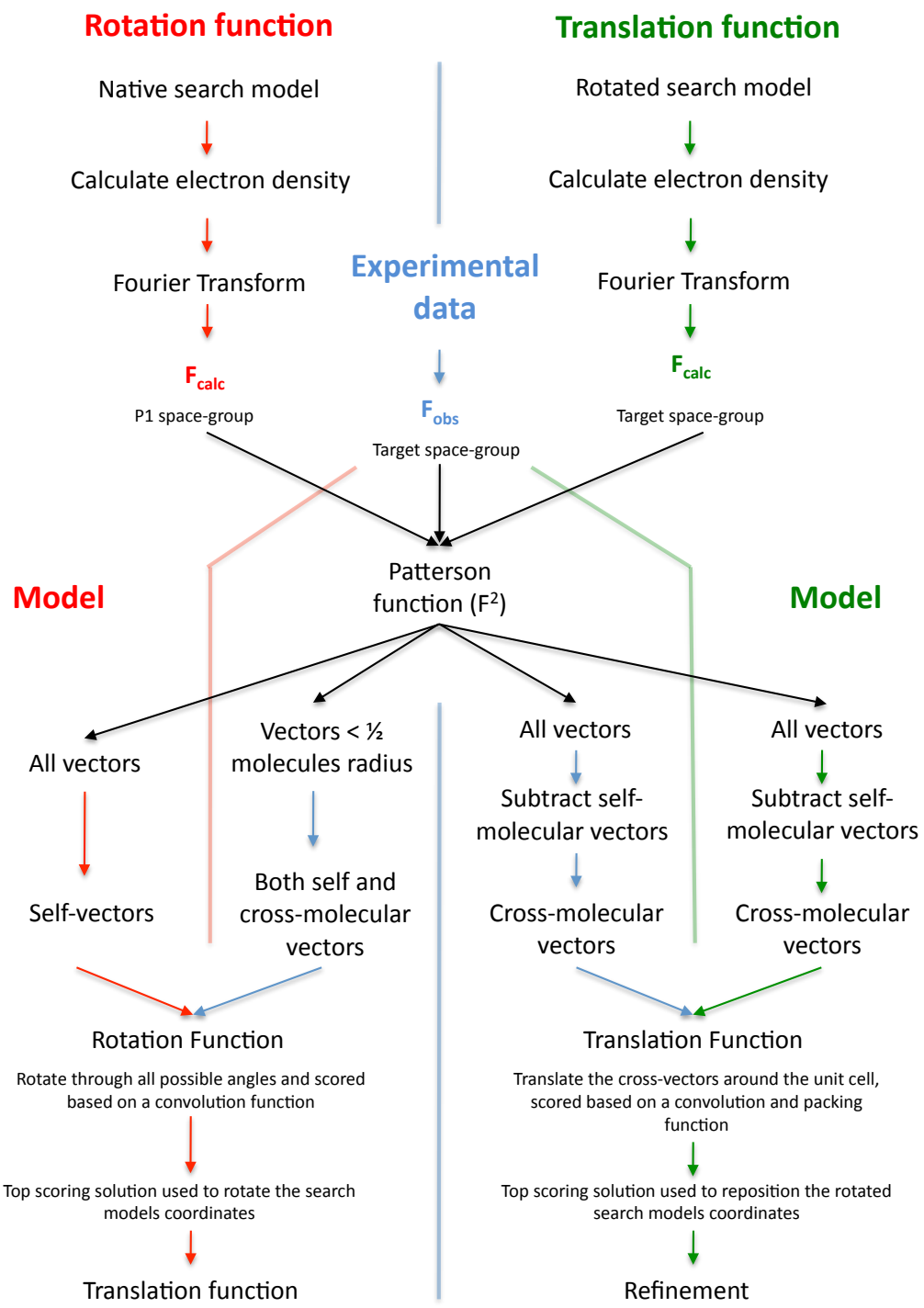


Figure 3.4.7 – Flow chart examining Molecular replacement. Molecular replacement fits a homologous search model into the experimental molecules unit cell. This is achieved by creating a Paterson function directly related to the experimental and calculated structure factors for the data and the search model respectively. A Patterson function allows fingerprinting of the molecule; the search models fingerprint is then matched first through rotation and then through translation functions with the target molecule.

However, histogram matching also inhibits negative density, which is important when model building so should be employed conservatively.

If the techniques above do not yield more accurate phases then non-crystallographic symmetry (NCS) between the intra-protein and inter-protein components of the asymmetric unit, can also be used. For example, if there is strong density assignment for one portion of the molecule, it is possible to average or transplant this density into a related region with poor density. In asymmetric units composed of proteins that exhibit a large proportion of intra-molecular symmetry this can be extremely powerful for improving electron density maps.

3.5 Model building and refinement

With a correctly phased set of reflections, the last requirement for constructing an electron density map is satisfied. The final stage of protein structure solution is building a model that matches the map whilst also conforming to chemical assumptions such as bond lengths, angles and steric clashes. Building a model is also a valid method for recalculating the phase of the measured reflections, producing a better map from which a more accurate model can be built. This process of iterative model building and phase recalculation is termed refinement. Thus the aim of refinement is to improve the phases, through increasing the level of agreement between the measured and calculated structure factors (Winn *et al.,* 2011).

The model can be improved by adjusting several parameters; the x, y and z location of the atoms termed co-ordinates, along with the B-factor of the atoms and the occupancy of selected positions. The B-factor ($Å^2$) is a measure of thermal vibration that accounts for the movement of atoms at each given position. Model accuracy is then quantified by comparing the structure factors calculated from the diffraction experiment (F_{obs}) with calculated structure factors derived from the model (F_{calc}), the aim is to minimise the difference. This comparison is expressed as an R_{factor}:

$$R_{factor} = \frac{\sum_{hkl}(|F_{obs}| - |F_{calc}|)}{\sum_{hkl}|F_{obs}|}$$

Equation 3.5.1 – Measuring model completeness via an R_{factor}.

During refinement it is possible to over build a model incorporating features that improve the overall R_{factor} without significant density evidence to justify their inclusion. False minima are monitored by separating out a small subset of the measured reflections (~5%) termed the R_{free}, which are not refined against. This subset is then available to compare with the refined data. If the R_{free} is significantly higher than the refined R_{factor} it indicates that modifications made to the model are not in line with the information provided by the data.

3.6 Model validation and deposition

Before the finished model can be deposited in the PDB it needs to be checked thoroughly for errors. There are two areas of the model that need to be considered; the first is the conformation of the model, the second involves the interpretation of the electron density map. Structural considerations include Ramachandran plots to check that the bond angles (*phi* and *psi)* for the peptide backbone are within allowed regions and do not clash. Another important consideration is the molecules general geometry with both bond angles and lengths taken into account (Chen *et al.,* 2010). Finally, any portions of the protein that are modeled for completeness sake, but without sufficient density evidence need to be either truncated or have their B-factors adjusted to 200 to indicate that the model at that point is unclear.

Chapter 4: Materials and methods

This chapter is going to explore the general methodologies that are encountered routinely during the production of recombinant proteins. Specific purification and over-expression conditions will be expanded on in a case-by-case basis. Instead this portion of the thesis summarises the techniques involved. Figure 4.1.1 details the stages involved in taking a hypothetical protein from the bioinformatics stage through to crystallisation and functional characterisation.

All the experiments described throughout this thesis were undertaken using equipment and materials common to most laboratories. The chemicals used were sourced predominantly from either Sigma-Aldrich or Fisher and unless specified were 99.9% pure analytical grade materials.

4.1: Production of genetic constructs

This section describes the standard methods used to produce genetic constructs suitable for the production of recombinant proteins.

4.1.1 – Plasmid vectors

X-ray crystallography requires large volumes of high concentration, homogenous protein. In the past it was common to purify proteins directly from their native organism. However these organisms had to be easy to culture, whilst exhibiting abundant expression levels of the target protein. Nevertheless this approach limits the crystallographer to a very small subset of a given organism's proteome and rules out targets from pathogenic microorganisms. Therefore, It is more common to produce proteins for crystallisation experiments via a recombinant expression system.

All the protein samples detailed in this thesis were produced using a bacterial *E.coli* expression system. Where the host cells are adapted to contain a foreign gene, through introduction of a genetically altered plasmid. These altered plasmids are termed constructs throughout this thesis. The pET system of expression plasmids were used exclusively in the construction of modified vectors (figure 4.1.2).

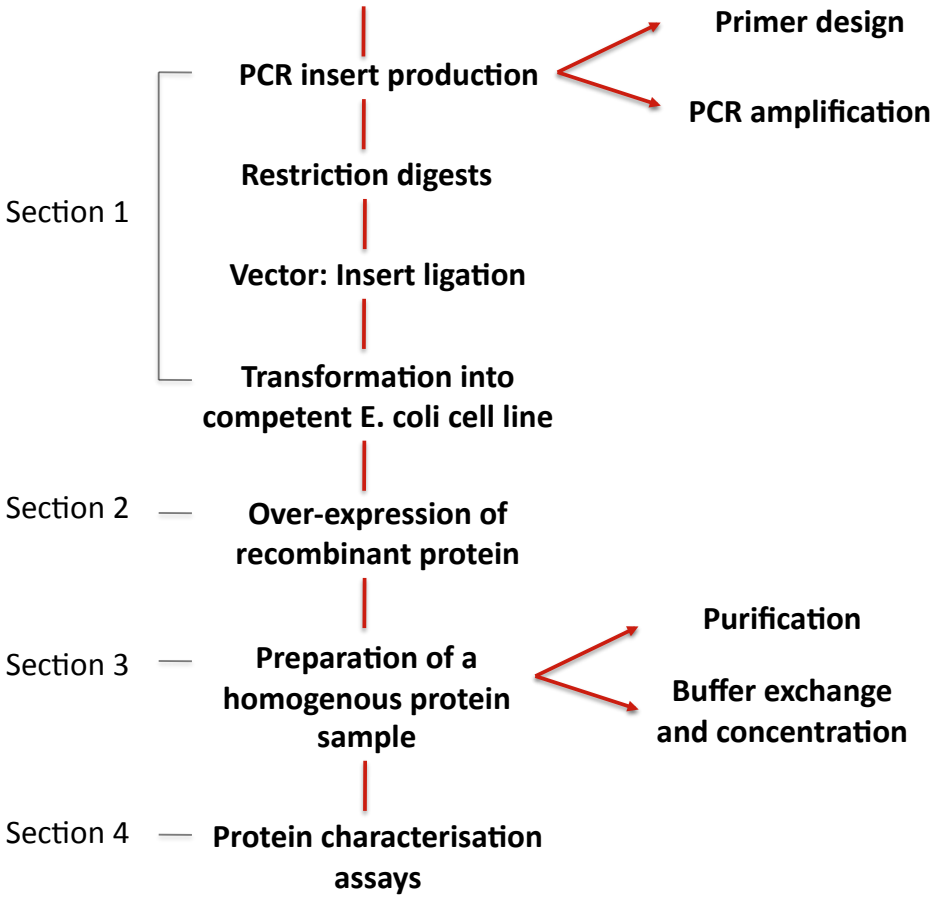


Figure 4.1.1 – Flow chart outlying recombinant protein production. All the recombinant protein samples discussed in this thesis have been produced in the same fashion. Through the construction of a modified plasmid, that when transformed into a genetically altered *E. coli* cell line promotes artificially high levels of foreign protein production. The target gene is isolated from its host microorganism's genome using polymerase chain reaction (PCR). PCR amplifies a specific gene through short sequence specific complimentary oligo-nucleotides called primers and thermal cycling exploiting heat stable DNA polymerases. The isolated gene and intended vector are then both digested with restriction enzymes specific to sites introduced by the PCR primers. These restriction enzymes introduce short single stranded overhangs. The plasmid and gene fragments samples are then mixed together and incubated across a range of temperatures in the presence of T4 Ligase; an enzyme that joins the complimentary overhangs together re-circularising the vector whilst incorporating the insert. This construct is then transformed into an engineered *E.coli* expression cell line, which induces over-expression of the target gene when IPTG is introduced. The final stage is to prepare a homogenous pure sample of the target protein using a variety of chromatography stages and buffer exchange into a low salt buffer suitable for crystallisation.

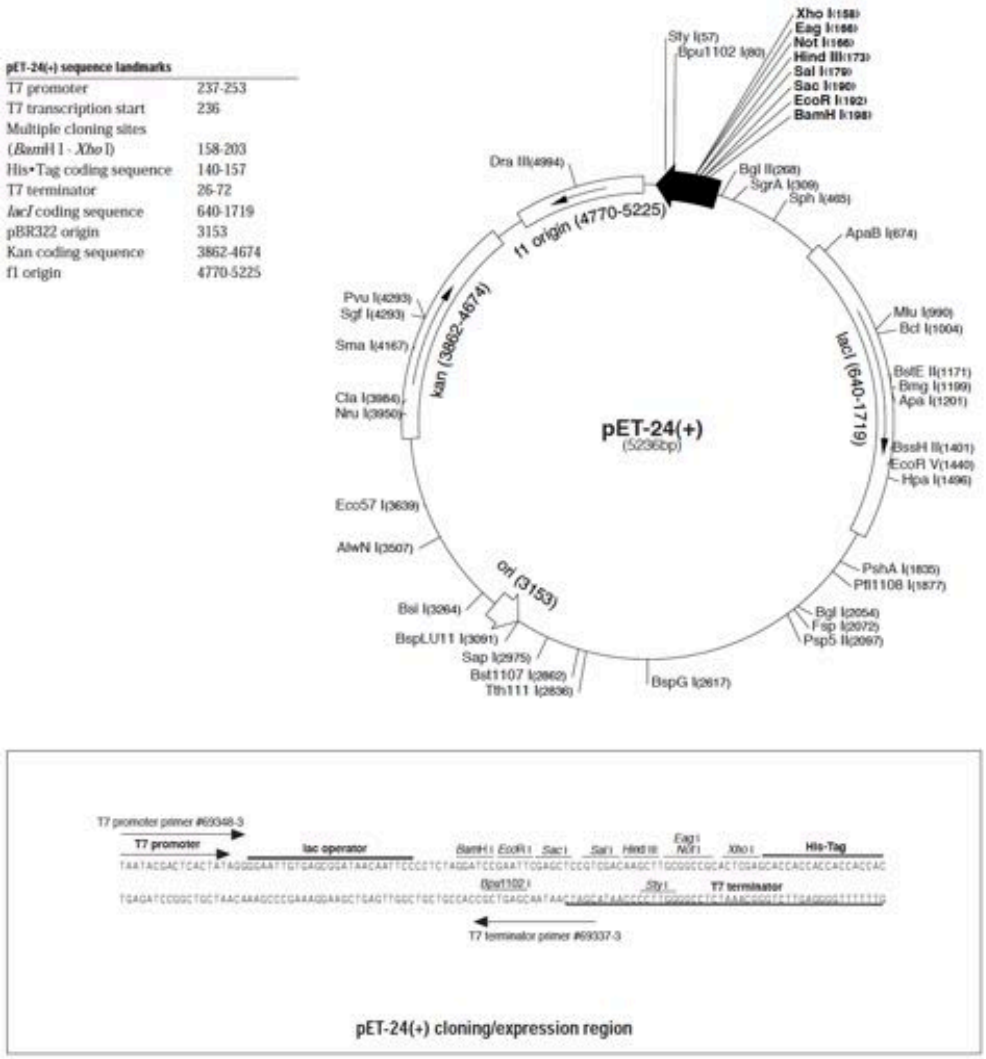


Figure 4.1.2 – Plasmid map for a pET24 vector. All the plasmids used in this study belong to the pET family of expression vectors. Exaggerated levels of protein production require the transcription of the target gene above levels commonly exhibited in prokaryotic cells. To achieve this expression plasmids incorporate a viral T7 RNA polymerase promoter, which cannot be down regulated by the host, with the corresponding T7 RNA polymerase encoded for by the expression cell line. As a result both the polymerase and target genes transcription are controlled by the *lac I* gene, which is ordinarily repressed apart from in the presence of lactose. Therefore, the introduction of a lactose analogue IPTG triggers T7 RNA polymerase production, which in turn transcribes the target gene. The remaining critical components in pET vectors are the antibiotic resistance cassette and poly-linker multiple cloning site. The antibiotic cassette serves a dual purpose. Allowing selection of cell lines incorporating the plasmid from background non-transformants, whilst also preventing the transformed cells from disposing of the plasmid. The poly-linker region provides a wide range of possible restriction sites for ligating into, essential as many genes can contain one or more internal restriction site.

4.1.2 Restriction site cloning into plasmid vectors

Production of a construct for recombinant protein production, involves producing a large quantity of exact repeats of the target gene. The amplified genes are modified with N and C terminal tags, containing specific sites where restriction endo-nucleases can bind and nick the DNA to create short overhangs of single stranded DNA. These restriction sites if correctly assigned should match a complimentary pair within the poly-linker region of the plasmid. When the plasmid has been digested this forms a site where the foreign gene can be inserted into the vector through ligation of complimentary single stranded overhangs, yielding a new construct.

4.1.3 Polymerase chain reaction

Amplification of the target gene was achieved using PCR (Bartlett and Stirling, 2003). First the double stranded host DNA is unzipped through exposure to boiling temperatures of 94 ^{o}C, leaving single stranded DNA that a short oligo-nucleotide primer can adhere to. This primer annealing process is temperature specific to each individual sequence and is a key primer design criterion. Once annealed the reaction proceeds through interaction with a heat stable DNA polymerase that extends the short region of double stranded DNA, starting from the primer before extending across the whole target gene, via complementary base pairing at a temperature of 72 ^{o}C. PCR reactions were performed using a Phusion (NEB^{TM}) cloning kit incorporating a proof reading DNA polymerase, in a final volume of 50 µl with the following components:

- 5 µl forwards primer 25 ng µl^{-1}.
- 5 µl reverse primer 25 ng µl^{-1}.
- 1 µl gDNA 100 ng µl^{-1}.
- 10 µl HF buffer.
- 1 µl 10 mM dNTP solution.
- 1 µl Phusion DNA polymerase.
- 0-5 µl DMSO.
- 25 µl sterilised MQ-H_2O

Standard cycle, repeated 30 times

- 94 ^{o}C hot start 5 minutes
- 94 ^{o}C melting 45 seconds
- 70-50 ^{o}C annealing 45 seconds
- 72 ^{o}C extending 1 minute
- 72 ^{o}C hold 10 minutes

Chapter 5: Protein production and structural determination of HCH_03101

This chapter details: the production, purification and crystallisation of recombinant protein for the candidates, identified in chapter 2. The later portions of this chapter will then move on to cover the diffraction data obtained. Finishing with the refinement and validation of the models built. This study will predominantly focus on the work undertaken solving the structure of HCH_03101, the highest priority Glutamine de-amidase target, produced by the marine bacterium *H. chejuensis*. With the current progress made on the alternative quartet of structural genomics targets detailed concurrently.

5.1 – Molecular cloning

All of the candidate proteins (chapter 2.7) required production through a recombinant over-expression system (chapter 4.1.1). The first step was to excise the target gene from its native organisms gDNA using PCR amplification (chapter 4.1.3). To facilitate the rapid amplification of multiple gene fragments, standardised thermal-cycle protocols and primer designs were employed. The constructs produced throughout this study are listed in table 5.1.1, with their related primers detailed in table 5.1.2. All of the primers described were designed to share a large consensus sequence with the target gene, approximately 15-25 nucleotides long. However, this level of consensus across a broad range of template sequences has resulted in a wide variety of annealing temperatures. Therefore, a straightforward approach that allows for the amplification of multiple genes (from inherently different template material) is the touchdown methodology. Touchdown PCR shifts from higher annealing temperatures (intended to encourage specific interactions) through to lower temperatures were amplification is more efficient. Consequently, this method relies upon a surplus of correctly amplified template being produced early in the reaction, which limits the quantity of non-specific template available at the less discriminating lower annealing temperatures.

5.1.1 Cloning HCH_03101 from *H. chejuensis*

The first gene cloned in this fashion was the top priority Glutamine de-amidase candidate, HCH_03101 from *H. chejuensis*. The initial amplification of the gene from genomic DNA (figure 5.1.1A), kindly supplied by the KRIBB institute, was initially extremely non-specific requiring the addition of 5 % DMSO.

Table 5.1.1 – List of genetic constructs produced and their experimental purpose.

Gene name and modification	Protein name and function	Cloning technique	Experimental purpose	Vector
BPSL_1549 C94S	BLF1 – Glutamine de-amidase toxin from *B. pseudomallei*	Quick-change SDM	Structural studies on an inactive mutant, published in Science see appendix.	pET-Blue
CNF1 C866S NTD 6xHIS	C-CNF1 – Glutamine de-amidase toxin from *E. coli*	Restriction cloning + Quick-change SDM	Proof of concept for the pull-down experiments	pET21a+
HCH_03101 WT	HCH_03101 putative Glutamine de-amidase from *H. chejuensis*	Restriction cloning	Structural genomics	pET24d+
HCH_03101 WT CTD 6xHIS	HCH_03101 putative Glutamine de-amidase from *H. chejuensis*	Restriction cloning	Structural genomics + pull-down assays	pET24d+
HCH_03101 C94S CTD 6xHIS	HCH_03101 putative Glutamine de-amidase from *H. chejuensis*	Quick-change SDM	Structural genomics + pull-down assays	pET24d+
PPUT_1063	PPUT_1063 putative Glutamine de-amidase from *P. putida*	Restriction cloning	Structural genomics	pET24d+
PPUT_1063 CTD 6xHIS	PPUT_1063 putative Glutamine de-amidase from *P. putida*	Restriction cloning	Structural genomics + pull-down assays	pET24d+
PSTAA_2862	PSTAA_2862 putative Glutamine de-amidase from *P. stutzeri*	Restriction cloning	Structural genomics	pET24d+
PSTAA_2862 CTD 6xHIS	PSTAA_2862 putative Glutamine de-amidase from *P. stutzeri*	Restriction cloning	Structural genomics + pull-down assays	pET24d+
HMPREF0758_5017	DNT - annotated Dermonecrotic toxin from *S. odourifera*	Restriction cloning	Structural genomics	pET24d+
HMPREF0758_5017 CTD 6xHIS	DNT - annotated Dermonecrotic toxin from *S. odourifera*	Restriction cloning	Structural genomics + pull-down assays	pET24d+
VSII_1134	VSII_1134 putative Glutamine de-amidase from *V. splendidus*	Restriction cloning	Structural genomics	pET24d+
VSII_1134 CTD 6xHIS	VSII_1134 putative Glutamine de-amidase from *V. splendidus*	Restriction cloning	Structural genomics + pull-down assays	pET24d+

Table 5.1.2 – List of primers used to amplify target genes from genomic template DNA.

Primer name	Primer sequence 5' to 3'	Tm (°C)
CNF1 NTD 6XHIS F	CGAAGC | CAT-ATG | CAT-CAC-CAT-CAC-CAT-CAC | AGT-ACT-GGA-AGC-ACC-TCC Nonsense NdeI NTD 6xHIS tag CNF1 ORF	78
CNF1 R	GGCGCG| GGA-TCC | TTA | AAA-TTT-TTT-TGA-AAA-TAC-CTT-C Nonsense Bam HI Stop CNF1 ORF	78
HCH_03101 F	GGTCCT | AC-ATG-T | GG-AGT-GAA-CTG-GAA-AGA-GT Nonsense NcoI HCH_03101 ORF	71
HCH_03101 R	GGAATA | GGA-TCC | TTA | GGC-CAA-CAC-AGC-CGT Nonsense Bam HI Stop HCH_03101 ORF	77
HCH_03101 CTD 6xHIS R	GGAATA | CTC-GAG | GGC-CAA-CAC-AGC-CGT Nonsense XhoI HCH_03101 ORF	82
PPUT_1063 F	TATATT | C-ATG | ACA-GGC-TTC-GAG-CG Nonsense FatI PPUT_1063 ORF	67
PPUT_1063 R	TGTATT | GGA-TCC | TCA | ATA-TTG-CAG-GAA-GCG-G Nonsense Bam HI Stop PPUT_1063 ORF	64
PPUT_1063 CTD 6xHIS R	TGTATT | CTC-GAG | ATA-TTG-CAG-GAA-GCG-G Nonsense XhoI PPUT_1063 ORF	65
PSTAA_2862 F	ATATTG | C-ATG | ACC-TTC-GGC-CGA Nonsense FatI PSTAA_2862 ORF	66
PSTAA_2862 R	TTATAT | GGA-TCC | CTA | ACG-GCA-CAT-GCC-G Nonsense Bam HI Stop PSTAA_2862 ORF	64
PSTAA_2862 CTD 6xHIS R	TTATAT | CTC-GAG | ACG-GCA-CAT-GCC-G Nonsense XhoI PSTAA_2862 ORF	65
DNT F	TATGTA | C-ATG | CAT-GAA-CAT-GGA-GTT-TA Nonsense FatI DNT ORF	65
DNT R	TATGTA | GGA-TCC | CTA | CCG-TCC-TAC-TAG-AAC-CTA-GAA-C Nonsense Bam HI Stop DNT ORF	67
DNT CTD 6xHIS R	TATGTA | CTC-GAG | CCG-TCC-TAC-TAG-AAC-CTA-GAA-C Nonsense XhoI DNT ORF	68
VSII_1134 F	TCGTCT | C-ATG | CCA-TTF-ACA-TTA-GAT Nonsense FatI VSII_1134 ORF	65
VSII_1134 R	TATGTA | GGA-TCC | CTA | CCG-TTC-TA-TAG-AAC-CTA-GAA-C Nonsense Bam HI Stop VSII_1134 ORF	67
VSII_1134 CTD 6xHIS R	TATGTA | CTC-GAG | CCG-TTC-TA-TAG-AAC-CTA-GAA-C Nonsense XhoI VSII_1134 ORF	68

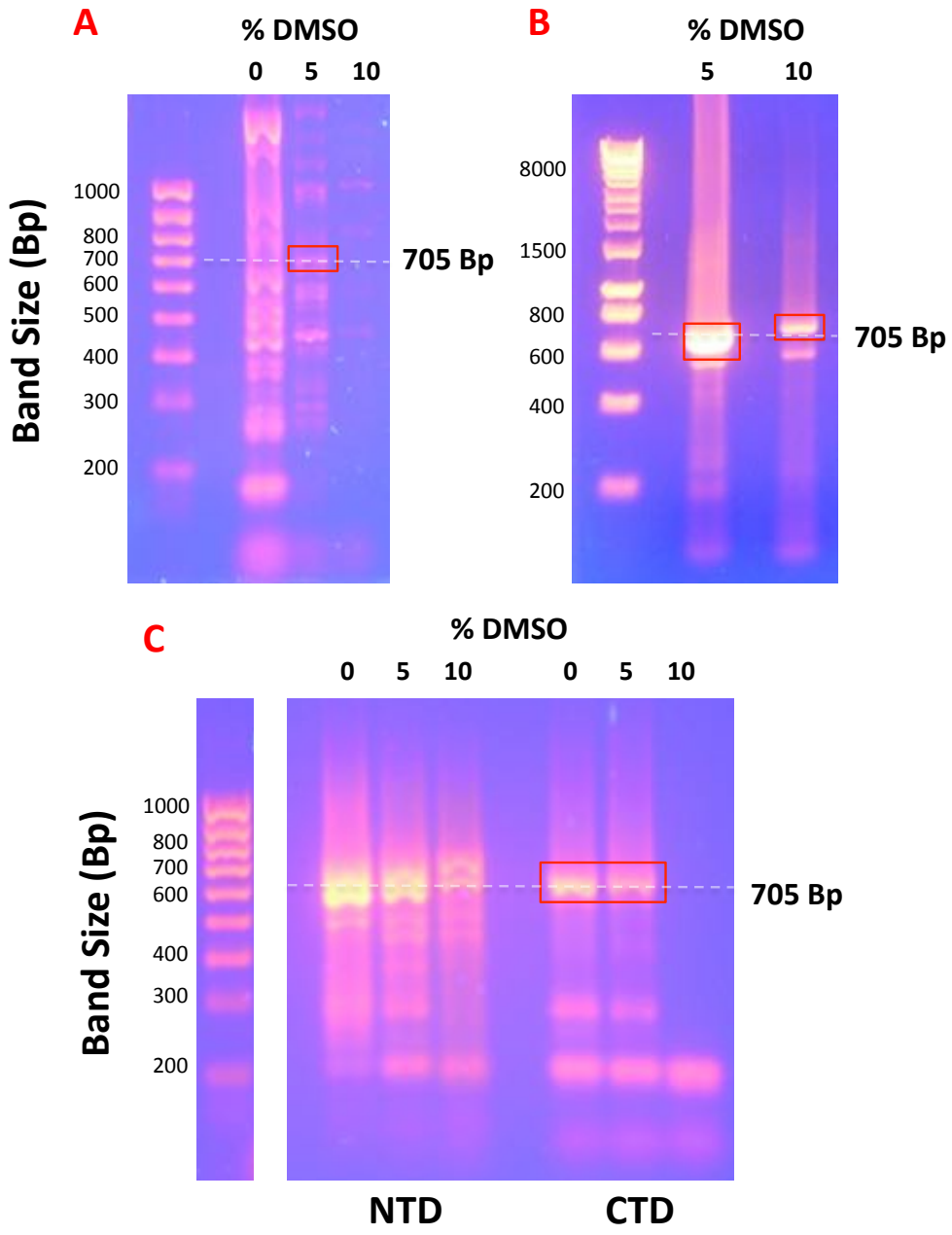


Figure 5.1.1 – PCR amplification of the HCH_03101 gene from genomic DNA. Panel **A** shows the initial amplification of the HCH_03101 gene, from 1 µl of genomic DNA (~ 50 ng μl^{-1}) supplied by the KRIBB institute. The non-specific amplification is initially over-whelming. However, with the addition of 5% (v/v) DMSO the non-specific amplification can be reduced to a level allowing excision of the correct 705 Bp sized band, highlighted in red. Panel **B** displays the amplification of the non-tagged HCH_03101 gene product, using the excised band from previous amplifications as template material (~ 50 ng). Despite being far less abundant, the fragment amplified in 10 % DMSO (v/v) provided the correct fragment. Panel **C** exhibits amplifications undertaken with the alternate 6x HIS tagged primers, both examples contain a fragment of the correct size (705 Bp). However, the only successful constructs were produced using fragments amplified with the CTD specific primer.

DMSO is a chemical that lowers the relative annealing temperature of the reaction, by un-folding super-coiled DNA, which in this case allowed identification of the correct sized fragment. This fragment was then excised from an agarose gel and purified using a gel extraction kit Qiagen^{TM} (section 4.1.5), for use as template material in subsequent PCR reactions. Figures 5.1.1B and C detail the amplification of both the native and tagged 6x HIS variants of HCH_03101 using this specific gene fragment as template material. This change in template leads to cleaner amplification of the target gene, which is suitable for purification by PCR cleanup Qiagen^{TM} (section 4.1.6).

Following amplification, the HCH_03101 gene fragments were then inserted into an expression plasmid, pET24d+, through complementary single stranded sticky ends, introduced by restriction endo-nucleases. Therefore, the native HCH_03101 construct was digested at the 5' end with NcoI and PciI for the vector and insert respectively; then digested at the 3' end with Bam HI for both the vector and insert. The HIS tagged variant of HCH_03101 was digested in a similar fashion but with the 3' Bam HI exchanged with an XhoI restriction site. Post digestion the DNA fragments were then ligated together, through incubation with T4 DNA ligase (section 4.1.8) at temperatures rising from 4 ^{o}C through to 20 ^{o}C across 24 hours. The ligated constructs were then sequenced by the Core genomics group (University of Sheffield), which confirmed them as an unaltered match for the sequence deposited in the NCBI non-redundant database.

5.1.2 – Cloning the remaining structural genomics targets

Once functional studies on HCH_03101 were well underway, the remaining target proteins were cloned simultaneously. The amplification of: PPUT_1063, VSII_1134, PSTAA_2862 and *S. odourifera* DNT was undertaken with genomic DNA provided by Dr Richard Eaton (Meredian Bioplastics ^{Inc}), Professor Didier Mazel (Pasteur institute) and the DSMZ repository respectively. Unlike HCH_03101, all but one of the aforementioned genes was amplified without excessive non-specific contamination (figure 5.1.2) then excised for secondary PCR reactions. However, despite repeated primer redesigns and the production of fresh genomic DNA it has not proved possible to amplify a fragment corresponding to VSII_1134. The 3 successfully amplified fragments where then digested with 5' Fat I and either 3' Bam HI or 3' Xho I, for the WT and 6xHIS tagged variant respectively. These digested fragments were then ligated into a similarly digested pET24d+ vector using the same methods described previously, prior to sequencing by the Core genomics group (University of Sheffield). All three constructs were confirmed to be unaltered matches from the sequences deposited in the non-redundant database.

5.1.3 – Producing a selection of active site mutants of both HCH_03101 and BLF1.

The final genetic constructs produced for this study were a variety of active site mutants. The first was a SER substitution at the essential CYS 94 position in BLF1, which was used for substrate binding assays and structure determination for a publication (appendix 1). This C94S mutant was also replicated in HCH_03101, which is hypothesised to conserve its catalytic CYS residue at the same sequence location, position 94. Therefore, mutation at this position should result in a inactive protein that broadly shares the same structural features as the WT. This construct could then be used for a range of assays and co-crystallisations, which require the irreversible capture of the enzymes substrate. The presented mutants were amplified using a variation of the Quick-change SDM methodology (section 4.1.12). Where a circularised plasmid template is amplified in its entirety with long complimentary primers encoding for the desired substitution. Initial, unsuccessful, attempts at producing these mutants with both BLF1 and HCH_03101 templates were undertaken with a stock Quick-change II kit (Agilent^{TM}). However, successful mutation was achieved with a modified protocol; which utilising the same overall scheme as the Quick-change protocol, whilst substituting the stock DNA polymerase Pfu turbo (Agilent^{TM}) with Phusion (NEB^{TM}) and reducing the annealing temperature in 2^{o} steps every 10 cycles from 68-58 ^{o}C. To reduce the chance of transforming non-modified template material these amplified samples are then digested with DpnI prior at 37 ^{o}C for 2 hrs, prior to transformation into DH5α competent cells. DpnI is a restriction enzyme that nicks methylated DNA, the amplified material is produced from non-methylated dNTP stocks, whereas the template material is sourced from an *E. coli* a strain that methylates its DNA. Therefore, the digestion of template material allows for selection of the mutated plasmids. DNA sequencing performed by the Core genomics group (University of Sheffield) confirmed that all the mutated constructs contained the desired substitution without additional modifications.

5.2 – Over-expression of recombinant protein samples

Having produced a selection of genetic constructs, trials were undertaken to determine the best conditions for producing large quantities of soluble recombinant protein. All the genes cloned were trialled for expression in an identical fashion, with the construct first transformed into Tuner (DE3) competent cells, then incubated at varying temperatures in the presence of IPTG. Expression trials were conducted in 5, 50ml flasks each inoculated with 2 ml of an overnight starter culture then grown to an OD_{600} of 0.6. This culture was then induced with an overall concentration of 1mM IPTG, before incubation at 15, 20, 25, 30 and 37 ^{o}C for 24 hours.

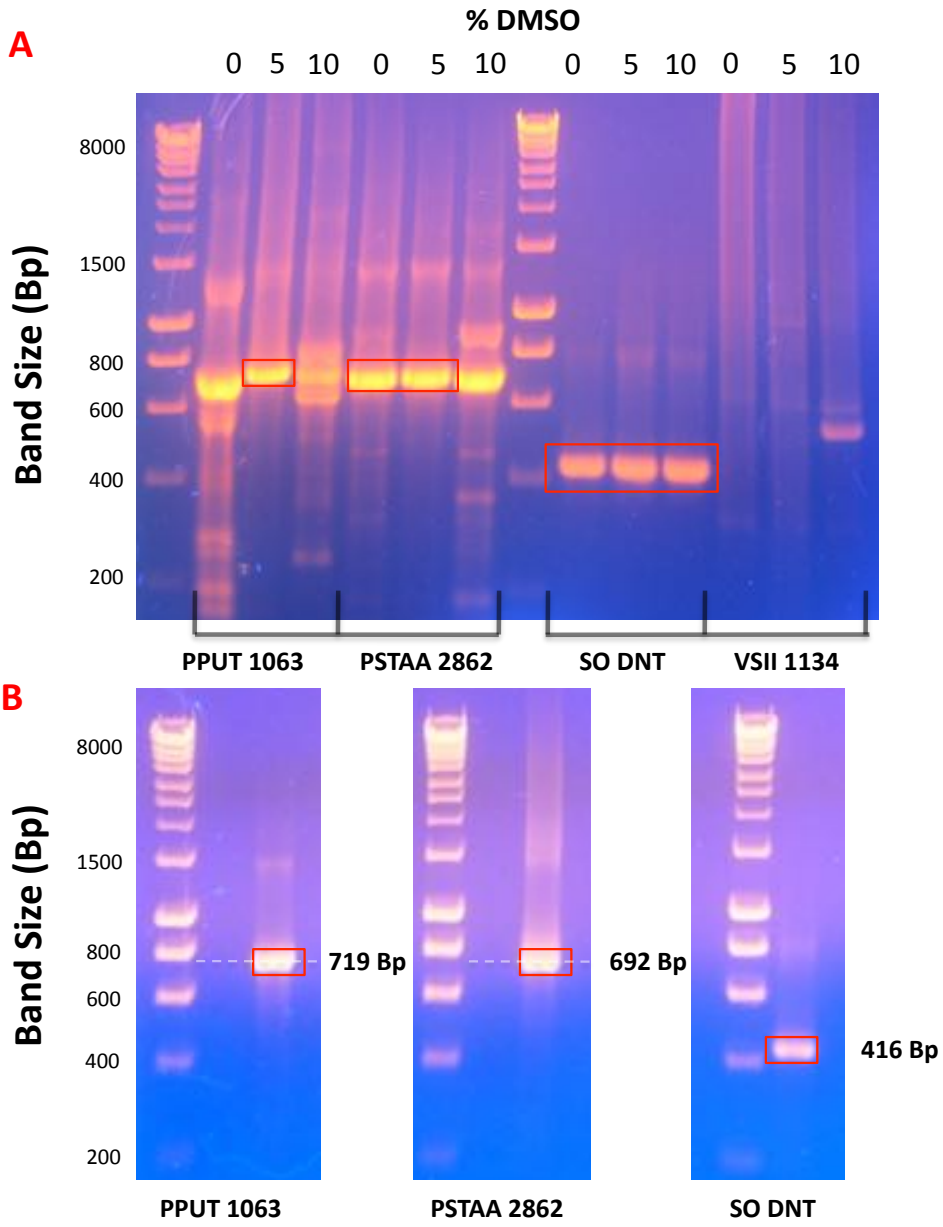


Figure 5.1.2 – Amplification of targets not taken through to structural determination. Panel **A** shows the PCR amplification of the 4 remaining Glutamine de-amidase candidates. All four genes were amplified simultaneously from ~ 50 ng of fresh template gDNA. PPUT_1063 and PSTAA_2862 amplified best in the presence of 5 % Dimethyl sulfoxide (v/v), yielding clear bands at 719 and 692 Bp respectively. Whilst the *S. odourifera* DNT did not require any additives to produce a single fragment at 416 Bp. However, despite repeated attempts and multiple primer re-designs it was not possible to amplify VSII_1134. Panel **B** displays the secondary PCR reactions, using 50 ng of the gel-extracted PCR product previously shown as template material. In all three cases the addition of DMSO was unnecessary with clear amplified bands present at the correct size locations.

5ml samples were then taken and centrifuged at 5000 *g* for 10 minutes prior to induction to act as a control. Additional 5 ml samples were then collected and centrifuged from each trial and corresponding temperature bracket at 4, 8 and 24 hr intervals, with the cell pellets frozen for storage. The level of expression from each trial was then tested by resuspending the cell pellet in 400 µl of 50 mM Tris pH 8, prior to cell breakage through 3, 4 second intervals of sonication. These broken cells were then centrifuged at 20, 000 *g* for 10 minutes and the soluble cell free extract removed. The insoluble portion contained within the pellet was then solubilised in a 50 mM Tris pH 8, 4 % SDS buffer and incubated at 20 ^{o}C under agitation for 20 minutes, prior to being centrifuged at 20, 000 *g* for 10 minutes. The protein concentration of the soluble cell free extract was then determined by Bradford assays (section 4.2.4), with 20 µg of the protein loaded onto an SDS-PAGE gel for analysis.

5.2.1 Over-expression trials of HCH_03101

The un-tagged WT HCH_03101 (figure 5.2.1A) expresses best at 25 ^{o}C for 8 hours. In these conditions it accounts for approximately 10 % of the cells total protein content, with negligible in-soluble expression. However, the tagged variants of HCH_03101 both favour lower expression temperatures. The native 6x HIS construct (5.2.1B) expresses best at 20 ^{o}C for 24 hours, with insoluble products outweighing the useful material within 8 hours at higher temperatures. Whilst the Se-MET derivative, which was trialled in minimal expression medium (figure 5.2.1C) expresses more slowly with the best expression obtained at 15 ^{o}C after 24 hours.

5.2.2 Over-expression trials of the structural genomics targets not pursued

At present none of the alternative Glutamine de-amidase constructs yield soluble protein, with PPUT_1063 producing vast amounts of insoluble protein, whilst the remaining constructs produced no detectable expression whatsoever. Sequencing of the constructs has confirmed that the foreign genes are present, un-altered, in the expression plasmid. However, it does not provide adequate coverage of the flanking region surrounding the gene. For example, the stock plasmid could have developed a mutation within the promoter region, or an insertion prior to the open reading frame. Therefore, all three genes require re-cloning into a different batch of pET24d+ or a different pET vector altogether, before being discarded as failures.

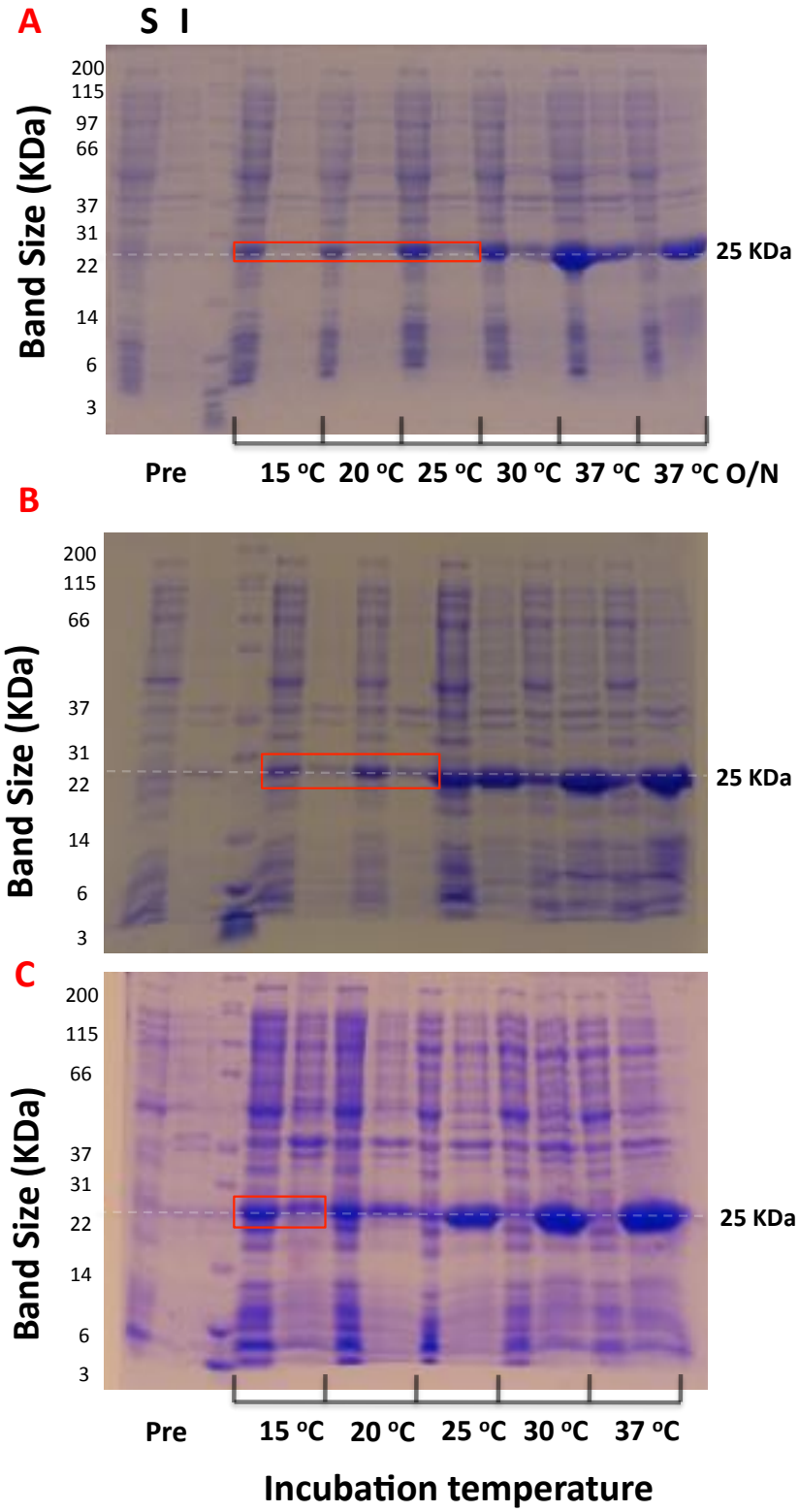


Figure 5.2.1 – Over-expression trials of HCH_03101. Panel **A** displays the 8 hour time point for the un-tagged WT HCH_03101 over-expression trial, with the best conditions highlighted in red. Panel **B** corresponds to the CTD 6x HIS variant at 24 hours and panel **C** the Se-MET derivative of the 6x HIS variant trailed in minimal medium at 24 hours. These tagged HCH_03101 constructs are shown to express best for 24 hours at 20 ^{o}C and 15 ^{o}C, for the native and Se-MET derivative samples respectively.

5.3 – Purification of HCH_03101 protein samples

5.3.1 Purifying tagged CTD 6xHIS HCH_03101

The tagged (CTD 6xHIS) WT construct of HCH_03101 was the first to be successfully over-expressed, its terminal HIS tag permits an efficient two-step purification (figure 5.3.1). 3 g of cell paste were resuspended in 35 ml of buffer A (50 mM Tris pH 8, 0.5 M NaCl) and the cell walls disrupted through 3, 20 second rounds of sonication. The soluble fraction was then separated by centrifuging the broken cells at 60,000 *g* for 20 minutes in a JLA-25-50 rotor (Beckman^{TM}), before being loaded onto a Ni-HP 5 ml column (GE healthcare^{TM}) at 5 ml min^{-1}. The protein was then eluted from the Ni-HP 5 ml column with a linear gradient moving from 0-70% buffer B (50 mM Tris pH 8, 0.5 M NaCl, 0.5 M Imadizole) in 15 column volumes, HCH_03101 is eluted with 0.3 M Imadizole. SDS-PAGE analysis of this purification step, shows that this chromatography stage achieves approximately 95 % purity. The final stage of the purification is size exclusion chromatography using a gel-filtration column, the selected Ni-HP fractions were concentrated to approximately 8 mg ml^{-1} in a 2 ml volume and then loaded onto a Superdex 200 16/60 Gel-filtration column (GE healthcare^{TM}) at 1.5 ml min^{-1}. The column was then run at a rate of 1.5 ml min^{-1} for 120 ml, with a single peak eluted at 90 ml. This corresponds to a K_{av} of 0.55, which is equivalent to a 25 KDa molecular weight, which in turn is close to the anticipated monomeric mass of HCH_03101. However, despite eluting in a single sharp peak the overall purity of the sample is no better as a result of Gel-filtration.

Throughout expression trials HCH_03101 behaved as expected for a readily soluble protein. However, during purification it became apparent that this recombinant protein is sensitive to both low salt and high concentration conditions, rapidly precipitating in buffers below 100 mM NaCl and at concentrations above 8 mg ml^{-1}. Therefore the useful lifetime of the protein in favourable (high salt) conditions is approximately three hours, severely limiting the scope of experiments possible. Nonetheless, removing the Gel-filtration stage (which offers little purity benefit) significantly increases the lifetime of the protein to approximately 5 hours. Analysis of this purified protein under a microscope. Shows that in certain buffering conditions (25 mM HEPES pH 7.5, 100 mM NaCl) and at concentrations > 7 mg ml^{-1} HCH_03101 is liable to form either small poorly diffracting crystals, or insoluble precipitate within a day of storage.

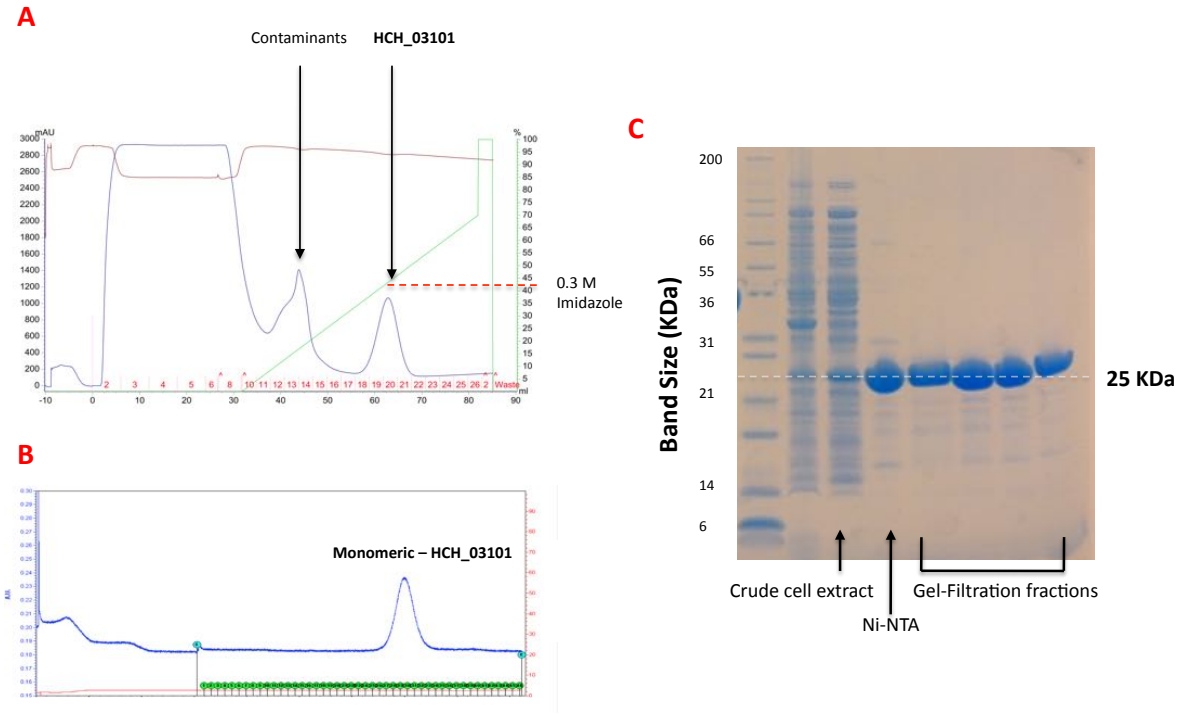


Figure 5.3.1 – Purification strategy for 6x HIS tagged HCH_03101 protein samples. Panel **A** displays the purification trace of a Ni-HP 5 ml column. There are two peaks eluted at 45 and 65 ml, corresponding to contaminants and HCH_03101 respectively. The column was run on a gradient from 0 – 70 % buffer B (50 mM Tris pH 8, 0.5 M NaCl, 1 M Imidazole) with HCH_03101 dissociating from the matrix with 0.3 M Imidazole. The Ni-NTA affinity column is followed by a Gel-filtration step, shown in panel **B**, were HCH_03101 elutes in a single peak at 90 ml corresponding to a monomeric 25 KDa molecular weight. Panel **C** is an SDS-PAGE gel for the above purification protocol: Lane 1, Mark 12 MW marker (Novex^{TM}); lane 2, 20 µg Ni-HP run off; lane 3, crude cell free extract; lane 4, 20 µg pooled Ni-HP elution and lanes 5 -8, 20 µg Gel-filtration fractions. This gel displays that a single Ni-HP stage is capable of achieving 95 % purity with no additional gains made with Gel-filtration.

5.3.2 – Purifying un-tagged WT HCH_03101

Having previously established that the tagged (CTD 6xHIS) HCH_03101 construct was unstable, led to the production of an un-tagged construct. A purification protocol was developed to provide pure HCH_03101 protein, in as few chromatography stages as possible. HCH_03101 has an iso-electric point of 9.12, predicted by PROTPARAM (Gasteiger *et al.,* 2003). As a result it is an ideal candidate for cation exchange chromatography. At neutral pH conditions the majority of proteins produced by *E. coli* exhibit an acidic pH, as a result the cation exchange matrix will only interact with a small subset of the natively produced contaminants. The first purification trial was with a CM-Sepharose (GE Healthcare^{TM}) column, a weak cation exchanger. However, In this case the matrix is too weak with a large proportion of the sample protein running straight through (figure 5.3.2). The remaining bound protein was eluted with a linear gradient across 10 column volumes (10 x 20 ml), moving from 100 % buffer A (50 mM MES pH 6.5) to 100% buffer B (50 mM MES pH 6.5, 0.5 M NaCl), with HCH_03101 eluting at 0.2 M NaCl. SDS-PAGE analysis of this chromatography step shows that the eluted protein is significantly contaminated (figure 5.3.2B). Therefore, a stronger cation exchange resin, SP-Toyopearl (Toyopearl^{TM}) was attempted next. When run using the same buffer and gradient conditions as the previous CM-Sepharose stage, the SP-Toyopearl column yields a single peak at 0.3 M NaCl, that is 95 % pure (figure 5.3.3). Unfortunately, further experiments to improve upon the purity, by adding a Resource S (GE Healthcare^{TM}) column (the strongest cation exchanger available) resulted in irreversible protein adherence to the matrix. Consequently, all future WT HCH_03101 purifications were undertaken using an SP cation exchange column.

The finalised protocol for purifying un-tagged HCH_03101 protein (figure 5.3.4), involves several minor alterations from the preliminary trials; Including the substitution of a self-packed SP-toyopearl column with a pre-packed SP-HP (GE Healthcare^{TM}) 5 ml column and the addition of a Gel-filtration stage. 3 g of cell paste was broken in 35 ml of buffer A (50 mM MES pH 6) through 3, 20 second rounds of sonication. The disrupted cells were then centrifuged at 60,000 *g* for 20 minutes in a JLA-25-50 rotor, before being loaded directly onto the SP-HP 5 ml column. This column was run at a rate of 5 ml min^{-1}, with a linear gradient from 0 -100 % buffer B (50 mM MES pH 6, 0.5 M NaCl) across 15 column volumes. HCH_03101 is eluted from this column at 0.2 M NaCl, with no significant contaminant peaks. SDS-PAGE analysis of the peak (figure 5.3.4A) shows that the protein is approaching >95 % purity from a single chromatography step.

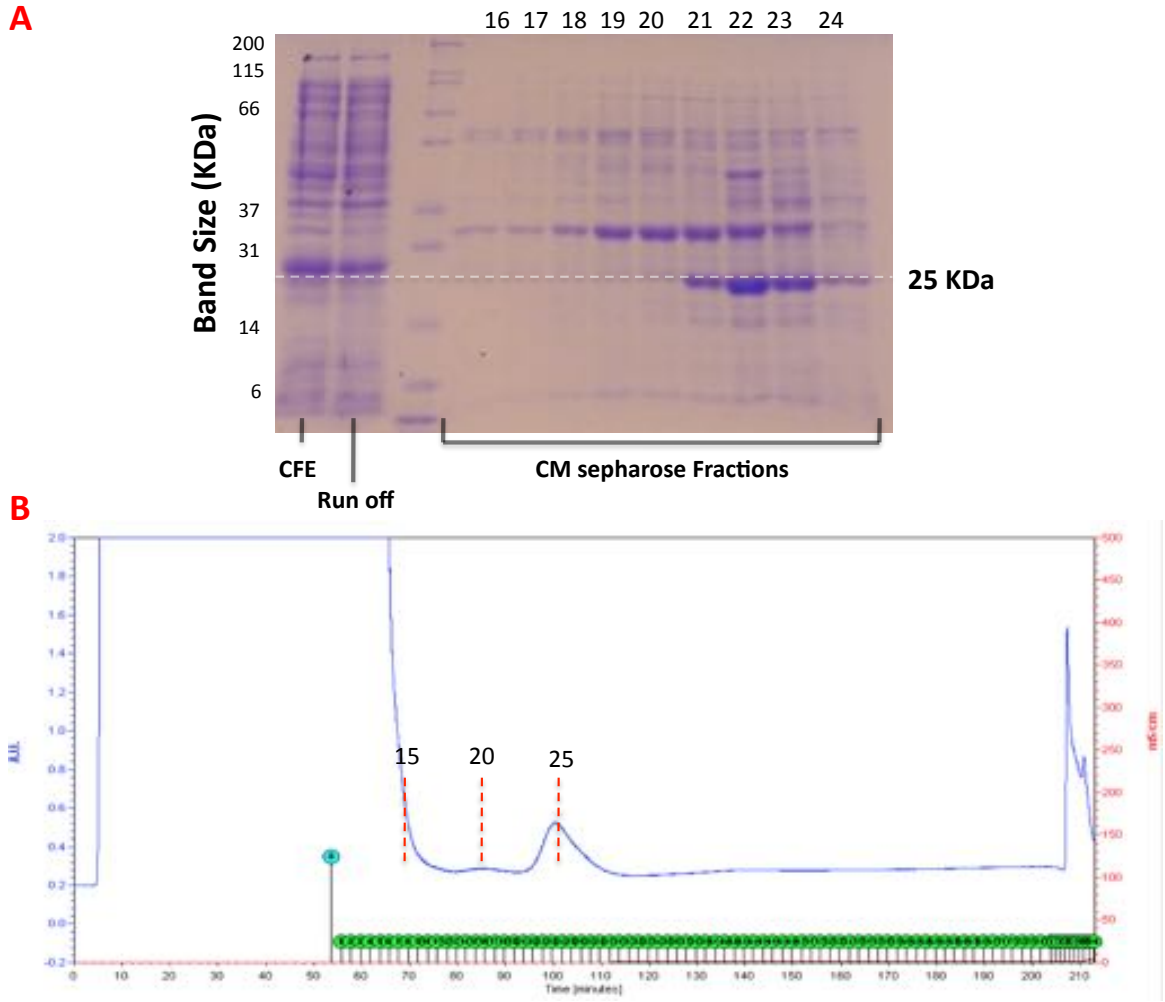


Figure 5.3.2 – Purification of un-tagged HCH_03101 using a CM-sepharose column. Panel **A** displays an SDS-PAGE analysis of a 20 ml CM-sepharose column: Lane 1, 20 µg of cell free extract; lane 2, 20 µg CM-sepharose of run off; Lane 3, mark 12 MW marker; lanes 4-12, 13 µl (~15 µg) of CM-sepharose fractions 16-24. The run off contains a large proportion of the soluble HCH_03101 loaded onto the column, indicating that the CM resin does not interact tightly to HCH_03101. Panel **B** shows the corresponding purification trace for the CM-sepharose column, with the only clear peak between fractions 20 -25 at approximately 0.2 M NaCl. The SDS-PAGE analysis shows that the large peak corresponds to HCH_03101. However, the fractions are heavily contaminated with a prominent band observed at 34 KDa.

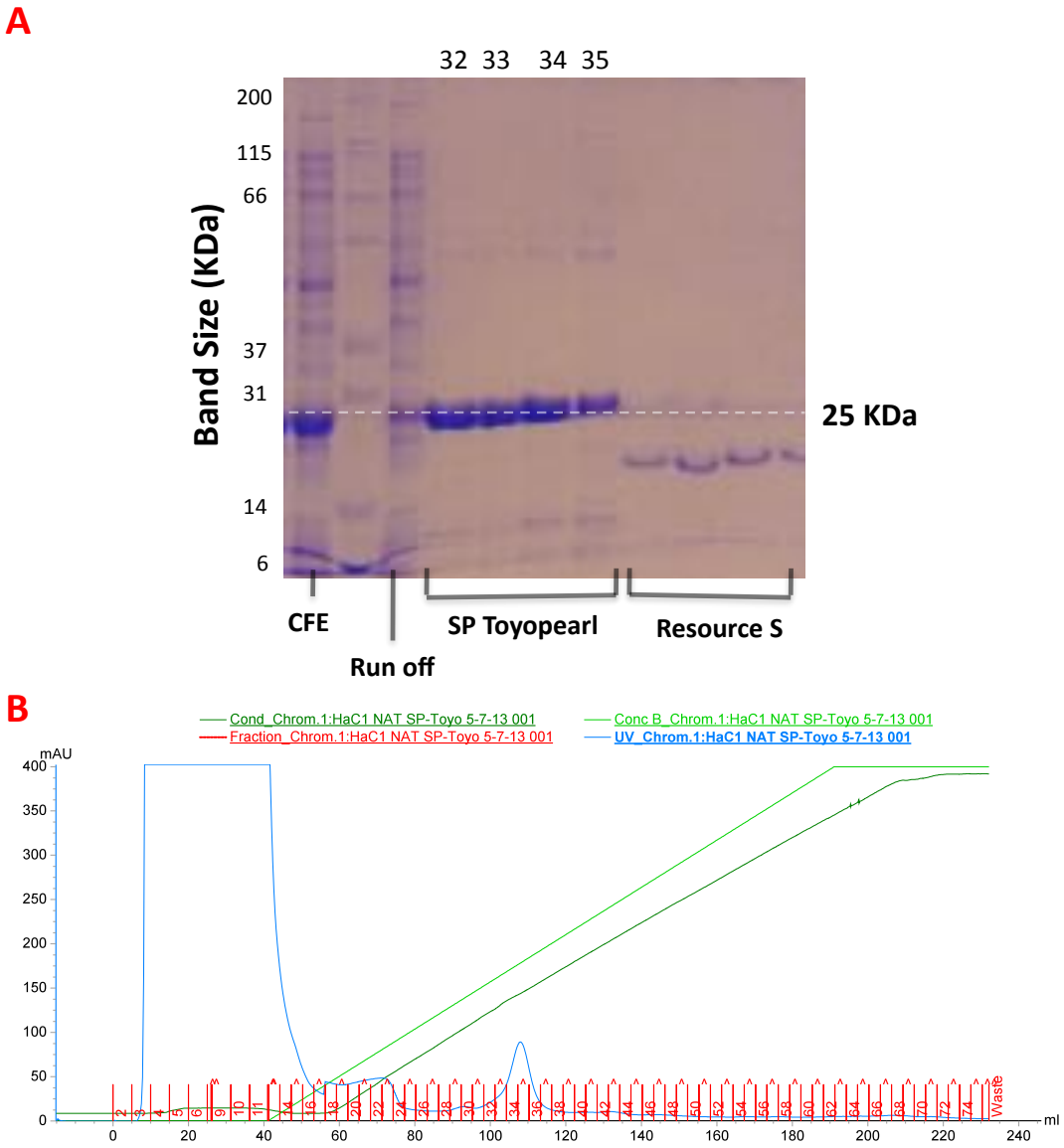


Figure 5.3.3 – Purification of un-tagged HCH_03101 with an SP-Toyopearl column. Panel **A** displays an SDS-PAGE analysis of the elution profiles from first a 20 ml SP-Toyopearl column and afterwards a Resource S 5 ml column. Lane 1, 20 µg soluble cell free extract; lane 2, Mark 12 MW marker; lane 3, 20 µg SP-Toyopearl run off; lanes 4-7, 20 µg SP-Toyopearl fractions 32-35, lanes 8-11, 13 µl Resource S fractions. Panel **B** is the purification trace from the SP-Toyopearl column showing a single peak at 110 ml, with 0.2 M NaCl. The SDS-PAGE gel highlights that SP type resins are well suited for purifying HCH_03101, with the majority of the protein interacting with the resin and the resultant HCH_03101 > 95 % pure. However, HCH_03101 binds irreversibly to Resource S columns removing any potential for improving purity through additional ion exchange chromatography.

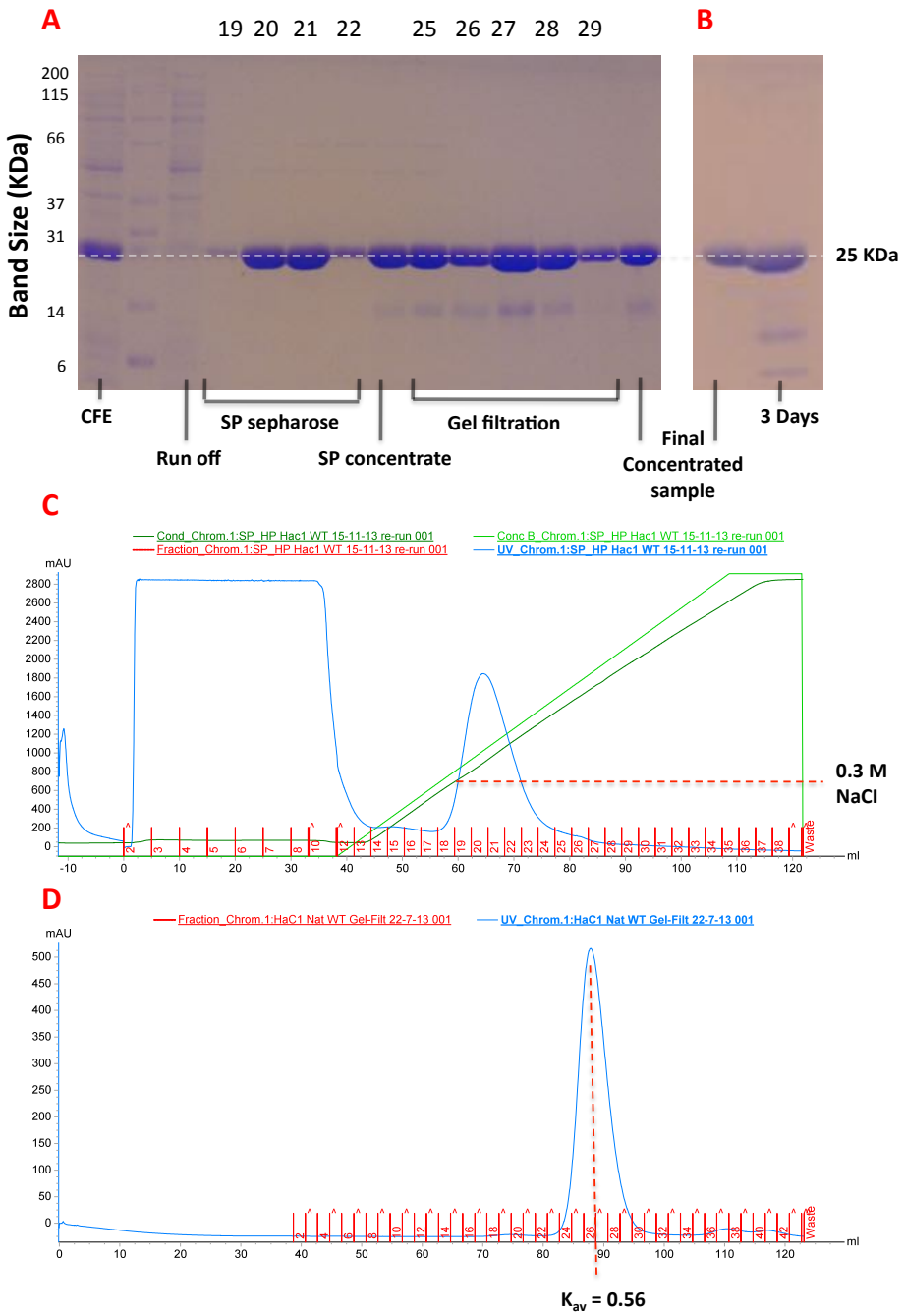


Figure 5.3.4 – The finalised untagged WT HCH_03101 purification strategy. Panel **A** shows an SDS-PAGE gel of the untagged WT HCH_03101 purification scheme. Lane 1, 20 µg cell free extract; lane 2, Mark 12; lane 3, 20 µg SP-HP run off; lanes 4-7, 5 µl SP-HP fractions 19-22; lane 8, 20 µg pooled SP-HP fractions; lanes 9-13, 8 µl Gel-filtration fractions 25-29; lane 14, concentrated 7 mg ml-1 HCH_03101 in crystallisation buffer (25 mM HEPES pH 7.5, 100 mM NaCl). Panel **B** SDS-PAGE test for protein degradation: lane 1, 15 µg freshly concentrated HCH_03101; lane 2, 15 µg after 72 hours incubation at 4^{o}C. The protein is >95 % pure and displays only minor degradation. Panel **C** and **D** are the purification traces from the SP-HP 5 ml and Superdex 200 16/60 Gel-filtration columns respectively. HCH_03101 is eluted from the Gel-filtration column as a single peak at 88 ml, consistent with a monomeric 25 KDa molecular weight.

The fractions where then pooled and concentrated to 7 mg ml^{-1} in a 2 ml volume and loaded onto a Superdex 200 16/60 Gel-filtration column, run at 1.5 ml min^{-1} (figure 5.3.4C). The non-tagged HCH_03101 is eluted in a single peak at 88 ml, which corresponds to a K_{av} of 0.56 and molecular weight of 25 KDa.

The above purification scheme has the advantage of speed, which is important because as none of the solubility issues affecting the tagged constructs were negated. To the contrary, the un-tagged protein is less stable, precipitating within 2-3 hours of concentration. The protein also takes approximately 3 hours to concentrate, and cannot go further than 7 mg ml^{-1}. However, the observed precipitation is not pH dependent; with low concentration 1 mg ml^{-1} stocks incubated overnight in pH 6, 7 and 8 (50 mM Tris, 0.5 M NaCl) buffers, all precipitating overnight. This decrease in solubility also cannot be explained by degradation. Figure 5.3.4B shows an SDS-PAGE analysis of a sample incubated at 4 ^{o}C for 3 days in crystallisation buffer (25 mM HEPES pH7.5, 100 mM NaCl), with the resultant precipitant re-solubilised in detergent (50 mM TRIS pH 8, 50 mM SDS). This gel shows that lower molecular weight degradation products are present. However, the majority of the sample remains at a native molecular weight. Curiously, the untagged HCH_03101 protein has not been observed forming crystals when stored, a common occurrence with the tagged protein. .

5.3.3 – Purifying the C94S point mutant of CTD 6xHIS HCH_03101

The final HCH_03101 construct to be purified was a CYS to SER point mutant believed to be located within the putative proteins active site. This CYS position is the proposed catalytic nucleophile in HCH_03101, assigned on the basis of sequence homology with BLF1. Therefore any mutation at this site would be anticipated to knock out activity, without necessarily modifying the enzymes ability to bind its substrate. The primary purpose of this active site mutant was to act as a bait protein in pull-down experiments to elucidate the binding partners of HCH_03101. However, structural validation that the enzyme had not been modified on a tertiary basis by this mutation was also desirable. It was logical to attempt crystallisation with the same CTD 6xHIS variant of the C94S mutant, as would be used in the pull-down experiments. Therefore, the purification protocol was the same as described in section 5.3.1 for the WT tagged HCH_03101 protein with a typical C94S purification detailed in figure 5.3.5.

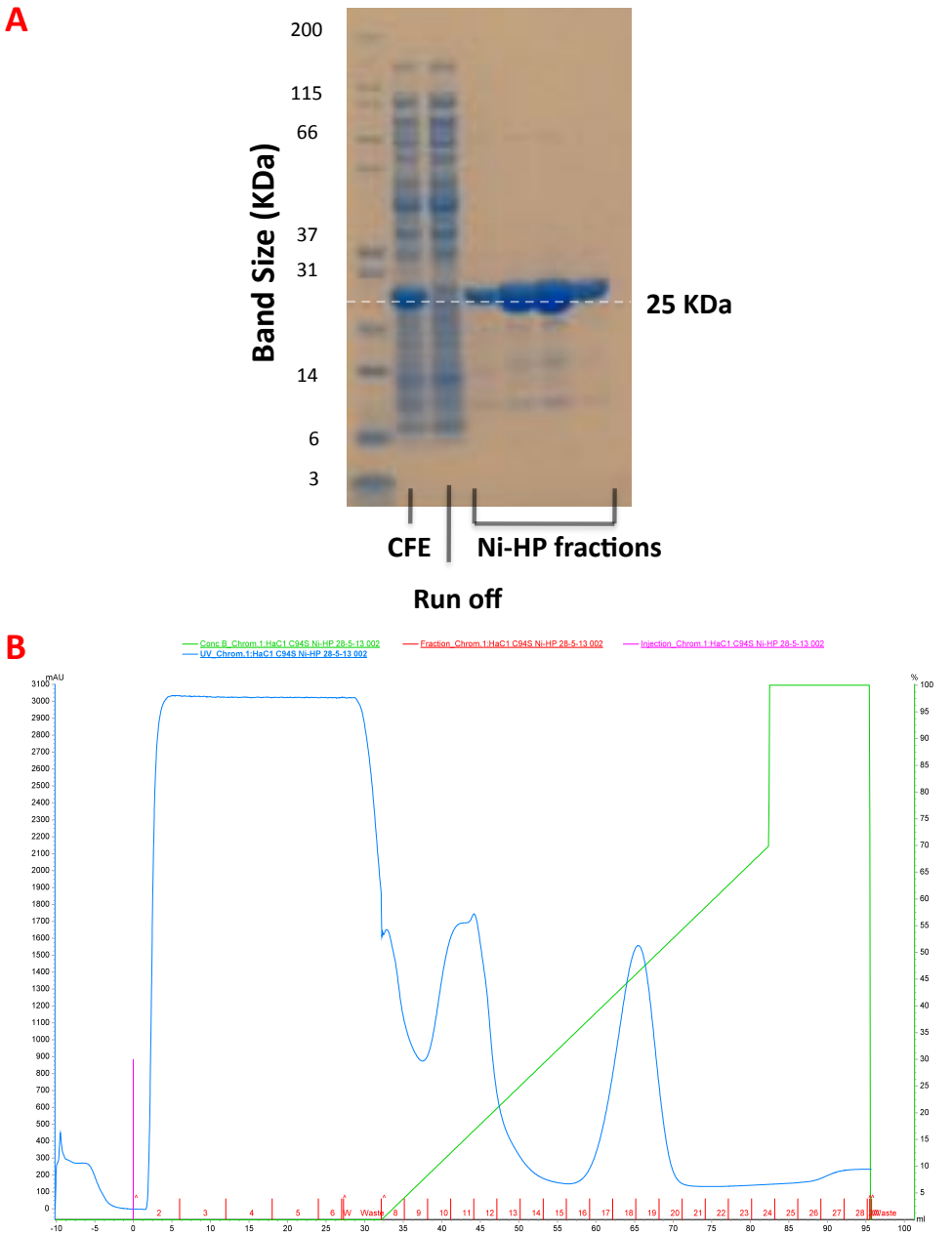


Figure 5.3.5 – Purification of the 6x HIS tagged C94S active site mutant of HCH_03101. Panel **A** is an SDS-PAGE analysis for the purification of HCH_03101 C94S (6xHIS tagged), using a Ni-HP 5 ml column. Lane 1, Mark 12 MW marker; lane 2, 20 µg cell free extract; lane 3, 20 µg Ni-HP run off; lanes 4-7, 15 µg Ni-HP fractions 17-20. Panel **B** is the purification trace for the Ni-HP 5 ml with the column run with a linear gradient moving from 0-70% buffer B (50 mM Tris pH 8, 0.5 M NaCl, 0.5 M Imidazole), showing that unlike the tagged WT HCH_03101 constructs the C94S mutant elutes with less Imadizole, at 0.25 M. Furthermore, the single Ni-HP stage yields protein >95 % pure. Therefore, skipping the Gel-filtration stage provides an extra 2 hours of working time prior to precipitation.

5.4 – HCH_03101 Crystallisation and diffraction tests

For crystallisation trials a sparse matrix screening strategy was employed, with the protein samples concentrated to between 7-9 mg ml^{-1} and trialled across 384 conditions encompassing the JCSG+, PACT, Classics (Qiagen™) and Proplex (Molecular Dimensions™) screens. These preliminary crystallisation conditions were set up in sitting drop vapour diffusion experiments (section 3.1.4), using a Hydra II Plus 1 crystallisation robot (Matrix™). With the robot programmed to administer 200 nl of protein into a 200 nl droplet of mother liquor, which was flanked by a 50 µl precipitant reservoir of precipitant. Favourable crystallisation conditions were then optimised upon by hanging drop vapour diffusion, with the protein and precipitant combined in 1 µl :1 µl and 1 µl :2 µl ratios on the same cover slip, above a 1 ml reservoir of precipitant. Both the sparse matrix and optimisation trials were produced in duplicate to be incubated at both 7 and 17 ^{o}C.

5.4.1 – Crystallising HCH03101 CTD 6xHIS

The tagged WT HCH_03101 crystallises in over 50 of the initial 384 conditions trialled, with little correlation observed between crystal formation and the specific salt and precipitant concentrations, or different buffering solutions. However, there was a single area of consistency, with the time frame required for crystallisation approximately a week. Despite crystallising across a range of conditions, the diffraction quality of the crystals varies dramatically from one condition to another in both resolution and mosaic spread. The hits that displayed the best diffracting crystals are displayed in figure 5.4.1A. All the crystals grown in sitting drop trials share eight-sided diamond morphologies, with uniform dimensions of 50 µm x 30 µm x 30 µm.

Crystals grown in PACT C9 (0.2 M Lithium chloride, 0.1 M Tris pH 8, 20 % (w/v) PEG 6000), Proplex A5 (0.1 M MES pH 6.5, 15 % (v/v) PEG 550 MME) and Proplex E9 (0.2 M Sodium Chloride, 0.1M HEPES pH 7.5, 12 % (w/v) PEG 8000) conditions were the first to be tested. These crystals were frozen prior to diffraction tests, in the same buffer and precipitant conditions present in the mother liquor with 30 % (v/v) Ethylene glycol. The PACT C9 crystals diffracted to 2.8 Å, whilst both Proplex conditions reached 3.5 Å. However, the quality of the diffraction in both cases was poor. The Proplex conditions exhibited figure of eight reflections, a common indicator of multiple lattices and macroscopic twinning. Whereas the crystals grown in PACT C9 appeared to diffract nicely out to 2.5 Å, but displayed a high mosaic spread of 6 ^{o}, which would hinder any phasing experiments downstream.

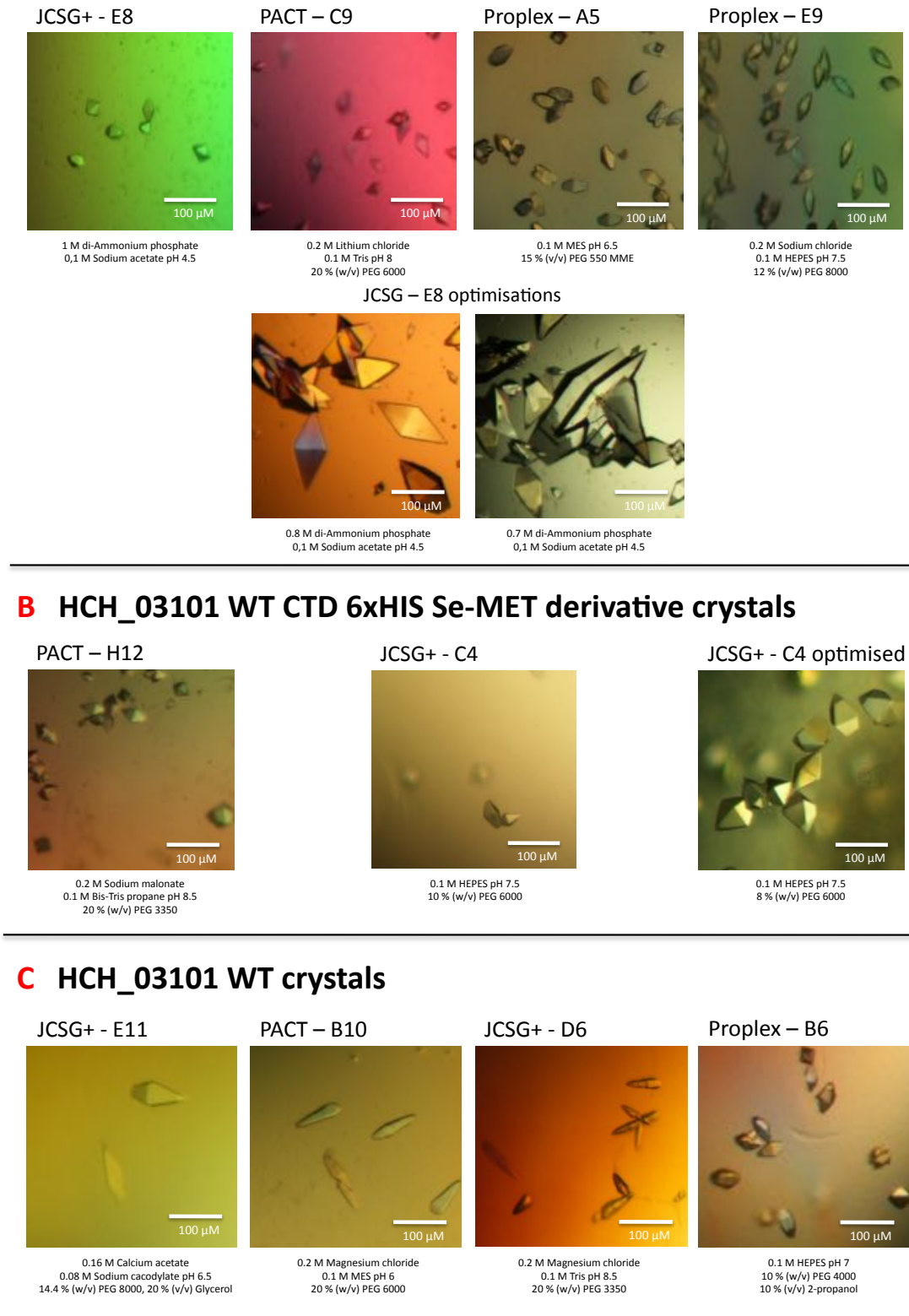


Figure 5.4.1 – Selection of crystallisation conditions uncovered for both tagged and non-tagged HCH_03101.

The only condition from the initial screens that combined high resolution with a low mosaic spread were grown in JCSG+ E8 (1 M di-Ammonium phosphate, 0.1 M Sodium acetate pH 4.5), diffracting out to 1.9 Å with a mosaic spread of 2 ^{o}. Optimising the JCSG+ E8 condition was attempted by varying the di-Ammonium phosphate concentration between 0.4 - 1.4 M and the Sodium acetate concentration between 0.05 - 0.2 M, in both cases this range is covered in 0.05 M steps. The largest crystals were grown in optimised conditions between 0.7 - 0.9 M di-Ammonium phosphate with 0.1 M Sodium acetate pH 4.5, measuring 150 x 70 x 70 µm. However, despite testing over 20 of the optimised crystals in a variety of cryo-protectants between: 15 – 35 % (v/v) Ethylene glycol, 25-30 % (v/v) Glycerol, 25-30 % (v/v) Sucrose and 25-30 % (v/v) PEG 400. None of these crystals diffracted above 2.15 Å. Therefore, the cryo-protectant does not appear to play a crucial role in determining diffraction quality.

5.4.2 – Crystallising the Se-MET phasing derivative of HCH_03101 CTD 6xHIS

With a selection of suitable native crystals of the tagged WT HCH_03101 produced and tested, a derivative incorporating heavy atoms was required to solve the phase problem (section 3.4). Attempts to soak in Mercury ethyl-compounds and salts proved unsuccessful. Therefore, with no Isomorphous replacement derivative possible, it was decided to attempt anomalous dispersion by incorporating Se-MET residues during over-expression (section 5.2.1). Therefore, HCH_03101 with 4 MET residues should provide 4 different heavy atom sites, plenty for phasing.

Expression trials for the Se-MET derivative, grown in minimal media supplemented with Se-MET, showed that the protein expresses at a lower level with more in the insoluble fraction. However, none of the previously identified conditions produced single crystals. Therefore, fresh sparse matrix trials where set up, with JCSG+, PACT and Classics screens laid down at 8 mg ml^{-1}. These trials identified two conditions: PACT H12 (0.2 M Sodium malonate, 0.1 M Bis-Tris propane pH 8.5, 20 % (w/v) PEG 3350) and JCSG+ C4 (0.1 M HEPES pH 7.5, 10 % (w/v) PEG 6000) (figure 5.4.1B). These crystals were diffraction tested frozen in mother liquor + 30 % (v/v) Ethylene glycol. With the PACT H12 samples not producing diffraction patterns but several crystals from the JCSG+ C4 condition reaching 2.6 Å resolution.

At a later date the C4 JCSG+ condition was optimised upon, by setting up hanging drop trials, which varied the precipitant concentration in 1 % steps between 4 - 20 % (w/v) PEG 6000 and buffering pH between 6 -8 in 0.5 steps. The optimised condition yielding the largest crystals was 0.1 M HEPES pH 7.5, 8 % PEG 6000. However, as with the native HCH_03101 (tagged) these larger crystals failed to diffract better than their sitting drop counterparts, with higher mosaic spread and lower overall resolution observed.

5.4.3 Crystallising the Native non-tagged HCH_03101

Despite the un-tagged HCH_03101 expression trials, showing less material in the insoluble fraction than the tagged variants. The Purification trials suggest that the untagged protein is significantly less stable. However, sparse matrix trials incorporating the JCSG+, PACT and Proplex screens were laid down for with the un-tagged HCH_03101 protein to isolate preliminary crystallisation conditions. The best hits identified from these screens (figure 5.4.1C) are spread across, the now customary, wide variety of conditions. However, unlike the un-tagged protein none of these conditions have produced crystals that diffract to a higher resolution then 3 Å. Moreover they also displayed significantly different unit cell dimensions to the previously discussed crystals, especially along the c axis where they are 13 Å longer than the tagged WT HCH_03101 (section 5.4.1).

5.4.4 Crystallising the C94S mutant of HCH_03101 CTD 6xHIS

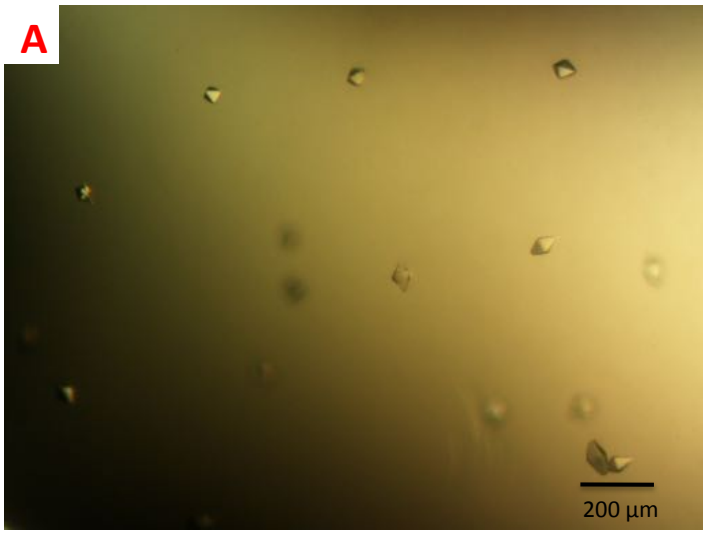
The C94S mutant of HCH_03101 predominantly exhibited crystallisation conditions in common with the WT 6xHIS construct. However, the condition that produced the best diffracting crystals, PACT E4 (0.2 M Potassium thiocyanate, 20 % (w/v) PEG 3350) was novel. This crystal was optimised along the same parameters as the derivative JCSG+ C4 crystals with the best crystals grown in 0.2 M Potassium thiocyanate, 12 % PEG 3350. Unlike the previous constructs the C94S mutant is the first to display better diffraction quality as a result of larger optimised crystals with data collected to 2.08 Å. These optimised crystals also display the same unit cell dimensions previously exhibited by the un-tagged crystals (section 5.4.3). With a c unit cell dimension, 12 Å longer than observed in any of the previously tested HIS tagged constructs (sections 5.4.1 – 5.4.2).

5.5 – HCH_03101 Se-Met derivative diffraction data

A wide variety of crystals were tested for their diffraction properties. The best crystals were grown in the sparse matrix condition JCSG+ C4 (0.1 M HEPES pH 6.5, 10 % PEG 6000) (figure 5.5.1A). These crystals were exposed to X-rays at the Diamond light source on beamline I03, diffracting out to a maximum resolution of 2.6 Å, which was comfortably the highest resolution observed from a Se-MET derivative crystal. They also exhibited tetragonal symmetry in space-group P 4_1 2_1 2. Prior to collecting any data these crystals were tested for heavy atom incorporation. A fluorescence scan identified a strong selenium signal from which the peak energy and atomic scattering factors (*f'* and *f''*) were estimated (figure 5.5.1B). The wavelength of the incident X-rays was consequently altered, to match the peak energy indicated from the above scan at 12660 eV (0.098 nm).

Phasing HCH_03101 was approached using a high redundancy, SAD data collection strategy. This was due to preliminary tests on the less well diffracting examples from this condition, which had shown that these crystals succumbed to radiation damage after approximately 400 ^{o} of rotation. Therefore, it is unlikely that a single crystal could survive the multiple exposures required to collect an effective MAD dataset. For the data collection a single crystal was tested, diffracting to 2.6 Å, with the anticipated P 4_1 2_1 2 space-group. A dataset incorporating 360 ^{o} of rotation in 0.2 ^{o} steps was collected (figure 5.5.2A). However, despite the data displaying a strong anomalous mid slope (1.45) and high multiplicity statistics (24 / 14), it was not possible to locate more than one of the heavy atom site using SHELX (Sheldrick, 2008). Undeterred, a second crystal was tested (figure 5.5.2B) using the same parameters, which diffracted to 2.8 Å resolution with similarly strong anomalous stats. However, this crystal also did not locate any additional heavy atom sites.

This was perplexing as every indicator pointed towards these data being capable of providing a solution. However, close inspection of the diffraction patterns revealed a troubling flaw with the diffraction. With both crystals displaying a high mosaic spread of 2 – 2.5 ^{o}. This mosaic spread value is unusually high, with each reflection persisting for upwards of 10-15 images, which has the effect of broadening the intensity peaks measured during the diffraction experiment. These broader intensity peaks make accurately determining the anomalous scattering extremely difficult. A potential solution to this problem would have been to attempt production of new crystals in an alternative condition. Unfortunately, in the case of HCH_03101, a 2.5^{o} mosaic spread was the lowest that was observed across any of the tested Se-MET derivatives and also many of the native crystals.



JCSG+ C4: 0.1 M HEPES pH 6.5, 10 % PEG 6000

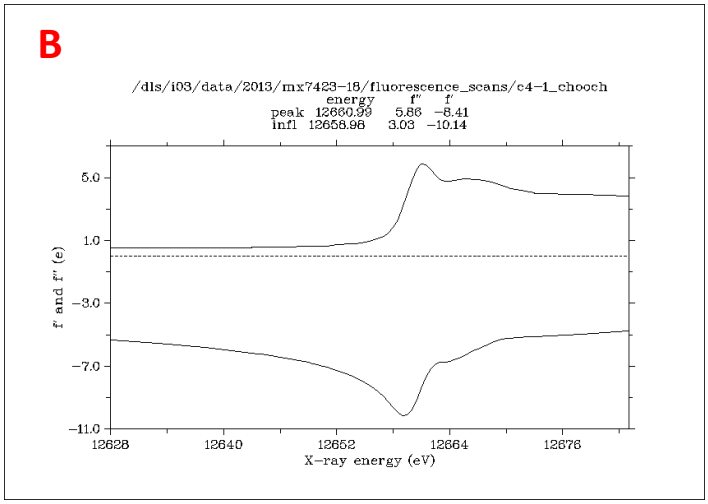


Figure 5.2.1 – Se-MET derivative crystals grown in JCSG+ C4 contain a heavy atom sites. Panel **A** displays the best diffracting Se-MET derivative crystals, grown in sitting drop condition JCSG+ C4 (0.1 M HEPES pH 6.5, 10 % PEG 6000). Panel **B** shows the CHOOCH output of a fluorescence scan, undertaken to identify the Selenium absorption edge prior to adjusting the X-ray's wavelength to 0.97930 Å for data collection.

A HCH_03101 Se-MET derivative – C4-1 SAD

	Overall	Low	High
High resolution limit (Å)	2.79	12.49	2.79
Low resolution limit (Å)	37.70	37.70	2.87
Completeness	99.9	97.9	100
Multiplicity	24.0	15.5	25.6
I / sigma	26.1	57.1	5.1
R_{pim} (I)	0.023	0.013	0.148
R_{pim} (I+/-)	0.029	0.011	0.201
Wilson B factor (Å²)	55.03		
Anomalous completeness	100.0	100.0	100.0
Anomalous multiplicity	13.8	13.7	14.1
Anomalous correlation	0.744	0.973	-0.025
Anomalous mid-slope	1.367		
Total observations	148,215	1538	11,422
Total unique observations	6176	99	447
Unit cell dimensions			
A, B, C (Å)	52.15	52.15	163.69
α, β, γ (°)	90.0	90.0	90.0
Space group	P 4_1 2_1 2		

B HCH_03101 Se-MET derivative – C4-2 SAD

	Overall	Low	High
High resolution limit (Å)	2.64	11.82	2.64
Low resolution limit (Å)	33.62	33.62	2.71
Completeness	99.8	96.3	99.6
Multiplicity	24.2	15.6	24.7
I / sigma	27.1	60.9	4.7
R_{pim} (I)	0.021	0.013	0.162
R_{pim} (I+/-)	0.026	0.011	0.221
Wilson B factor (Å²)	50.8		
Anomalous completeness	99.8	100.0	99.3
Anomalous multiplicity	13.8	13.4	13.3
Anomalous correlation	0.786	0.986	0.013
Anomalous mid-slope	1.446		
Total observations	174,057	1734	12,519
Total unique observations	7205	111	506
Unit cell dimensions			
A, B, C (Å)	52.15	52.15	163.59
α, β, γ (°)	90.0	90.0	90.0
Space group	P 4_1 2_1 2		

C HCH_03101– Se-MET derivative – C4-C SAD

	Overall	Low	High
High resolution limit (Å)	2.65	11.85	2.65
Low resolution limit (Å)	40.91	40.91	2.72
Completeness	99.8	98.3	99.6
Multiplicity	44.8	30.5	24.8
I / sigma	33.8	76.6	4.8
R_{pim} (I)	0.016	0.009	0.158
R_{pim} (I+/-)	0.021	0.009	0.216
Wilson B factor (Å²)	51.6		
Anomalous completeness	99.8	100.0	99.3
Anomalous multiplicity	25.5	27.2	13.4
Anomalous correlation	0.853	0.980	0.053
Anomalous mid-slope	1.560		
Total observations	321,057	3505	12,438
Total unique observations	7159	115	502
Unit cell dimensions			
A, B, C (Å)	52.15	52.15	163.63
α, β, γ (°)	90.0	90.0	90.0
Space group	P 4_1 2_1 2		

Figure 5.5.2 – Data collection statistics for the Se-MET derivative. Panels **A** and **B** show the processing statistics from 2 individual 360 ^{o} datasets each collected from a single JCSG+ C4 crystal. Panel **C** shows the processing statistics produced when the two datasets above are combined as a single dataset, by re-processing them in Xia2. Interestingly outside of the increase in the anomalous values, there is a broad improvement observed in both the merging (R_{pim}) and signal to noise (I/σI) stats as a result of the increased redundancy.

Therefore it was decided to maximise the statistical significance of the poor anomalous signal available, by increasing the redundancy of the data. However, these crystals had previously identified radiation damage limitations, which precluded the use of alternative kappa angles or longer rotations to increase the redundancy possible from a single crystal. Therefore, the best option available was to combine multiple datasets. The best two datasets, selected based on their anomalous mid-slope and isomorphous unit cell dimensions were combined. This was achieved by re-processing the data collected for each crystal in Xia2 (Winter, 2010), with each independent 1800 image batch treated as a separate wave (figure 5.5.2C). Whilst this was a slightly unusual way of producing a highly redundant dataset for SAD phasing, the statistics outputted from AIMLESS indicate that this combined dataset has a positive effect on the overall quality of the data. There are clear improvements observed in both the Merging (R_{pim}) and signal to noise (i/σi) statistics, besides the marked improvement to the anomalous mid slope up from 1.45 to 1.56 in the new dataset.

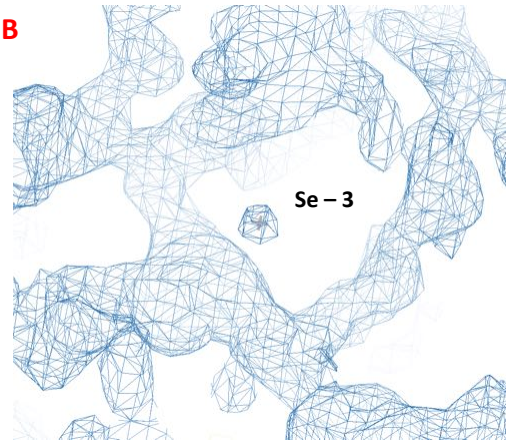
The processed combined dataset was then used to calculate the experimental phases using Autosol within the Phenix suite (Adams *et al.,* 2010). Autosol starts by determining the location of the heavy atoms within the derivative unit cell (figure 5.5.3). However, Autosol identifies 6 sites, which was unexpected as HCH_03101 only codes for 4 possible MET positions. Examination of these sites (figure 5.5.3) indicates that sites 1-3 are high occupancy with a fourth Selenium split between sites 4 and 6, which are in close proximity to one another with a combined occupancy of 0. 95. The 5th and final site has an extremely low occupancy of 0.17 and possibly represents either an error or a low occupancy location. Overall the Figure of merit calculated for these sites is 0.33, which leads onto the poor initial electron density map (figure 4.5.B) calculated using only the heavy atom sites. However, the following panel (5.5.3B) shows the same portion of electron density, post density modification in DM, highlighting how inaccurate the initial experimental phases were.

Having identified the heavy atom sites and calculated a set of experimental phases, Autosol was then used to build a model into the density-modified map. This model building is guided by prior knowledge of the primary sequence, provided by a .seq file, and the location of the heavy atom sites within this sequence, at the corresponding MET residues. Autosol outputs an incomplete HCH_03101 model, building 202 of the 234 expected residues with 152 correctly assigned in chain A and a further 50 positions built as poly-ALA backbone in chain B.

A

Selenium site	Occupancy	B-factor	X,Y,Z location
Se – 1	0.97	57.4	12.33 13.71 125.67
Se – 2	1.07	56.9	4.07 19.52 28.00
Se – 3	0.78	56.2	0.71 4.71 12.19
Se – 4	0.59	80.5	11.63 12.84 19.35
Se – 5	0.17	80.1	1.26 4.32 151.31
Se – 6	0.36	83.4	11.65 12.96 17.61

B



C

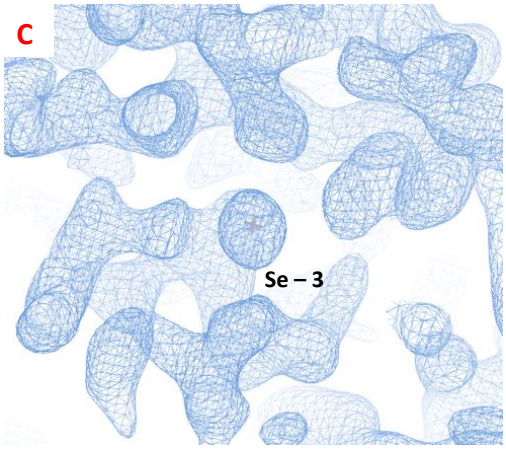


Figure 5.5.3 – Identifying heavy atom sites and constructing an initial electron density map. The Se-MET derivative of HCH_03101 encodes for 4 Selenium sites, but Autosol identifies 6. Chart A details the occupancy, temperature factor and real-space locations for each potential site. Selenium's 1, 2 and 3 couple high occupancy with low B-factors, site 5 is likely an error and sites 4 and 6 with their close proximity and otherwise high occupancies most likely constitute a single 4th selenium. The initial electron density map was calculated using phases obtained from the above sites. Panel **B** shows the electron density map calculated from the experimental heavy atom phases alone. The resulting map is of poor quality, as SAD data is rarely conclusive until some degree of density modification has been employed. Panel **C** shows the same portion of map, post density modification.

Residues built	202 / 234
Residues assigned	152
Chains / fragments	2 / 11
Map – model CC	0.77
R_{work}	0.33
R_{free}	0.41

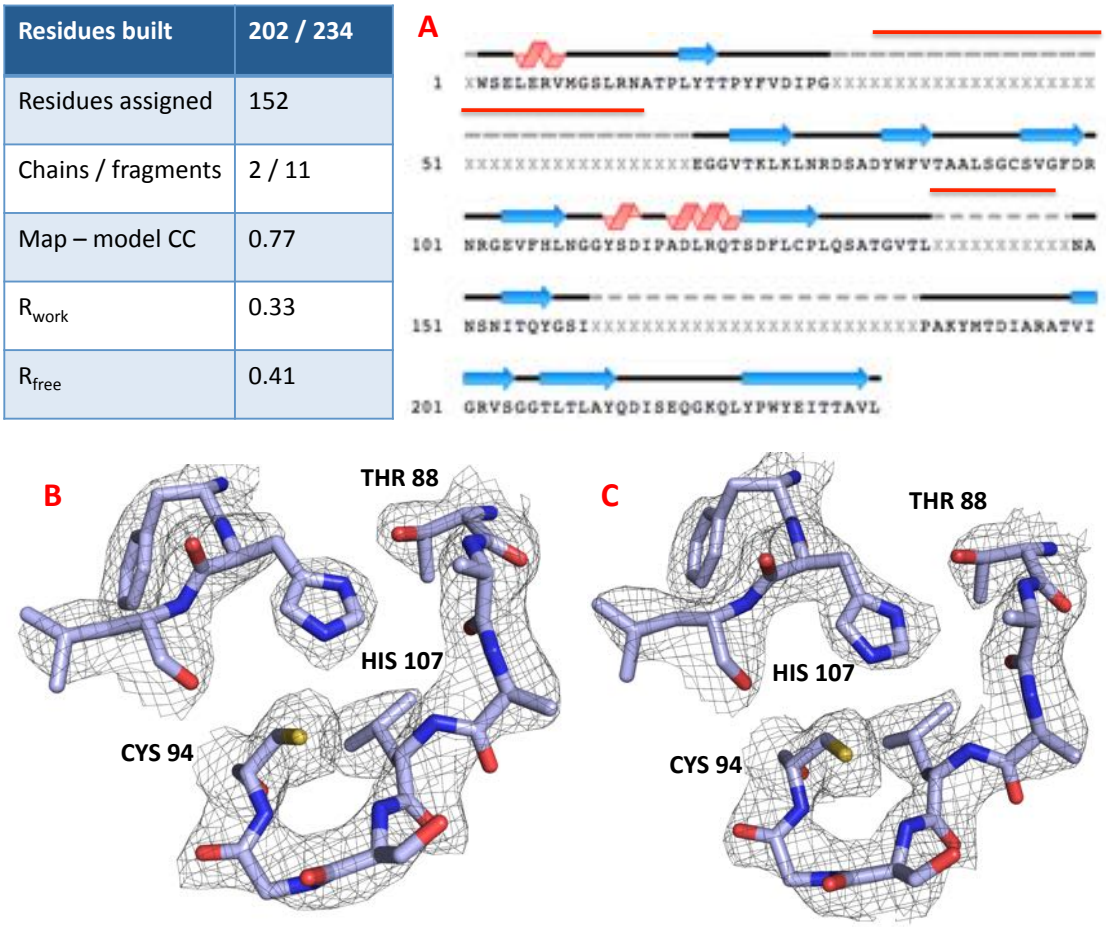


Figure 5.5.4 – Automated model building of Se-MET derivative HCH_03101 CTD 6x HIS. Having calculated experimental phases, Autosol builds a model which it refines the initial phases against prior to iterative model rebuilding. The chart to the top left hand corner summarises the progress of the automated model building. HCH_03101 is composed of 234 residues 152 of which are automatically placed, with a further 50 positions built as unassigned poly-ALA fragments. Taken together the model accounts for 86 % of the anticipated protein backbone with a map-model correlation of 0.77. Panel **A** displays the assigned residues with 3 key gaps in the model. However, 2 of these large gaps have been built in as a poly-ALA backbone in a separate chain indicated in red. Both N and C-terminal domains are built and complete, leaving a 25 residue gap unaccounted for. Panel **B** shows the active site residues THR 88, CYS 94 and HIS 108 built into the initial density modified map contoured at 1 σ. Panel **C** shows the same region after 5 cycles of model building and iterative refinement, with the map once again contoured at 1 σ. Whilst this map represents a significant improvement it is not possible to locate any unassigned residues.

Between these two chains there are 11 fragments accounting for 86 % of the anticipated protein backbone, with an map-model correlation of 0.77. Analysis of the partial model indicated that chain B had been built incorrectly. This region was entirely rebuilt using Coot (Emsley and Cowtan, 2004), but despite iterative refinement against this new model yielding significant improvements upon the initial phases, shown in Figure 5.5.4B-C, it was not possible to produce a model accounting for HCH_03101's full protein backbone.

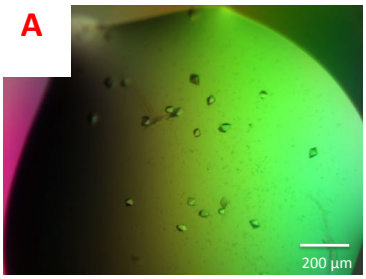
5.6 – HCH_03101 Native diffraction data

With a partial model accounting for over 80 % of HCH_03101's primary sequence, the next challenge was to optimise the diffraction quality from the native CTD 6xHIS crystals. The crystals with the best diffraction properties were harvested from a sitting drop trial in condition JCSG+ E8 (1 M di-Ammonium phosphate, 0.1 M Sodium acetate pH 4.5), detailed in figure 5.6.1. The diffraction from this particular condition was far stronger than previously observed, as shown in figure 5.6.2, it also exhibited a lower mosaic spread of 2 ^{o}. A 120 ^{o} dataset was collected from a single crystal at the Diamond synchrotron on Beam-line I03, diffracting to 1.88 Å with a P 4_1 2_1 2 space-group.

An initial electron density map was produced using phases calculated from a Molecular replacement of the incomplete derivative model into the improved native data, using PHASER-MR in the CCP4i suite (McCoy *et al.,* 2007). This route was chosen in place of phase extension because of the incomplete nature of the previous model. The resulting electron density map was substantially improved from the 2.6 Å Se-MET derivative, with substantial difference density observed at both previously unresolved fragments of the backbone and for incorrectly modelled portions of the search model, as shown in figure 5.6.3B. This superior map was subsequently used to model a complete HCH_03101 structure, with all 234 residues assigned in a single continuous chain, along with 154 water molecules and a single phosphate ion.

Optimisation to produce superior crystals was attempted around the JCSG+ E8 condition. The best crystals were grown in 0.7 M di-Ammonium phosphate, 0.1 M Sodium acetate and diffracted to 2.15 Å resolution. However, closer inspection of the diffraction patterns, figure 5.6.2C, showed that the reflections are not as well defined as previously collected datasets.

A



Condition JCSG+ E8:
1 M di-Ammonium Phosphate, 0.1 M Sodium acetate pH 4.5

B

HCH_03101 CTD 6xHIS – Native			
	Overall	**Low**	**High**
High resolution limit (Å)	1.88	8.41	1.88
Low resolution limit (Å)	51.96	51.96	1.93
Completeness	96.5	99.4	71.5
Multiplicity	8.1	6.2	7.4
I / sigma	21.9	55.0	4.3
R_{pim} (I)	0.018	0.010	0.224
Wilson B factor (Å²)	30.25		
Total observations	151,649	1808	7392
Total unique observations	18,619	290	997
Unit cell dimensions			
A, B, C (Å)	51.96	51.96	164.25
α, β, γ (°)	90.0	90.0	90.0
Space group	P 4_1 2_1 2		

Figure 5.6.1 - Crystal conditions and processing statistics for the native 6x HIS HCH_03101. Panel **A** shows the crystal droplet from which the best diffracting HCH_03101 crystals were harvested. The condition is JCSG+ E8 (1 M di-Ammonium phosphate, 0.1 M Sodium acetate pH 4.5). Chart **B** details the processing statistics for the dataset collected from crystals from JCSG+ E8, which were used to build the current best model of HCH_03101.

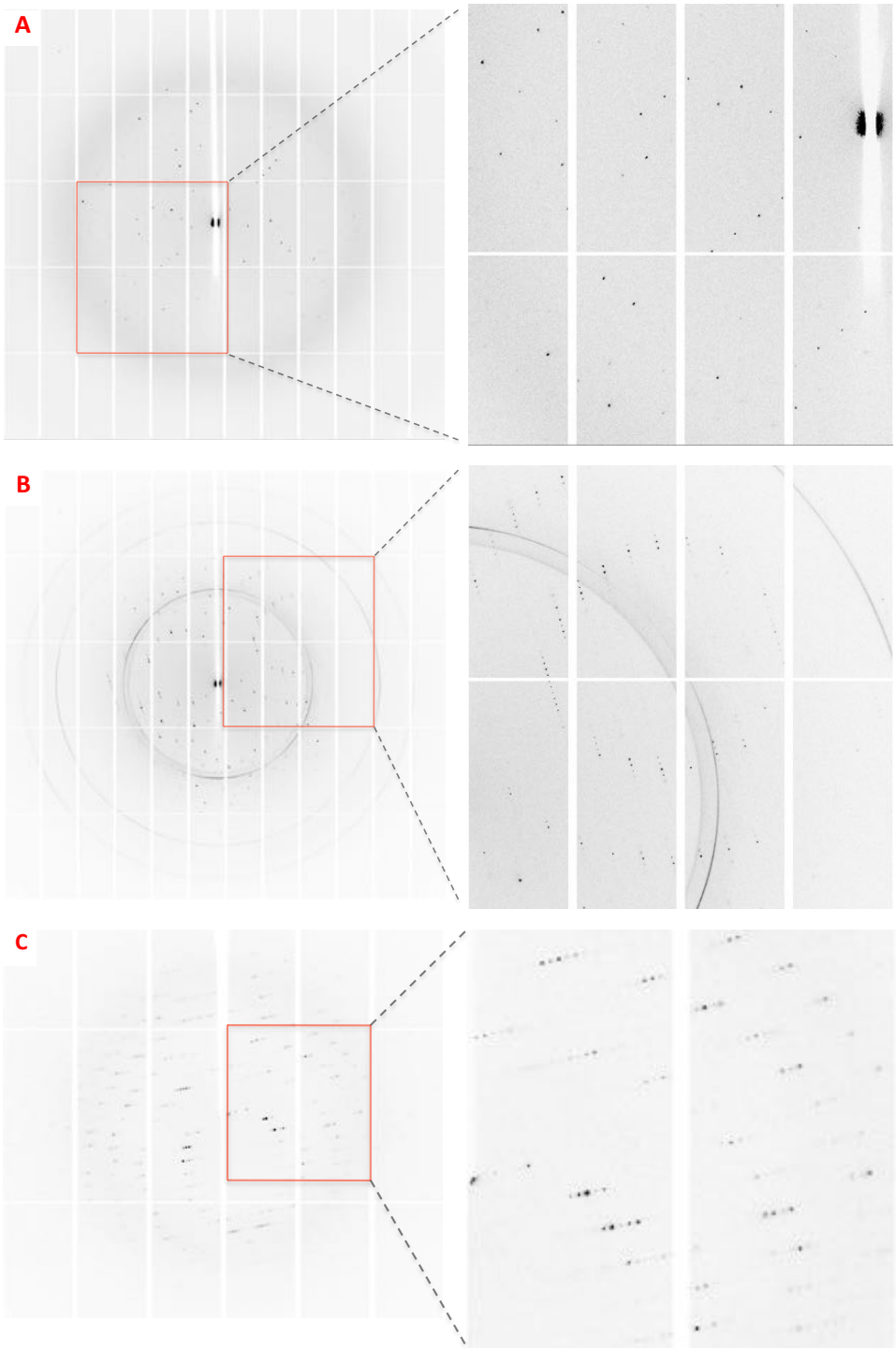


Figure 5.6.2 – Comparison of diffraction quality across the HCH_03101 crystals. Panel **A** shows a typical diffraction pattern taken from the Se-MET derivative, with panels **B** and **C** displaying diffraction from sitting drop and optimised crystals of tagged WT HCH_03101 respectively. Panel **B** is from the dataset used to model HCH_03101 and exhibits multiple ice rings.

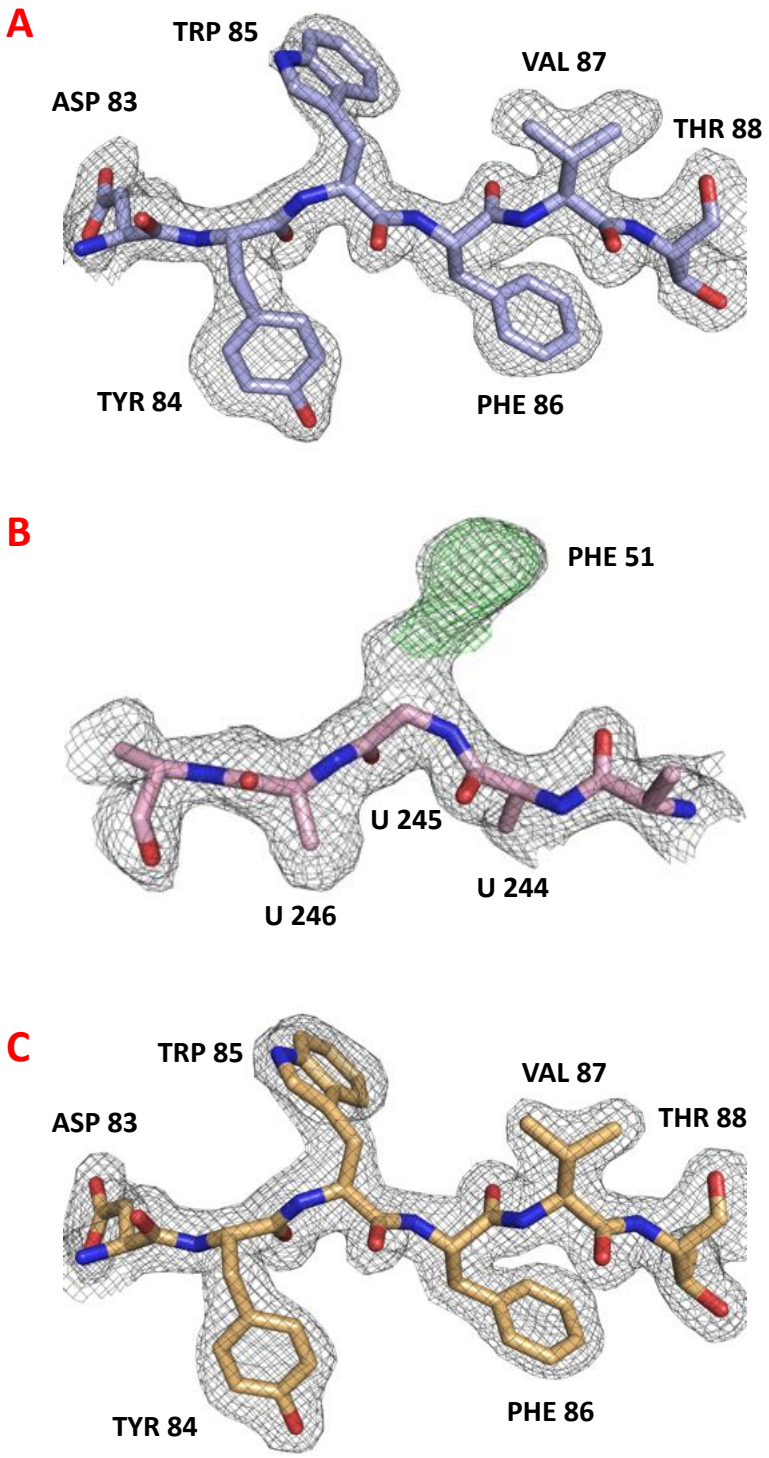


Figure 5.6.3 – Examining the initial and final electron density maps from tagged WT HCH_03101. Panel **A** shows a strongly aromatic portion of the structure between residues 83 and 88 with the initial electron density map, calculated directly after Molecular replacement, overlayed and contoured to 1 σ. Panel **B** exhibits an unassigned poly-ALA portion of chain B in the incomplete starting model, where the green Fo-Fc difference map is contoured at 3 σ provides new information, in the final model U 245 corresponds to PHE 51. Panel **C** is the same portion of the structure as panel **A** but overlaid with the final refined map contoured at 1 σ.

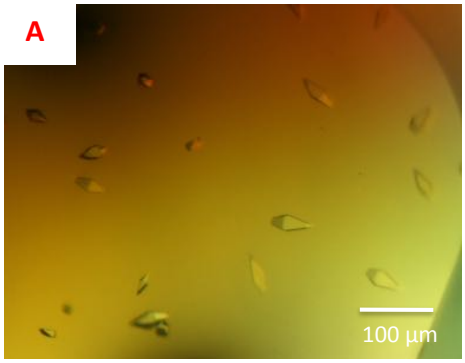
A final further attempt to produce crystals that diffracted to higher resolution, involved crystallising the WT construct. The statistics from the best data obtained from crystals of WT HCH_03101, is shown in figure 5.6.4. This data was collected from sitting drop crystals produced in JCSG+ E11 (0.16 M Calcium acetate, 0.08 M Sodium cacodylate pH 6.5, 14.4 % (w/v) PEG 8000, 20 % Glycerol). However, they only diffract to 3 Å resolution and so any resulting model would provide little insight into either the structure or function of HCH_03101.

5.7 – HCH_03101 C94S diffraction data

Data were also collected from a C94S mutant of the tagged HCH_03101 construct; this was to show that the expected inactive mutant displayed the same tertiary fold as the WT. The best diffracting crystals (figure 5.7.1), were an optimisation from the sparse matrix condition PACT E4 (0.2 M Potassium thiocyanate, 20 % (w/v) PEG 3350) grown in 0.2 M Potassium thiocyanate, 12 % (w/v) PEG 3350. A full dataset was collected at the Diamond synchrotron on beam-line I03 to a resolution of 2.08 Å, encompassing a full 360 ^{o} rotation in 0.2 ^{o} steps. These crystals display a characteristic P 4_1 2_1 2 space-group, but unlike the tagged examples the unit cell dimensions are extended by 13.6 Å along the C axis, in a similar fashion to the un-tagged crystals that diffracted to low resolution. This crystal structure has therefore been solved from a crystal displaying a different crystal form, which will be examined in chapter 6.

An electron density map was calculated for the above dataset through Molecular replacement using the finalised WT 6xHIS HCH_03101 model as a search ensemble, displayed in figure 5.7.2A. The search ensemble was not modified to incorporate the desired mutation; therefore analysis of the initial map refined against an incorrect CYS 94 model could be used to check for model bias. Figure 5.7.B shows HCH_03101's active site, the incorrect CYS 94 has not been modified and the difference density map post refinement indicates an excess of atoms modeled at the side-chains thiol termini, validating the successful mutation. The model was then modified to a SER at position 94 and fully rebuilt in Coot. Under cursory inspection there where substantial portions of the protein, in particular a 20 residue stretch between positions 165 and 185 which occupied a different conformation to the WT model. This deviation in the global structure is likely a consequence of the alternative crystal form.

A



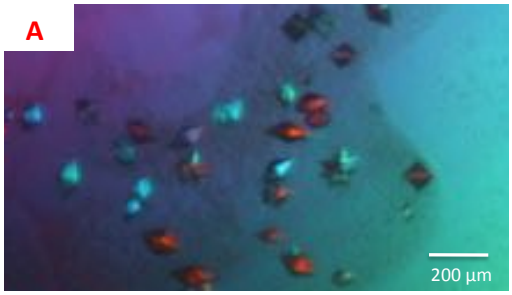
JCSG+ E11: 0.16 M Calcium acetate, 0.08 M Sodium cacodylate pH 6.5, 14.4 % (w/v) PEG 8000, 20 % Glycerol

B

HCH_03101 – Native			
	Overall	**Low**	**High**
High resolution limit (Å)	2.99	13.37	2.99
Low resolution limit (Å)	33.49	33.49	3.07
Completeness	99.9	93.9	100.0
Multiplicity	11.6	7.1	12.7
I / sigma	22.8	39.5	3.2
R_{pim} (I)	0.021	0.014	0.272
Wilson B factor (Å²)	98.371		
Total observations	62,353	578	4833
Total unique observations	5354	81	382
Unit cell dimensions			
A, B, C (Å)	51.59	51.59	176.15
α, β, γ (°)	90.0	90.0	90.0
Space group	$P\,4_1\,2_1\,2$		

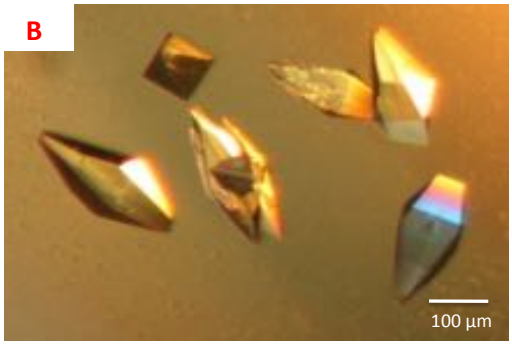
Figure 5.6.4 - Crystal conditions and processing statistics for the native non-tagged HCH_03101. Panel **A** shows the sitting drop crystals obtained from condition JCSG+ E11 (0.16 M Calcium acetate, 0.08 M Sodium cacodylate pH 6.5, 14.4 % (w/v) PEG 8000, 20 % Glycerol). These were the best diffracting crystals produced from un-tagged WT HCH_03101 protein with a 360 ^{o} dataset collected in 0.2 ^{o} slices at Diamond synchrotron, beam-line I02, to 2.99 Å resolution. Chart **B** displays the processing statistics for the above dataset. Outside of differing unit cell dimensions along the C axis the dataset was considered to be of little interest due to the low resolution achieved.

A



PACT E4: 0.2 M Potassium thiocyanate, 20 % (W/V) PEG 3350

B



Optimised: 0.2 M Potassium thiocyanate, 12 % (W/V) PEG 3350

C

HCH_03101 CTD 6xHIS C94S – Native			
	Overall	**Low**	**High**
High resolution limit (Å)	2.08	9.30	2.08
Low resolution limit (Å)	31.22	31.22	2.13
Completeness	99.9	97.9	100.0
Multiplicity	12.3	8.4	13.2
I / sigma	27.3	51.9	4.4
R_{pim} (I)	0.014	0.016	0.199
Wilson B factor (Å^2)	55.02		
Total observations	191,317	1932	12,345
Total unique observations	15,513	229	1090
Unit cell dimensions			
A, B, C (Å)	51.95	51.95	177.82
α, β, γ (°)	90.0	90.0	90.0
Space group	P 4_1 2_1 2		

Figure 5.7.1 - Crystal conditions and processing statistics for the C94S 6x HIS HCH_03101.

Panel **A** details the sitting drop crystals obtained from condition PACT E4 (0.2 M Potassium thiocyanate, 20 % (w/v) PEG 3350). This condition was optimised upon with the best diffracting crystals growing in 0.2 M Potassium thiocyanate, 12 % (w/v) PEG 3350 shown in panel **B**. Chart **C** details a full dataset incorporating 360 ^{o} of rotation with a 0.2^{o} oscillations, collected at the Diamond synchrotron on beam-line IO3 to 2.08 Å resolution. The data collected for the mutant bears a close resemblance to the un-tagged WT data (figure 5.6.4) with a C axis 13.6 Å longer than previously exhibited in the tagged WT HCH_03101 crystals (164 Å).

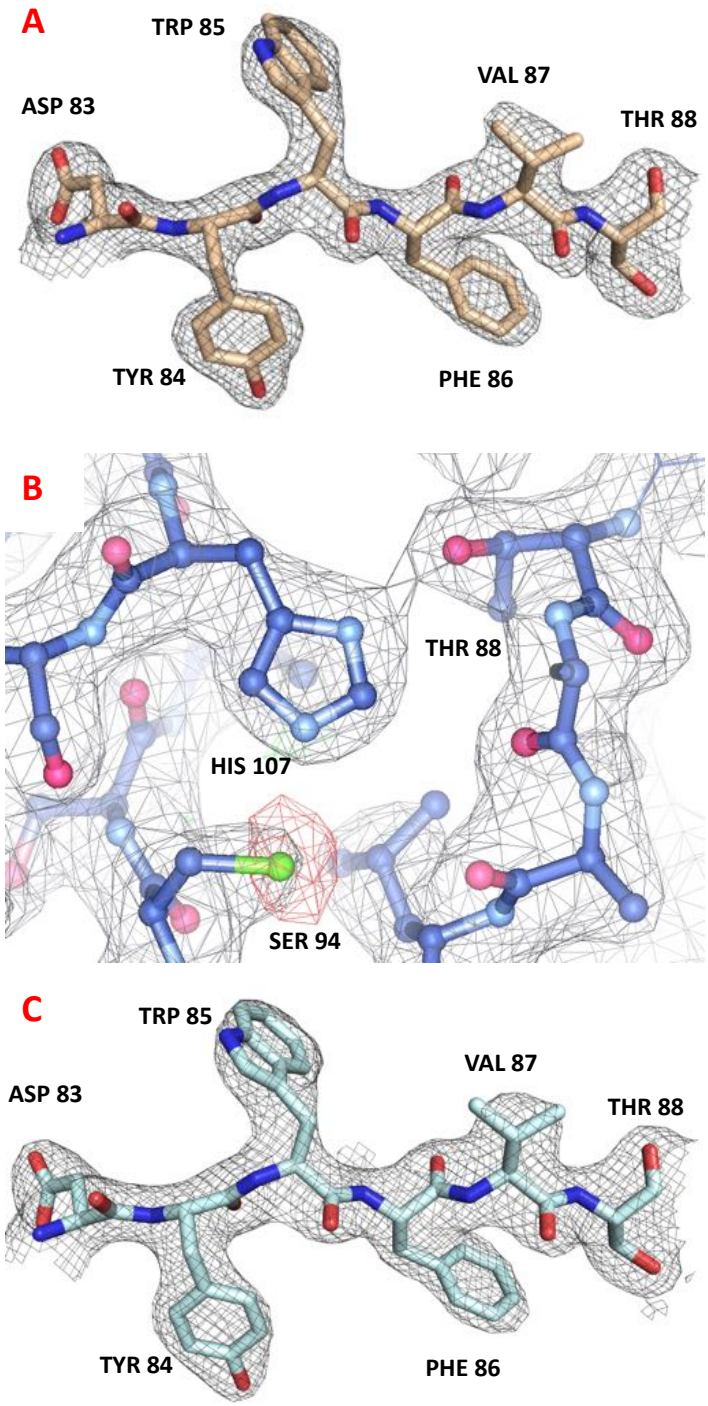


Figure 5.7.2 - Examining the initial and final electron density maps for the tagged C94S HCH_03101 mutant. Panel **A** shows a strongly aromatic portion of the structure between residues 83 and 88 placed in the initial electron density map calculated directly after Molecular replacement, contoured at 1 σ. Panel **B** displays the putative active site, with an incorrectly modelled CYS 94. However, the Fo – Fc difference map coloured red (contoured to 3 σ) is providing new information. In the finished model position 94 corresponds to a SER. Panel **C** is the same portion of the structure described in panel **A** but overlaid with the final refined map contoured at 1 σ.

5.8 Model refinement and validation

5.8.1 – Refinement and $R_{factors}$

Both the WT and C94S tagged HCH_03101 models were iteratively rebuilt and refined using Phenix Refine and Coot, until no further improvements to either the R_{factor} or the models geometry were identified. Each round of refinement carried out 5-10 cycles of maximum likelihood restrained refinement, allowing un-restrained individual isotropic B-factors. Isotropic B-factors assume that the thermal vibration of an atom within a molecule can occur in every direction around a mean position.

Both HCH_03101 structures were then refined with a constant 5 % sub set of the reflections withheld as a free R_{factor} (R_{free}), for comparison with the refined R_{factor} (R_{work}) to guard against model bias (Brunger, 1993). The R_{factor} correlates how closely the model corresponds to the experimental data, by comparing structure factors backwards calculated from the model (F_{calc}) with experimental structure factors calculated from the measured intensities (F_{obs}). Therefore, a round of refinement was considered a success provided both the R_{work} and R_{free} values reduced in tandem. Table 5.8.1 details the refinement of both the WT and C94S models of HCH_03101 CTD 6xHIS. With both of these models fully accounting for all the expected macromolecular contents of the crystal. Each model was then refined to the point where R_{work} and R_{free} no longer reduced in tandem, aiming for above average R_{factor} values for the resolution range (Kleywegt and Jones, 2002). The WT model was refined from a starting R_{work} of 0.28 to 0.22, with the R_{free} concurrently dropping from 0.33 to 0.27. Whereas, the C94S model started at an R_{work} of 0.28 reducing to 0.24, with an initial R_{free} of 0.33 dropping to 0.29.

5.8.2 – Model validation using Mol-probity

The final refined model was analysed in Mol-probity (Chen *et al.,* 2010), to highlight issues regarding steric clashes and general geometry such as poor rotamers, Ramachandran outliers and anomalous bond lengths. Mol-probity calculates two statistics quantifying model quality; both rank the query model by comparison with structures submitted to the PDB within a 0.25 Å resolution range. The first is the Clash score, which measures the number of atoms modeled within clashing distance of one another. However, this clashing analysis requires the incorporation of hydrogen atoms, which until this point have not been modeled. The second metric is the Mol-probity score, which accounts for both the steric and geometric quality of the model.

The overall refinement and validation statistics for both models are displayed in table 5.8.1. Both the WT and C94S HCH_03101 models acquit themselves favourably with Clash scores ranking within the top 2 % and Mol-probity scores in the top quarter of all structures deposited in the 1.88 and 2.09 Å (±0.25) resolution ranges respectively. Individual Ramachandran plots for each model are shown in figures 5.8.2 and 5.8.3 for the WT and C94S mutant respectively. 94 % of the WT and 93 % of the C94S mutants 234 residues occupy favorable regions. However, there is a single residue in the WT HCH_03101 structure that is modeled marginally outside the permitted region. This residue is PHE 47, comparison between the two structures of HCH_03101 reveals that this residue is located on a sharp turn between two secondary structure components that is not conserved in the mutant. Moreover deleting the entire region and refining the map returns exactly the same density features it was previously built into, culminating in the same outlier.

Model	HCH_03101 CTD WT	HCH_03101 CTD C94S
Resolution (Å)	49.54 - 1.88	31.22 - 2.08
Total unique observations	18, 552	15, 442
Protein molecules per asymmetric unit	1	1
Number of atoms	1984	1880
Number of waters	154	51
Number of ions	1 – PO4	N/A
Number of residues / modeled residues	234 / 234	234 / 234
Truncated residues (Cα)	N/A	N/A
Ramachandran favoured (%)	94 %	93 %
Ramachandran outliers (%)	0.4 %	0 %
Poor rotamers	4 – 2.1 %	2 – 1.1 %
RMSD bond length (Å)[1]	0.005	0.002
RMSD angle (°)[1]	1.04	0.73
Average B-factors (Å)	39.2	67.8
Protein	38.9	67.8
Waters	55.8	67.4
Ions	55.8	N/A
R_{work}	0.22	0.24
R_{free}	0.27	0.29
MolProbity score	1.9 (75 %)[2]	1.47 (98 %)[2]
MolProbity Clash score	5.25 (96 %)[2]	2.21 (99 %)[2]

[1] The RMSD values are for the deviation from an ideality
[2] The 100th percentile is the best among structures of comparable ± 0.25 Å resolution.

Table 5.8.1 – HCH_03101 refinement statistics. Both the WT and C94S models of HCH_03101 CTD 6xHIS were refined in Phenix Refine (Adams *et al.,* 2010) and validated using Mol-probity (Chen *et al.,* 2010).

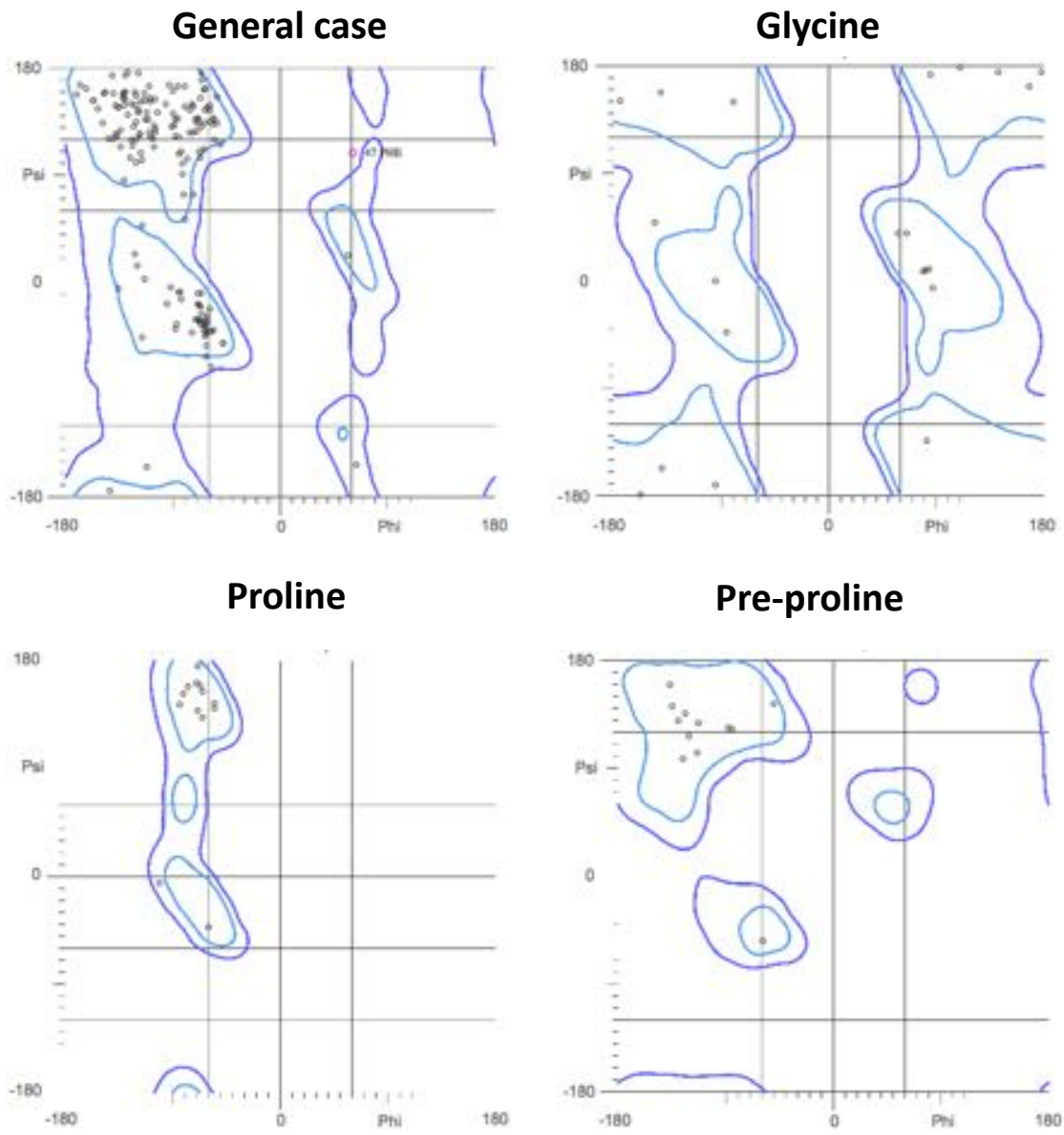


Figure 5.8.2 – Ramachandran plot for the WT 6xHIS HCH_03101 model. Ramachandran analysis of the WT HCH_03101 model, 94 % of the 234 residues have been modeled in favorable positions with only a single residue, PHE 47, built marginally outside the permitted region as shown in pink. Adapted from Mol-probity output (Chen *et al.,* 2010).

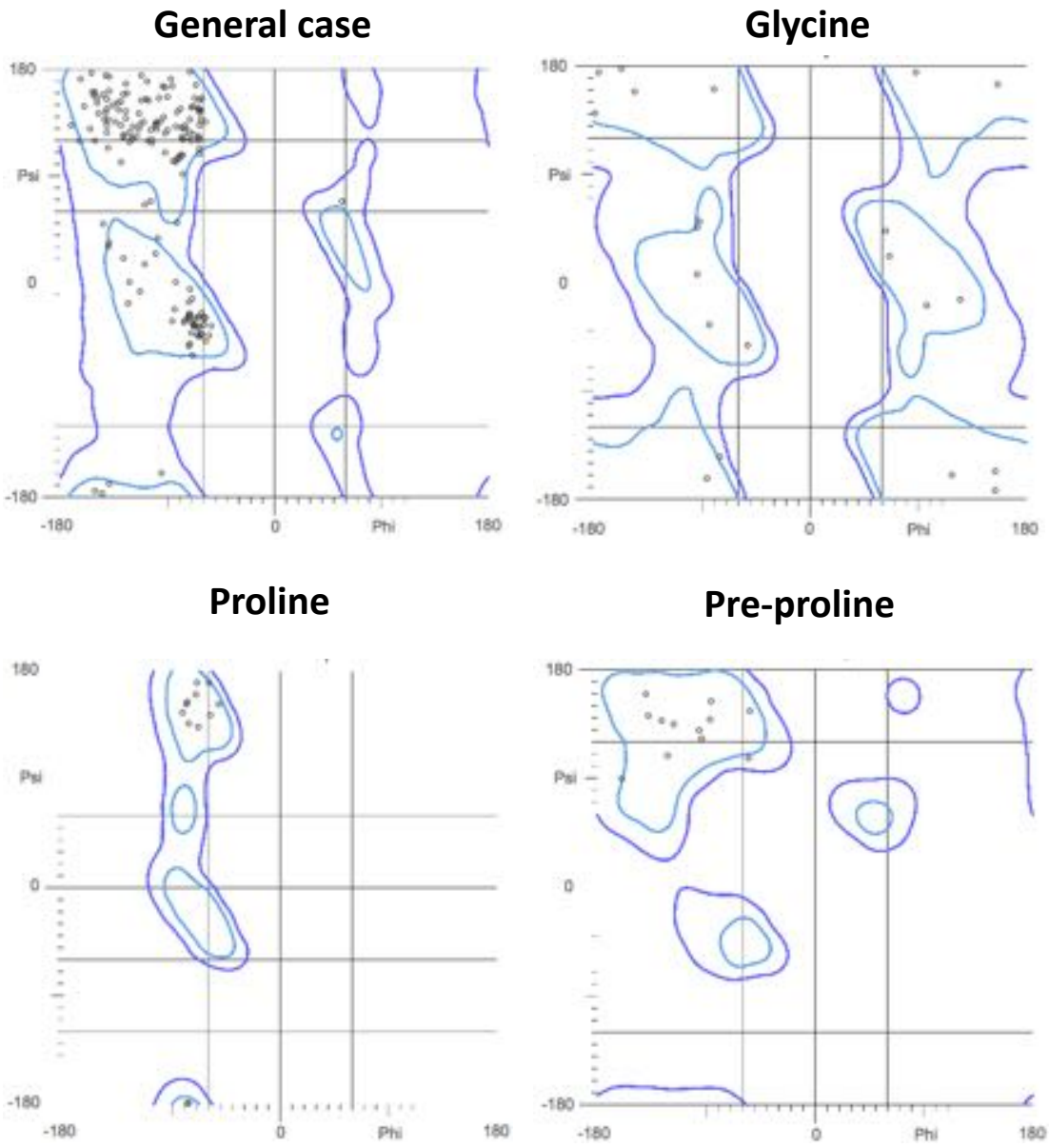


Figure 5.8.3 – Ramachandran plot for the C94S 6xHIS HCH_03101 model. Ramachandran analysis of the C94S HCH_03101 model, 93 % of the 234 residues have been modeled in favorable positions with no residues occupying the forbidden regions of steric hindrance. Adapted from Mol-probity output (Chen *et al.,* 2010).

Chapter 6: Structural analysis of HCH_03101

6.1 – HCH_03101 structural description

6.1.1 – Gross structure of HCH_03101

Two HCH_03101 structures have been solved from related crystal forms in a P 4_1 2_1 2 space group, with differing unit cell dimensions along the c axis. The highest resolution structure was solved with crystals grown from WT tagged protein (6xHIS), with a 164 Å c unit cell dimension. Whereas the active site mutant C94S, also tagged, has a c dimension of 178 Å. Cursory inspection reveals only minor differences in the overall fold of HCH_03101 between the two crystal forms, none of which alter the general flow of the tertiary structure. Therefore, for the initial description of the gross structure, only the highest resolution WT HCH_03101 structure is described.

The WT HCH_03101 structure was solved using data collected to 1.88 Å resolution (section 5.6), in a P 4_1 2_1 2 space-group, with a single molecule per asymmetric unit. HCH_03101 is a 25 KDa large and composed of a single domain, displaying a tertiary structure resembling the previously characterised Glutamine de-amidase enzymes BLF1, C-CNF1 and CheD (figure 6.1.1). The secondary structure of HCH_03101 is dominated by 14 β-sheets, which are flanked by 3 short α helices. The gross structure of HCH_03101 can be neatly divided into two portions (figure 6.1.2). The first is the β sandwich region, which is composed of two mixed β-sheets each containing 5 β strands, which is flanked on both sides by α helices. This β-sandwich is located centrally in the globular body of the protein, a trait shared with the Glutamine de-amidase family. The second region of interest is a long loop, which extends outwards from the main body of the protein through a pair of β-strands. One of the β-strands anchoring this loop in position, β16, is deeply embedded into the β-sandwich. Whilst the other strand, β14, is located in a novel position. This long loop is the most distinctive structural feature, with no equivalent within the Glutamine de-amidase family and from this point onwards it is termed the β-protrusion.

The core of HCH_03101 is composed of a β-sandwich. One sheet contains β-strands 5, 7, 16, 17 and 18, whereas the other sheet is composed of β-strands 6, 8, 9, 11 and 13 (figure 6.1.2). Within this β-sandwich the strand length varies significantly, with some strands like β16 spanning 50 Å while shorter examples like β2 only cover 8 Å. This central β-sandwich is flanked on the one side by a single helix α1, whilst the other sheet is flanked by shorter helices, α10 and α12.

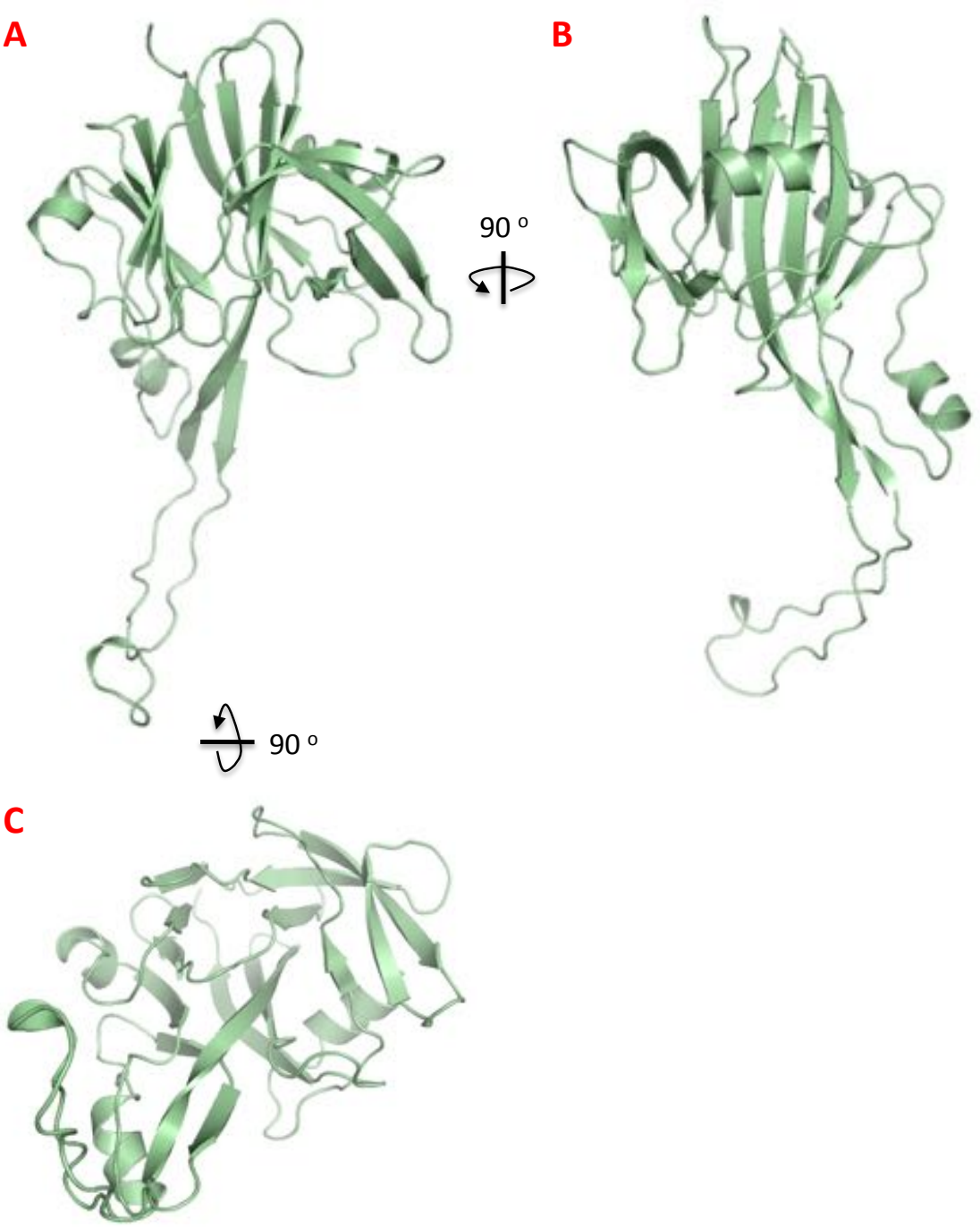


Figure 6.1.1 – 3D representation of HCH_03101. HCH_03101 is a 25 KDa single domain protein composed of 234 residues. It shares significant structural similarity with the Glutamine de-amidase super family. In common with these enzymes it exhibits a characteristic β sandwich comprising 2 mixed β-sheets, each containing 5 β-strands. This central β-sandwich is flanked by 3 α helices, 2 on one hand side and a single more substantial α helix on the other (panel **A**). The most striking structural feature is the long loop extending outwards from the globular body of the protein. Panels **A** and **B** show that this β-protrusion does not project straight outwards, instead curving back towards the globular body of the protein. Diagram produced using Pymol.

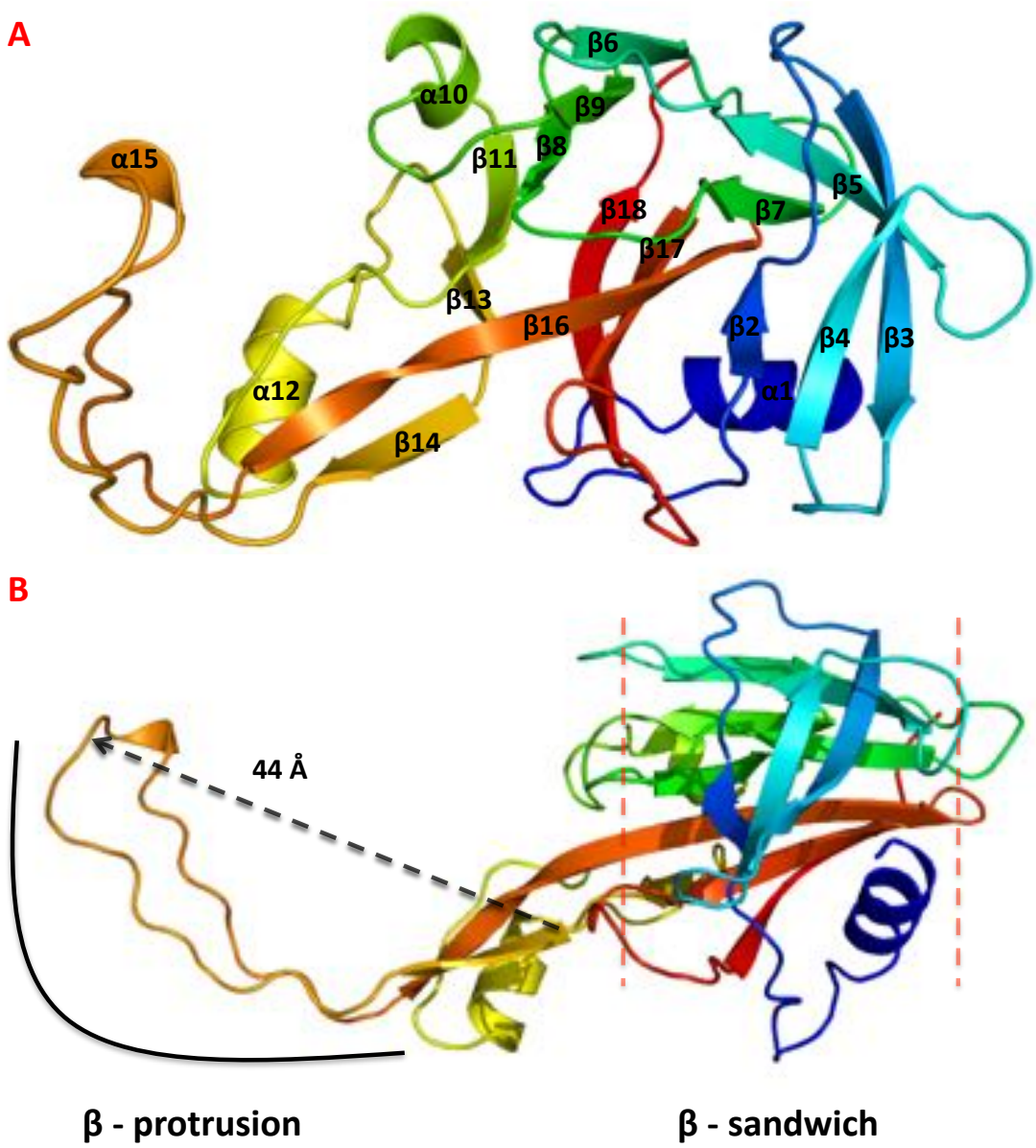


Figure 6.1.2 – Secondary structure elements of HCH_03101. Both panels **A** and **B** show a ribbon representation of HCH_03101, with rainbow colours tracking the progression from the N-terminus (blue) through to the C-terminus (red). Panel **A** shows HCH_03101 with each of its 18 secondary structure elements labelled in order from the N-terminus. HCH_03101 is composed of 14 β strands and 4 α helices, with a continuous run of β-strands between β2 and α10 covering 94 residues (which is 40 % of the total length of the protein). Panel **B** shows HCH_03101 from a side on perspective, which highlights the β-protrusion extending outwards from the main globular body of the protein. It also highlights the β-sandwich region in red dashed lines. The β-protrusion is particularly interesting as it extends for an unusually long distance. 44 Å from base to tip, equal to the width of the globular body of the protein. This protrusion initially projects out from the β-sandwich but halfway along it curves sharply by 70-80 °, with the tip of the loop facing the top face of the protein. Diagram produced using Pymol.

The second note worthy feature is the aforementioned β-protrusion, which extends from β14 (position 157) and terminates at β16 (position 195). With an extended 3-10 bend that has been interpreted as a short α helix (α15) located at the tip of the loop. In total the β-protrusion accounts for 38 residues, which is 16 % of the primary sequence. Despite only accounting for a small portion of the primary sequence this protrusion extends out an unusual distance from the globular body of HCH_03101, measuring 44 Å from TYR 157 at the base of β14 through to PRO 176 at the tip. Indeed the protrusion is longer than the globular body of HCH_03101 is wide. It is also noteworthy that the loop makes very few backbone interactions with the rest of the protein.

6.1.2 – HCH_03101 active site arrangement

The active site region of HCH_03101 has been identified through analogy with BLF1 and C-CNF1. The putative site is located on the surface of the protein with key residues placed on both sheets of the β-sandwich. The overall arrangement of the β-sandwich in HCH_03101 is reminiscent of the Glutamine de-amidase super-family. This resemblance extends to the proposed active site (figure 6.1.3), where the characteristic CYS – HIS dyad is held in a similar environment. The hypothetical CYS nucleophile is located at position 94 on a short loop between β6 and β7, supported from underneath by TYR 157 which is located on β13. TYR 157 shares a hydrogen bond with CYS 94, between its terminal carbonyl and the backbone amide of CYS 94. CYS 94 also forms a thiol-imidazolate pair with HIS 107, which is located on β9, through a hydrogen bond formed between the terminal thiol of CYS 94 and Nδ of the imidazole ring 3.2 Å away. On the other side of the imidazole ring the Nε of HIS 107 is supported from the other direction, by a hydrogen bond with the terminal carbonyl of THR 88 located 2.9 Å away on β7. If as appears likely the above residues form the basis of the active site in HCH_03101, then they are arranged in the same orientation as the previously characterised Glutamine de-amidase enzymes.

6.1.3 – Analysis of the active site cleft in HCH_03101

The hypothetical active site in HCH_03101 is located close to the surface. However, out of the three conserved residues shared between HCH_03101, BLF1 and C-CNF1 (section 2.5.5) only CYS 94 is solvent accessible (figure 6.1.4). The active site cleft is a deep circular crater surrounded by charged ridges, the top ridge is negative, whilst the bottom edge is positively charged.

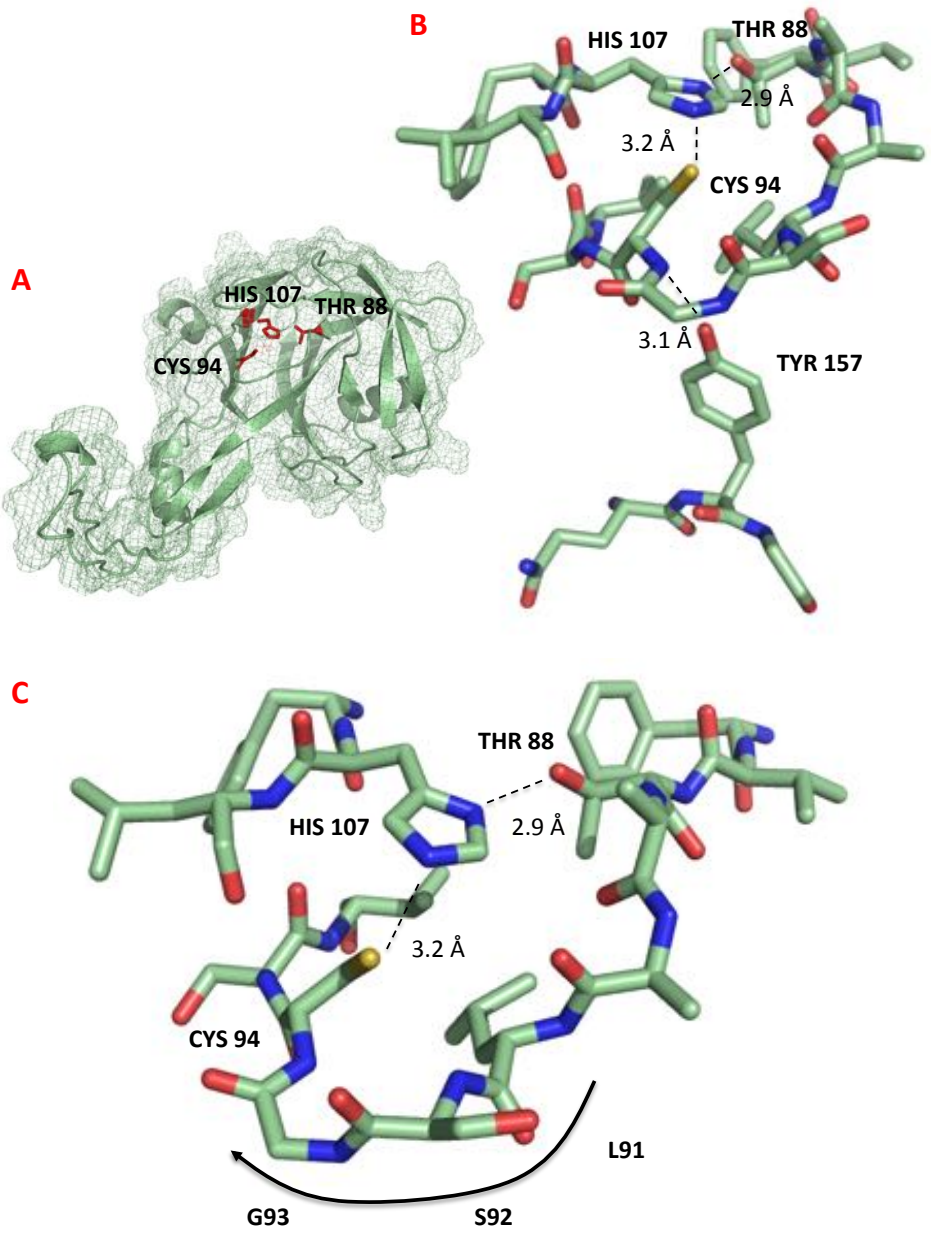


Figure 6.1.3 – Examining the active site of HCH_03101. Panel **A** is a surface mesh diagram of HCH_03101 with the proposed catalytic CYS-HIS dyad and nearby THR highlighted in red. Panel **B** zooms into the active site showing the entire region, with panel **C** focusing on the catalytic dyad. The essential CYS 94 position is located on a loop, a trait shared with the other Glutamine de-amidase enzymes. The active site contains a dyad believed to be catalytically essential; produced by the formation of a thiol-imidazolate pair between CYS 94 and HIS 107. This dyad is supported by a hydrogen bond shared between the thiol of CYS 94 and the imidazole ring Nδ. The loop containing CYS 94 is supported from underneath by a hydrogen bond between the side chain carbonyl of TYR 157 and the backbone amide of CYS 94 (panel **B**). Whilst HIS 107 is oriented in position to interact with CYS 94 through a hydrogen bond between the side chain carbonyl of THR 88 and Nε of HIS 107 (panel **C**). Diagram produced in Pymol.

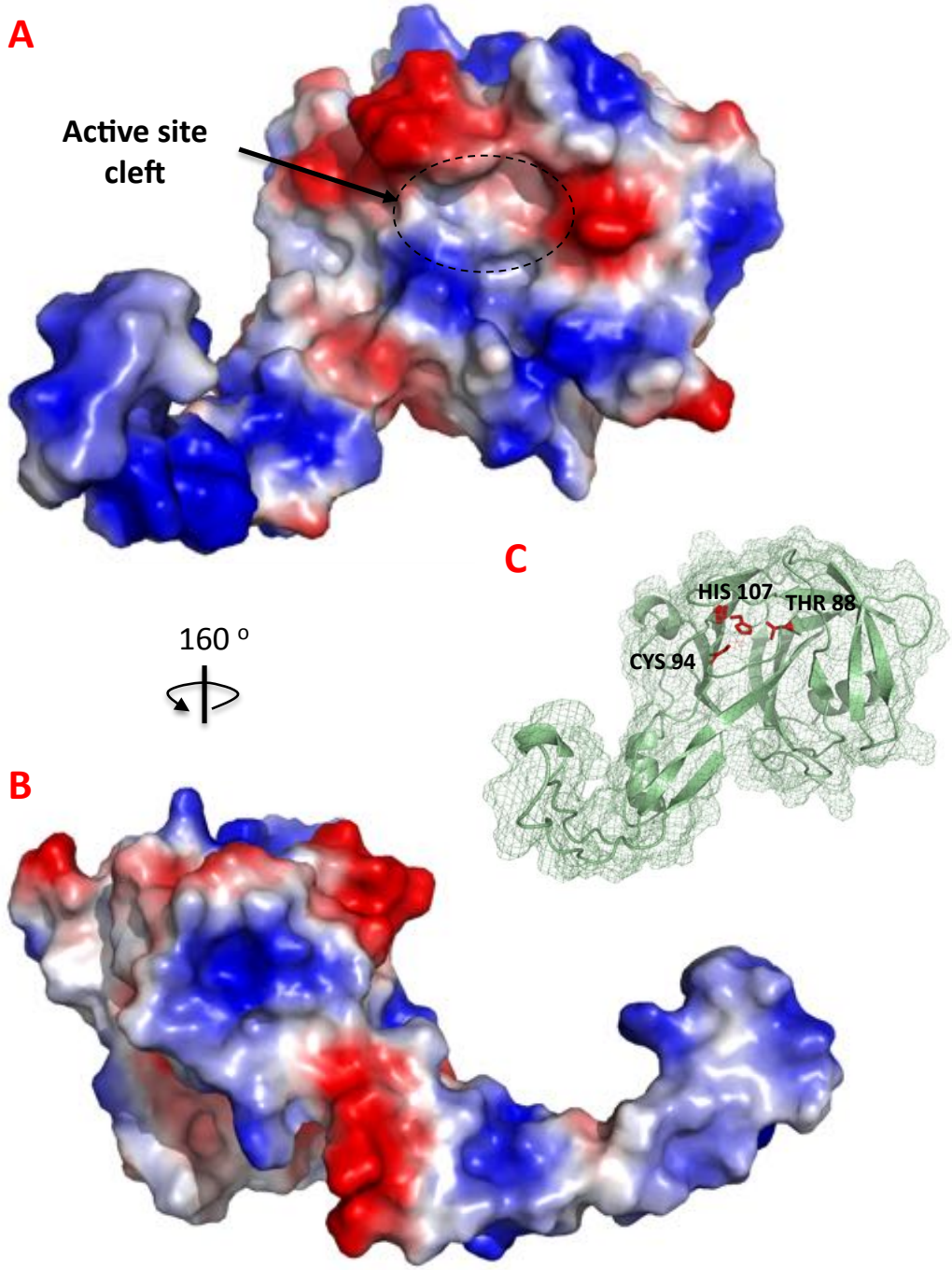


Figure 6.1.4 – The active site cleft in HCH_03101. Panels **A** and **B** are electrostatic surface charge diagrams of the front and rear faces of HCH_03101, with red and blue regions indicating negative and positive surface charges respectively. HCH_03101 exhibits a deep active site cleft, identified by the black dashed circle, located within a circular crater surrounded by a charged ridge. However, this ridge is not uniformly charged, with the top and bottom edges displaying different charges. Unlike the ridge surrounding it the active site itself is only moderately charged with a weakly negative surface. Panel **C** is a surface mesh representation of HCH_03101, with the active site residues THR 88, CYS 94 and HIS 107 coloured red. This mesh diagram shows that of the conserved active site residues, only CYS 94 is surface accessible. Diagram produced in Pymol.

At the bottom of this crevice the active site is held in a mild negatively charged environment. The distinctive β-protrusion is located in front of the active site, separated by 30 ~Å, with a largely positive charge distribution. This close proximity matched with the curvature of the protrusion, pointing it towards the active site, suggests it may play a role in substrate binding.

6.1.4 Regions of disorder

The majority of HCH_03101 was straightforward to model, owing to its highly ordered β-sandwich tertiary structure. However, there were three regions that proved challenging to interpret. These regions included the longer loops between the N-terminus and β-sandwich, particularly the loop between α1 and β2 and the entire β-protrusion. The aforementioned loops whilst challenging to interpret, were eventually modelled giving B-factors only slightly above the average (39 $Å^2$).

However, the initial map used to model the β-protrusion (figure 6.1.5A) did not provide density information for a continuous polypeptide backbone, requiring several rounds of model building and refinement to resolve the necessary density (figure 6.1.5B). In the final model the peptide backbone is well supported by extensive density evidence, with clearly defined density features representative of backbone carbonyl moieties. The same cannot be said for the side chains, with several positions displaying poor density and high B-factors post the Cβ atom (figure 6.1.5C). This disparity suggests that the β-protrusion is either a flexible component of the protein, or located in poorly supported region of the crystal lattice.

6.2 – Comparing HCH_03101 with the characterised Glutamine de-amidase enzymes

To determine structural relatives of HCH_03101, it was compared with all the structures deposited in the PDB using the Dali-lite server (Rosenstrom, 2010). The top 40 structural matches are displayed in figure 3.2.1.

6.2.1 – Comparing HCH_03101 with BLF1

Dali-lite structure comparison indicates that HCH_03101 is most closely related to BLF1 (chapter 1.4). Inspection of the gross structural alignment (figure 6.2.2) shows that BLF1 and HCH_03101 share a conserved β-sandwich fold. With their respective secondary structures following an identical route with a RMSD of 2.9 Å along the entire region.

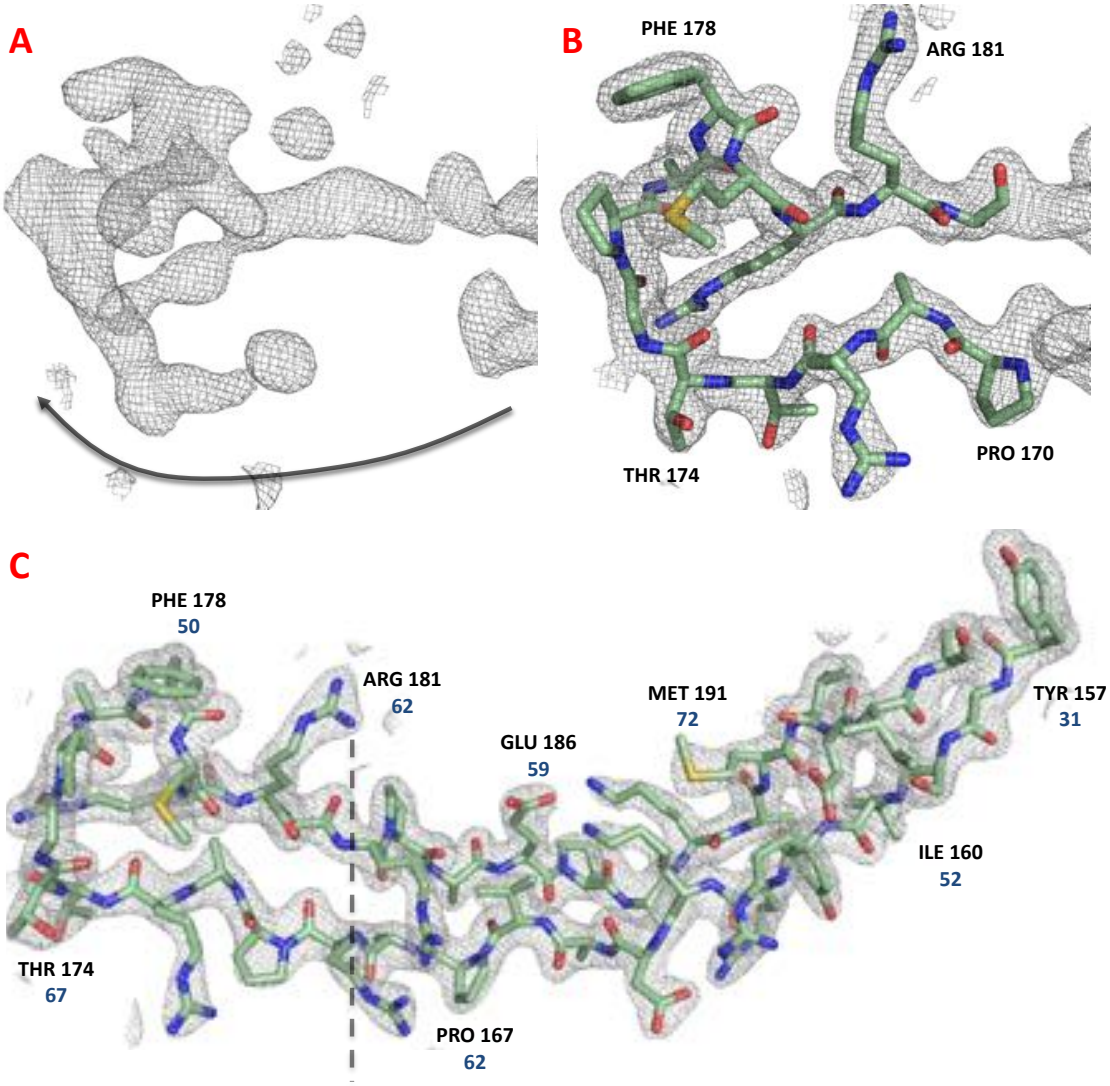
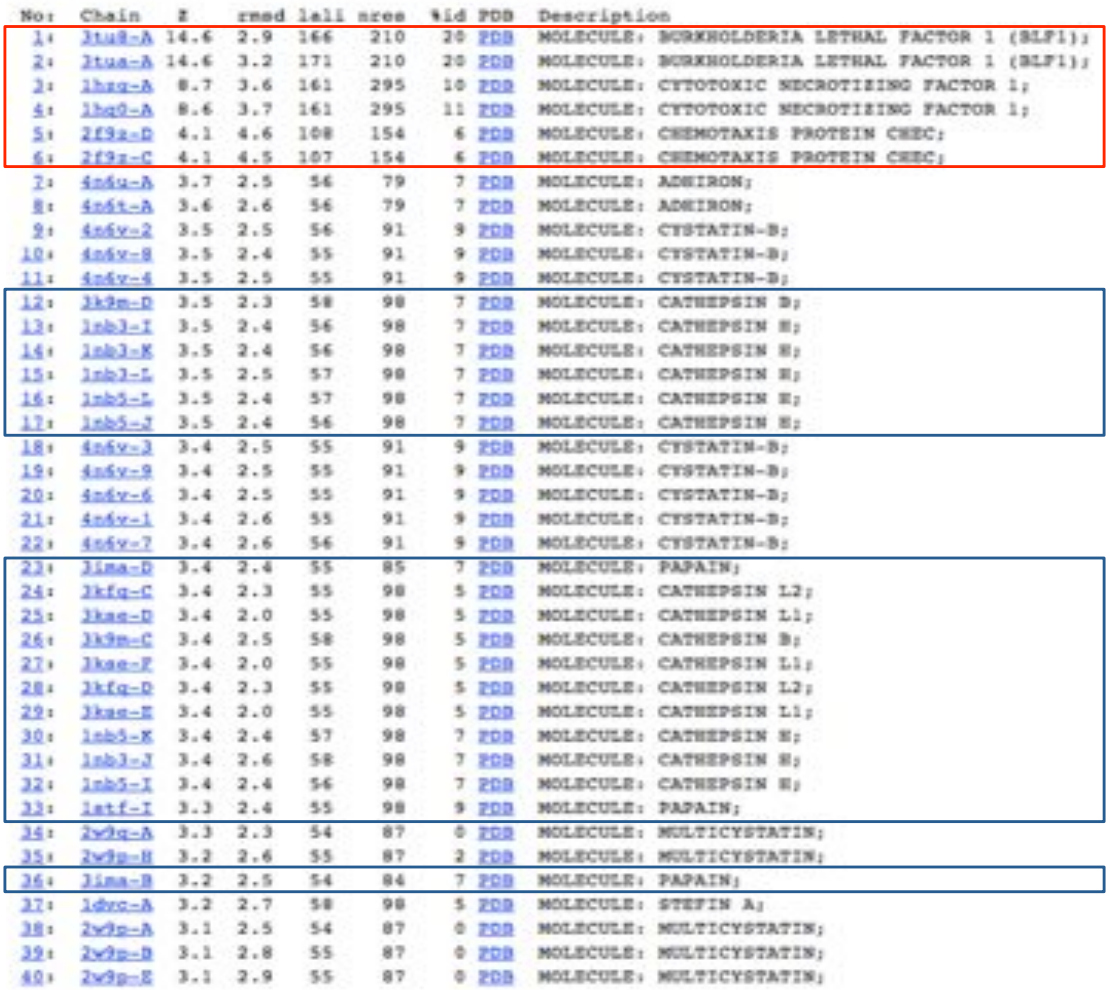


Figure 6.1.5 - Modelling the β-protrusion of HCH_03101. The region of HCH_03101 that posed the greatest challenge to model was the β-protrusion. Panel **A** shows the electron density available to build the furthest extremity of the protrusion contoured at 1 σ, directly after molecular replacement. This density provides ample evidence of the protrusions turning point, but little information for the backbone leading up to it (black arrow). Panel **B** is the same region, but with the refined model overlaid with its electron density map contoured at 1 σ. Placement of the polypeptide backbone within this improved map is secure, with well-distinguished density features corresponding to backbone carbonyl groups. Panel **C** zooms out to display the whole β-protrusion, with the map contoured at 1 σ. A selection of residues has been highlighted, with their highest atomic B-factor displayed in blue. This diagram shows that even the side chains built into well-defined electron density exhibit higher than average (39 $Å^2$) B-factors. Diagram produced in Pymol.



No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
1:	3tu8-A	14.6	2.9	166	210	20	PDB	MOLECULE: BURKHOLDERIA LETHAL FACTOR 1 (BLF1);
2:	3tua-A	14.6	3.2	171	210	20	PDB	MOLECULE: BURKHOLDERIA LETHAL FACTOR 1 (BLF1);
3:	1hzq-A	8.7	3.6	161	295	10	PDB	MOLECULE: CYTOTOXIC NECROTIZING FACTOR 1;
4:	1hq0-A	8.6	3.7	161	295	11	PDB	MOLECULE: CYTOTOXIC NECROTIZING FACTOR 1;
5:	2f9z-D	4.1	4.6	108	154	6	PDB	MOLECULE: CHEMOTAXIS PROTEIN CHEC;
6:	2f9z-C	4.1	4.5	107	154	6	PDB	MOLECULE: CHEMOTAXIS PROTEIN CHEC;
7:	4n6u-A	3.7	2.5	56	79	7	PDB	MOLECULE: ADHIRON;
8:	4n6t-A	3.6	2.6	56	79	7	PDB	MOLECULE: ADHIRON;
9:	4n6v-2	3.5	2.5	56	91	9	PDB	MOLECULE: CYSTATIN-B;
10:	4n6v-8	3.5	2.4	55	91	9	PDB	MOLECULE: CYSTATIN-B;
11:	4n6v-4	3.5	2.5	55	91	9	PDB	MOLECULE: CYSTATIN-B;
12:	3k9m-D	3.5	2.3	58	98	7	PDB	MOLECULE: CATHEPSIN B;
13:	1nb3-I	3.5	2.4	56	98	7	PDB	MOLECULE: CATHEPSIN H;
14:	1nb3-K	3.5	2.4	56	98	7	PDB	MOLECULE: CATHEPSIN H;
15:	1nb3-L	3.5	2.5	57	98	7	PDB	MOLECULE: CATHEPSIN H;
16:	1nb5-L	3.5	2.4	57	98	7	PDB	MOLECULE: CATHEPSIN H;
17:	1nb5-J	3.5	2.4	56	98	7	PDB	MOLECULE: CATHEPSIN H;
18:	4n6v-3	3.4	2.5	55	91	9	PDB	MOLECULE: CYSTATIN-B;
19:	4n6v-9	3.4	2.5	55	91	9	PDB	MOLECULE: CYSTATIN-B;
20:	4n6v-6	3.4	2.5	55	91	9	PDB	MOLECULE: CYSTATIN-B;
21:	4n6v-1	3.4	2.6	55	91	9	PDB	MOLECULE: CYSTATIN-B;
22:	4n6v-7	3.4	2.6	56	91	9	PDB	MOLECULE: CYSTATIN-B;
23:	3ima-D	3.4	2.4	55	85	7	PDB	MOLECULE: PAPAIN;
24:	3kfq-C	3.4	2.3	55	98	5	PDB	MOLECULE: CATHEPSIN L2;
25:	3kaa-D	3.4	2.0	55	98	5	PDB	MOLECULE: CATHEPSIN L1;
26:	3k9m-C	3.4	2.5	58	98	5	PDB	MOLECULE: CATHEPSIN B;
27:	3kaa-F	3.4	2.0	55	98	5	PDB	MOLECULE: CATHEPSIN L1;
28:	3kfq-D	3.4	2.3	55	98	5	PDB	MOLECULE: CATHEPSIN L2;
29:	3kaa-E	3.4	2.0	55	98	5	PDB	MOLECULE: CATHEPSIN L1;
30:	1nb5-K	3.4	2.4	57	98	7	PDB	MOLECULE: CATHEPSIN H;
31:	1nb3-J	3.4	2.6	58	98	7	PDB	MOLECULE: CATHEPSIN H;
32:	1nb5-I	3.4	2.4	56	98	7	PDB	MOLECULE: CATHEPSIN H;
33:	1atf-I	3.3	2.4	55	98	9	PDB	MOLECULE: PAPAIN;
34:	2w9q-A	3.3	2.3	54	87	0	PDB	MOLECULE: MULTICYSTATIN;
35:	2w9p-H	3.2	2.6	55	87	2	PDB	MOLECULE: MULTICYSTATIN;
36:	3ima-B	3.2	2.5	54	84	7	PDB	MOLECULE: PAPAIN;
37:	1dvc-A	3.2	2.7	58	98	5	PDB	MOLECULE: STEFIN A;
38:	2w9p-A	3.1	2.5	54	87	0	PDB	MOLECULE: MULTICYSTATIN;
39:	2w9p-B	3.1	2.8	55	87	0	PDB	MOLECULE: MULTICYSTATIN;
40:	2w9p-E	3.1	2.9	55	87	0	PDB	MOLECULE: MULTICYSTATIN;

Figure 6.2.1 – Structural comparison of HCH_03101 with structures deposited in the PDB. A Structural comparison against the PDB was undertaken using Dali-Lite, to identify proteins that share structural similarity with HCH_03101. The top 40 hits are shown above, with the currently characterised Glutamine de-amidases BLF1, C-CNF1 and CheD ranking highest (red). BLF1 is the closest identified structural match for HCH_03101, with 166 of its 211 residues aligning to within 2.9 Å RMSD, with 20 % sequence conservation. The remaining Glutamine de-amidase enzymes also aligned with HCH_03101, particularly in the β-sandwich region. 161 / 295 C-CNF1 residues align to within 3.6 Å RMSD, with 108 / 154 CheD residues aligning to within 4.5 Å RMSD. However, despite similar levels of structural conservation neither C-CNF1 nor CheD share the same level of sequence conservation with HCH_03101 at 10 % and 6 % respectively. Interestingly the Papain like family of Cysteine proteases is also identified as a possible match (blue). However, the Cysteine protease inhibitors Cystatin and Stefin are also identified. Inspection of the Papain matches shows that the Cystatin subunits, included in the co-ordinate file, are the molecules aligning with HCH_03101. Cystatin aligns with HCH_03101 along the NTD α-helix and a single sheet of the β-sandwich, with none of the catalytic residues conserved. Diagram adapted from the Dali-lite output (Rosenstrom, 2010).

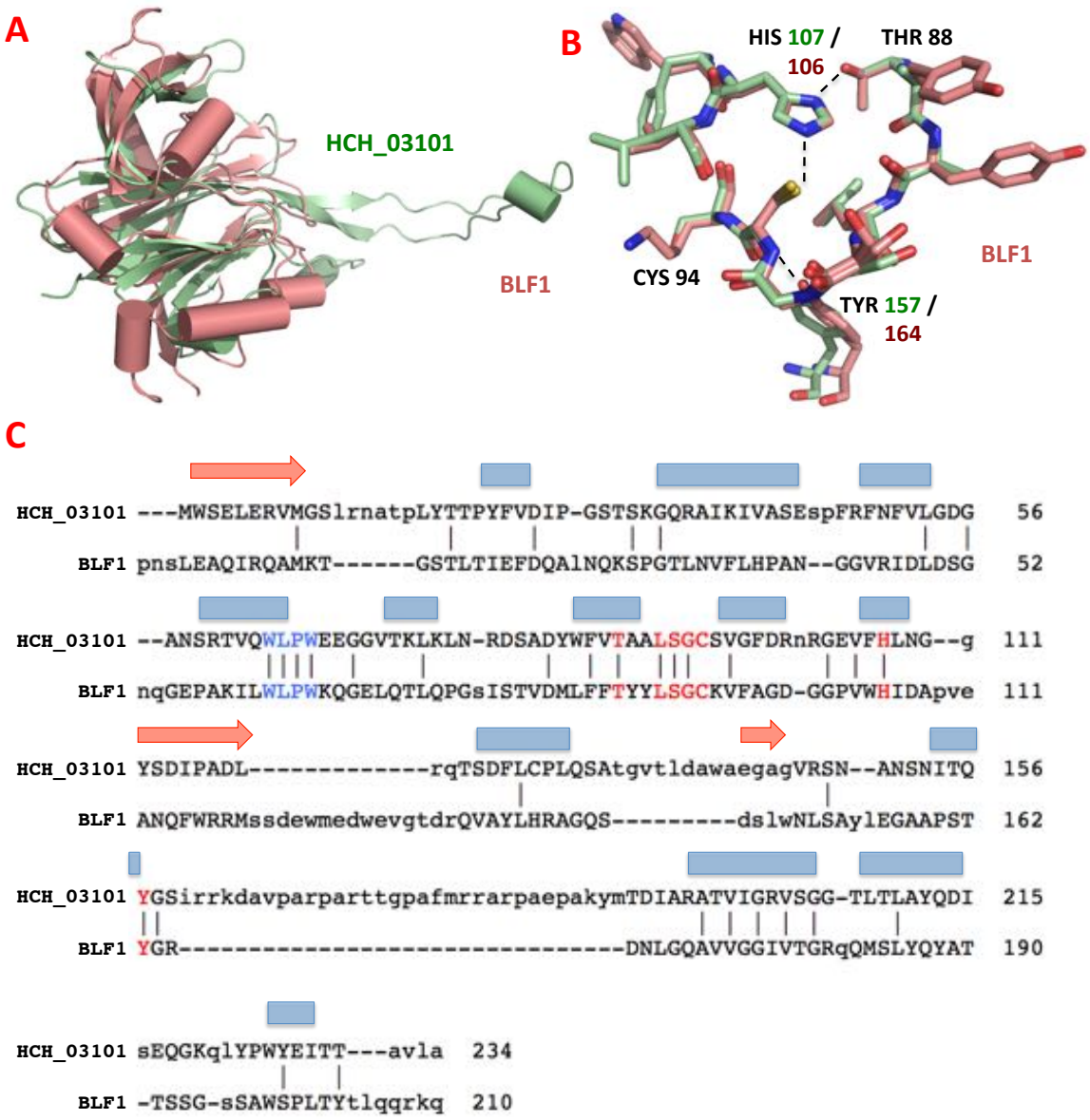


Figure 6.2.2 – Structural alignment of HCH_03101 with the Glutamine de-amidase toxin BLF1. Panel **A** shows the structural alignment of HCH_03101 (green) with BLF1 (red). Both proteins exhibit a conserved β-sandwich, which align with one another closely. However, the flanking α helices are located in non-conserved locations. The alignment also highlights the novel nature of the β-protrusion. With BLF1 not displaying either the protrusion or the β-strands forming the base. Panel **B** zooms into the active site, showing the structural conservation of the catalytic CYS-HIS dyad and its surrounding residues. Panel **C** is a structure-based sequence alignment. The blue bars represent β-strands and the red arrows α helices shared in common. A short region of continuous sequence conservation has been identified (blue), giving a WLPW motif. The conserved active site residues are highlighted in red, displaying how close the novel WLPW motif lies in relation to the current search motif. Diagram produced using Lsqkab alignments and Pymol.

However, there are also clear differences between these two proteins, particularly the flanking α helices and connecting loops. With the only structurally conserved loop lying between β7 and β8, which contains the conserved LSGC search motif. This structural alignment also highlights the novel nature of the β-protrusion, which represents a major insertion in HCH_03101.

BLF1 consists of 211 residues whilst HCH_03101 has 234. In both cases the active site is located between residues 88-107. There are several short insertions and deletions between these two enzymes, but in both cases the putative catalytic CYS residue is located at position 94. Unsurprisingly, given the strong conservation of the β-sandwich between BLF1 and HCH_03101, the aligned active sites display significant levels of sequence and structural similarity (figure 6.2.2B). This level of conservation also extends to the location and orientation of catalytic side chains, particularly the CYS-HIS dyad which is held in an identical orientation between both proteins. The obvious conservation of catalytic residues, which are held in functionally active orientations shared with BLF1, strongly suggests that HCH_03101 will exhibit Glutamine de-amidase activity.

The sequence conservation shared between HCH_03101 and BLF1 across the aligned β-sandwich region is 20 %. However, this conservation is not just localised at the active site, but spread broadly across the whole of HCH_03101, apart from at the novel β-protrusion where no conservation is observed. Given the sequence and structure similarities between BLF1 and HCH_03101, it is clear they are related. The major question at this stage, especially given the strong sequence conservation, is whether HCH_03101 is a functional homologue of BLF1.

Given the significant sequence conservation observed with BLF1, HCH_03101 may be helpful in determining a better Glutamine de-amidase search motif in the hunt for further examples. There is a single portion of continuous sequence similarity observed close to the active site in both enzymes, between residues 64 -67 giving a WLPW motif (figure 6.2.2C, blue). This motif is structurally conserved in both HCH_03101 and BLF1 (figure 6.2.3A). This is of interest because these residues are located on a β-strand directly above the active site and are solvent accessible (figure 6.2.3B-C). The close proximity to the active site, in particular the solvent accessible cleft, suggests that the conserved WLPW region may play a role in substrate binding.

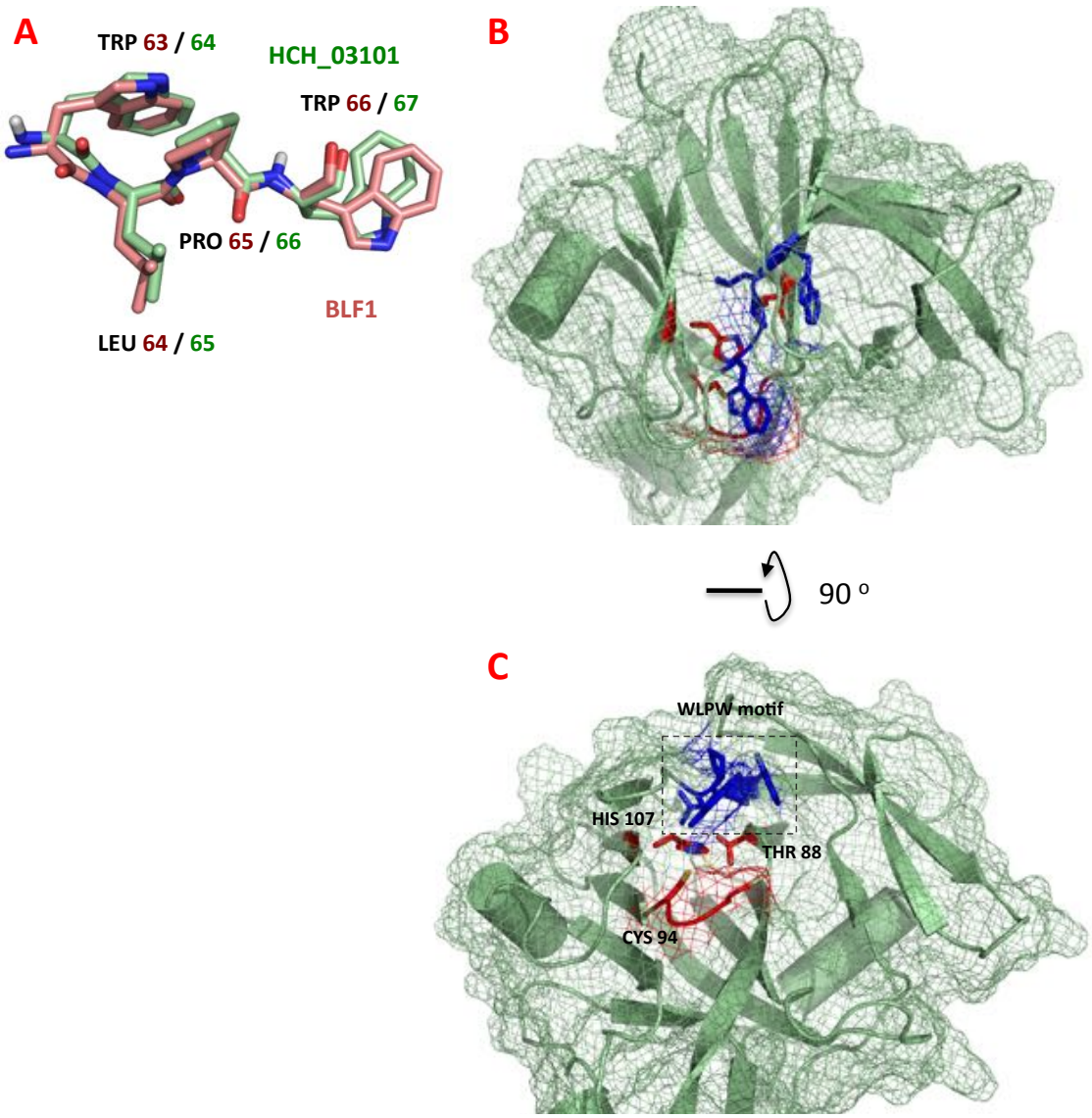


Figure 6.2.3 – Inspection of the conserved WLPW motif. Panel **A** is a structural alignment of HCH_03101 and BLF1, displaying the WLPW motif between positions 63-67. This alignment shows that the motif is both sequentially and structurally conserved. Panels **B** and **C** display a surface mesh representation of the top and front faces of HCH_03101 respectively. The conserved WLPW motif is highlighted in blue and the active site residues THR 88, CYS 94 and HIS 107 are coloured red. This novel motif sits directly above the active site and forms no explicit interactions with any of the functional residues. However, it is located close to the active site in a solvent accessible location, which may indicate that it plays a role in substrate binding. Diagram produced in Pymol.

Electrostatic surface diagrams (figure 6.2.4) show that the active site cleft in HCH_03101 is located within a small but deep circular crevice. Surrounded by a strongly charged ridge with both positive (bottom) and negatively (top) charged edges. BLF1 on the other hand has a far shallower cleft, which is surrounded by a narrow neutrally charged ridge. These diagrams suggest, that despite significant sequence conservation surrounding the active site, HCH_03101 and BLF1 likely bind to differing substrates.

One of the curious features of BLF1 is that it lacks a signal peptide or any alternative components to enable secretion. This trait is shared with HCH_03101, which may provide an insight into the lifestyle of *Hahella. Chejuensis*. BLF1 is produced by the intracellular pathogen *B. pseudomallei* and is believed to be secreted post internalisation. If HCH_03101 is a toxin, then it is a distinct possibility that *H. chejuensis* is also an intracellular pathogen. This is significant because *H.chejuensis* has previously been characterised as an extracellular bacterium, which secretes algaecidal secondary metabolites (Lee *et al.,* 2001; Jeong *et al.,* 2005). Therefore, if HCH_03101 is as critical to this organism's pathogenesis, as BLF1 is for *B. pseudomallei,* then this discovery could represent a major step towards characterising *H. chejuensis* a poorly understood marine bacterium.

6.2.2 – Comparing HCH_03101 with the remaining Glutamine de-amidase family members

The Dali-lite search against HCH_03101 (figure 6.2.1) also identified both the remaining Glutamine de-amidase enzymes, albeit with reduced sequence similarity. C-CNF1 (chapter 1.2) when aligned with HCH_03101 matches across 161 of its 295 residues to 3.7 Å RMSD, accounting for the majority of the β-sandwich. However, in comparison with BLF1, it shares significantly less sequence similarity with HCH_03101 at 10 %. Yet despite lower sequence conservation there is a strong level of structural similarity, particularly within the active site region where the backbone aligns closely (figure 6.2.5B). When structure-based sequence alignments of C-CNF1 with both BLF1 (chapter 2.4) and then HCH_03101 are compared the level of conservation is equivalent across both examples. This alongside strong sequence conservation with BLF1, further suggests that HCH_03101 is potentially a member of the Glutamine de-amidase toxin sub-family.

The third and final Glutamine de-amidase enzyme available for comparison is CheD (chapter 1.3), a non-toxic distant relative of BLF1 and C-CNF1. Of the 154 residues present in CheD, 108 align with HCH_03101, accounting for 70 % of its primary sequence and the entire β-sandwich region (figure 6.2.6).

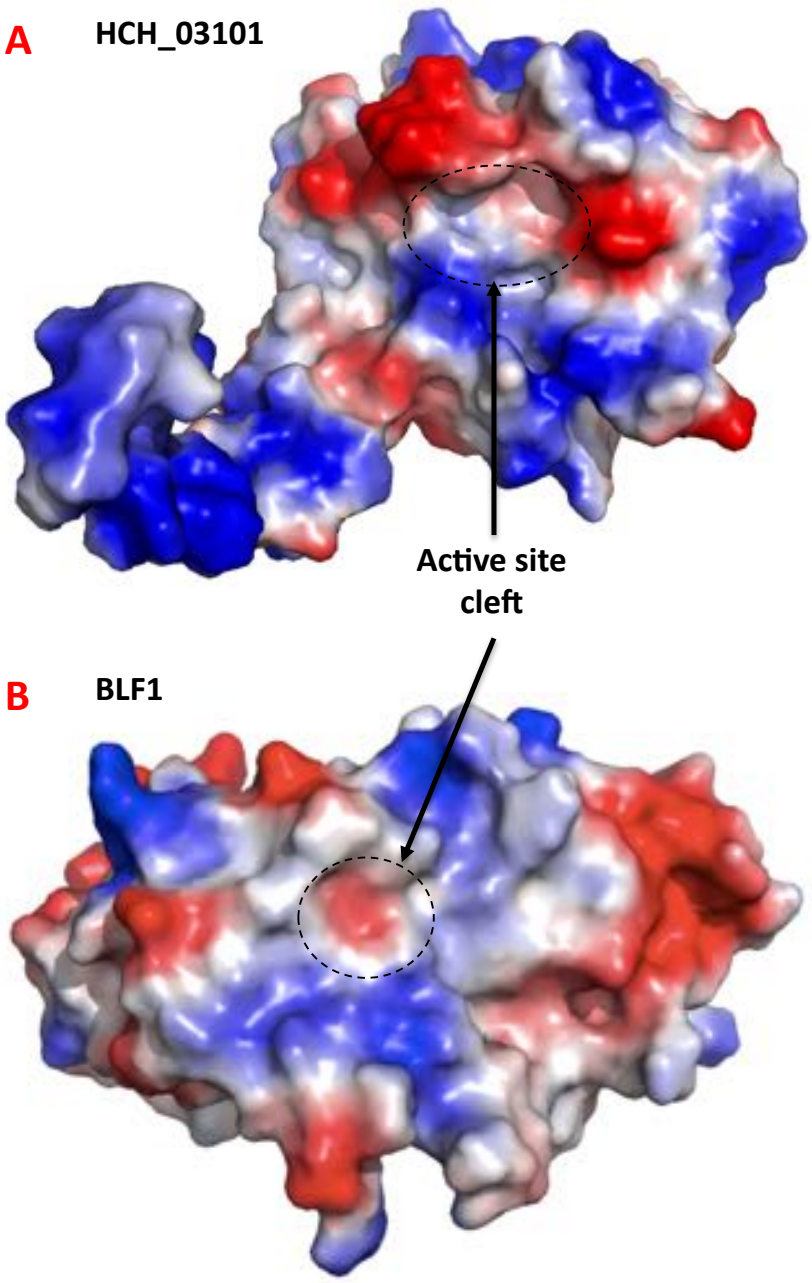


Figure 6.2.4 – Comparison between the active site clefts of BLF1 and HCH_03101. Panels **A** and **B** are electrostatic surface charge diagrams of HCH_03101 and BLF1 respectively, with positive charges shown in blue and negative charges shown in red. On both diagrams a black dashed circle identifies the active site cleft. In HCH-03101 the active site is located in a deep cavity surrounded by a strongly charged ridge. Whereas, the active site in BLF1 is located in a shallow crater surrounded by a neutral ridge. The two proteins also display different charge distributions surrounding the cleft. For example, HCH_03101 is flanked by strong negative charges, whilst BLF1 is encircled by weaker positive charges. However, one aspect that both active sites share in common is a mild negative charge at the centre of the cleft. These active site environments do not suggest that HCH_03101 will share substrate specificity with BLF1. Diagram produced in Pymol.

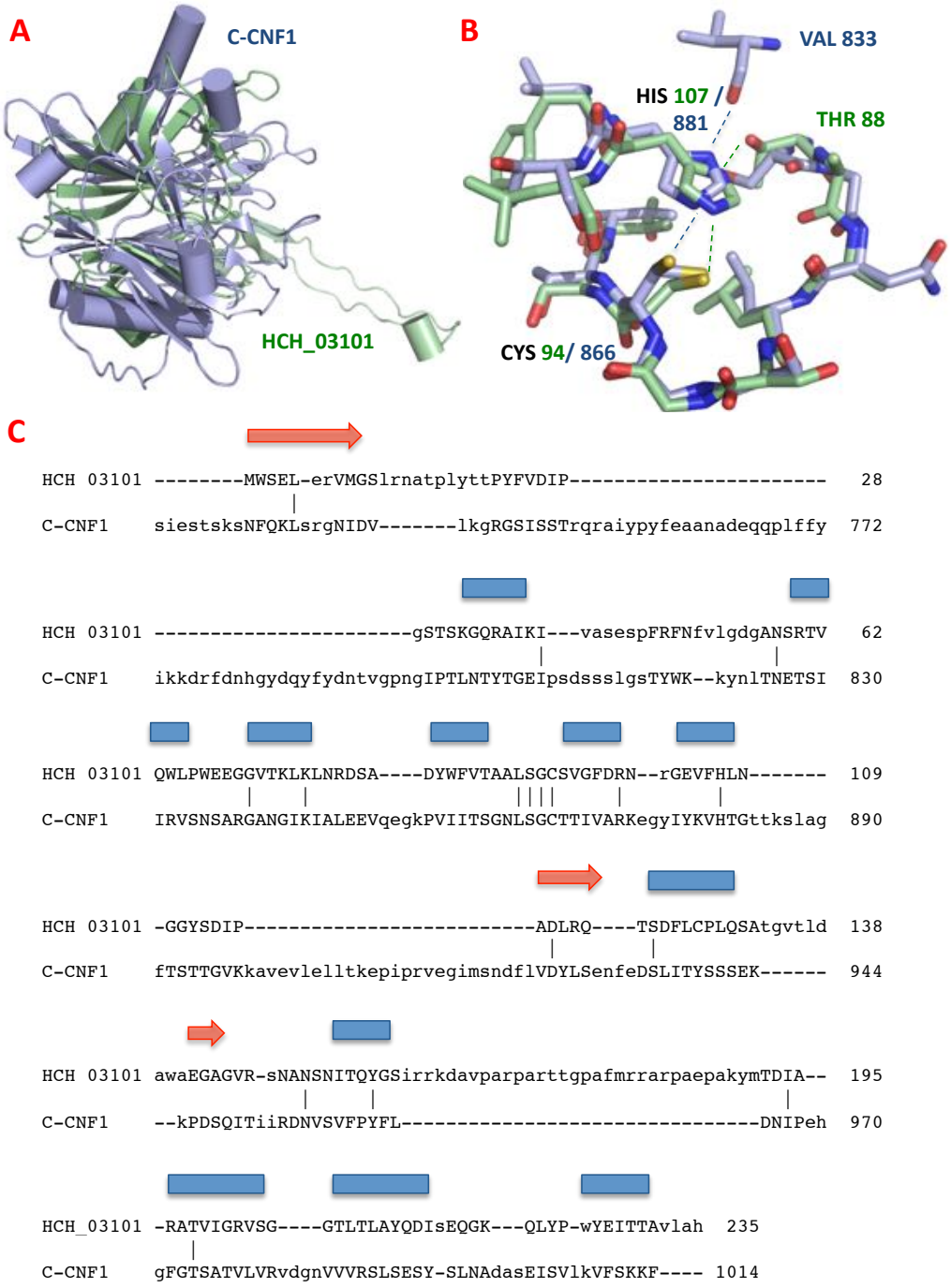


Figure 6.2.5 - Structural alignment of HCH_03101 with C-CNF1. Panel **A** is the structural alignment of HCH_03101 (green) with C-CNF1 (blue). C-CNF1 matches with HCH_03101 across the β-sandwich region, with 161 of its 295 residues aligning to within 3.6 Å RMSD. Panel **B** zooms into the active site region, with both enzymes exhibiting a well conserved CYS-HIS dyad arrangement. The minor differences observed in active site orientation caused by the differing HIS 107 Nε co-ordinating positions, VAL 833 in C-CNF1 as opposed to THR 88 in HCH_03101. Panel **C** is a structure-based sequence alignment. The blue bars represent β-strands and the red arrows α helices shared in common. Outside of the highlighted catalytic residues (panel **B**), there are 11 conserved positions spread evenly across the entire primary sequence. Diagram produced using Lsqkab alignments and Pymol.

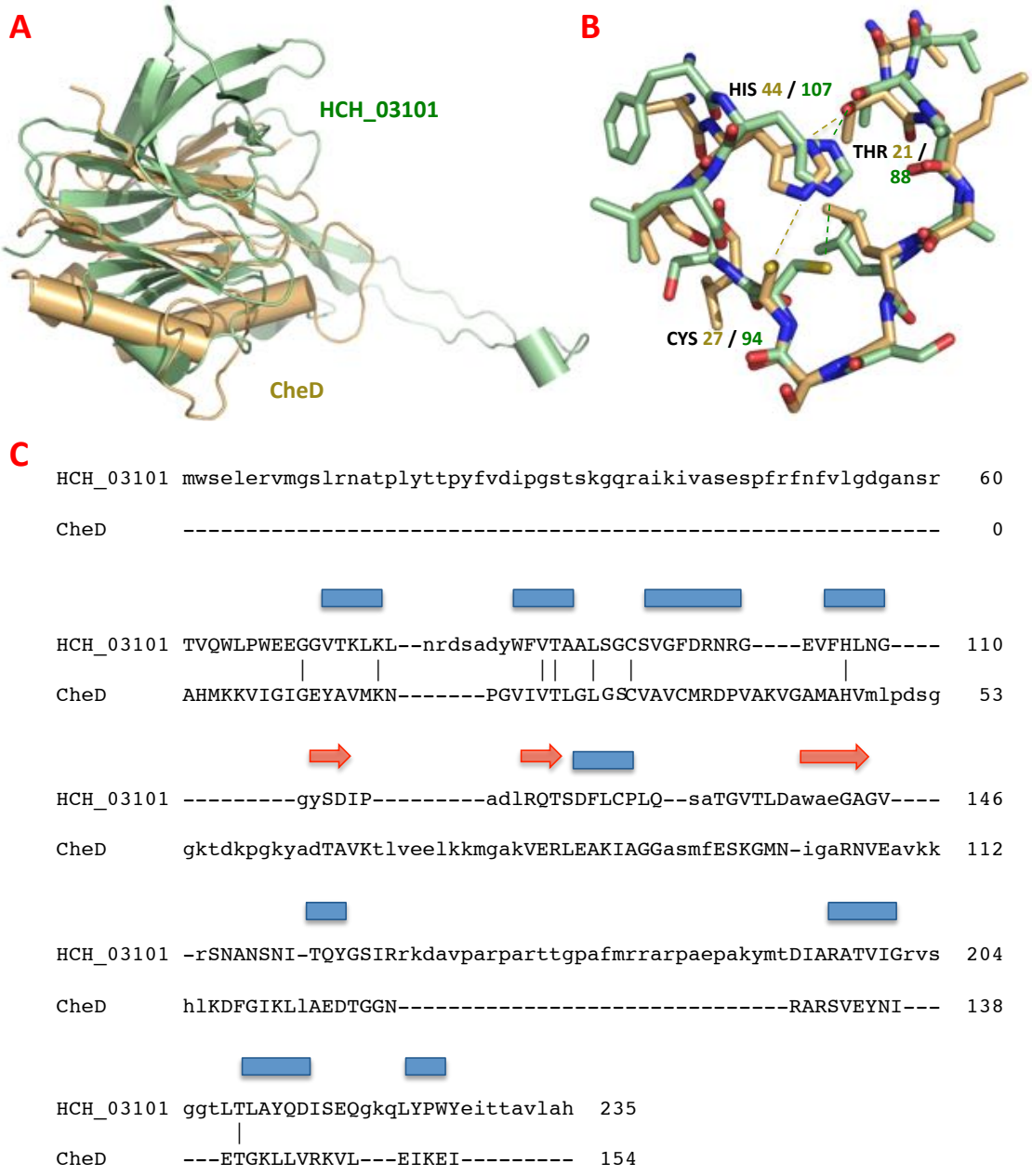


Figure 6.2.6 - Structural alignment of HCH_03101 with CheD. Panel **A** is the structural alignment of HCH_03101 (green) with CheD (orange). CheD aligns with HCH_03101 across 108 of its 154 residues, to within 4.5 Å RMSD. This region accounts for two thirds of the entire primary sequence of CheD and is located predominately within the β-sandwich. Panel **B** zooms into the active site region. CheD does not align with HCH_03101 closely along its peptide backbone, diverging in several places and orienting the essential CYS position in a different direction. Panel **C** is a structure-based sequence alignment. The blue bars represent β-strands and the red arrows α helices shared in common. Outside of the conserved catalytic residues far less sequence similarity is observed than with the toxin type Glutamine de-amidases, with only 4 positions shared between HCH_03101 and CheD. Diagram produced using Lsqkab alignments and Pymol.

Despite two thirds of the polypeptide backbone broadly aligning with HCH_03101, the gross structural alignment is far less strongly conserved, with CheD deviating along the backbone by upwards of 4.6 Å RMSD. This deviation is particularly noticeable at the active site where the backbone shifts both before and after the conserved catalytic loop (figure 6.2.6B).

This reduction in structural similarity from 3-3.7 Å RMSD along the peptide backbone in the toxin type Glutamine de-amidases to 4.6 Å in CheD, is matched by a reduction in primary sequence conservation. With only 6 % sequence similarity observed between HCH_03101 and CheD, as opposed to 10 and 20 % with C-CNF1 and BLF1 respectively. This is further supported by active site comparisons, where there is a clear resemblance. However, unlike the toxin types that match HCH_3101 closely, CheD exhibits small deviations in both the peptide backbone and CYS-HIS dyad (figure 6.2.6B). Comparison of structural alignments made between CheD and C-CNF1 (figure 2.4.3) show that CheD shares a similar level of conservation with HCH_03101 as it does with the toxin type Glutamine de-amidases. HCH_03101 is evidently far less closely related to CheD than it is with BLF1 and C-CNF1. Therefore, increasing the likelihood that it is a member of the toxin sub-family.

Examination of the structure-based sequence alignments across all 3 glutamine de-amidases and HCH_03101, exposes two conserved positions close to the active site that could be utilised in future search motifs (figure 6.2.7). These locations are a GLY residue conserved approximately 19-27 residues upstream from the essential CYS across all four examples and a LYS residue 13-22 residues upstream from CYS 94, which is absent in BLF1. Whilst the location of these conserved positions at first appears too broad for inclusion in a motif, when you discount the distantly related CheD these distances are sharpened to between 23-27 and 18-22 for the conserved GLY and LYS respectively. With more specific primary sequence locations in place, both conserved positions make a compelling case for inclusion in future search motifs. The above comparison also discounts the aforementioned WLPW region, which shares no conservation with either CheD or C-CNF1.

Having previously determined that HCH_03101 does not share similar chemical characteristics in the cleft with BLF1 comparison with the remaining Glutamine de-amidase enzymes was carried out (figure 6.2.8). Comparing HCH_03101 with both C-CNF1 and CheD in this region emphasises that there is very little similarity between any of the glutamine de-amidases. With all 4 examples displaying different cleft shapes and surrounding charge distributions. This raises the possibility that the small patch of mildly negative charge, shared between HCH_03101 and BLF1 at the heart of their active site clefts (figure 6.2.4) may be more significant than it first appears. Particularly as the surface accessible WLPW motif is specifically conserved with BLF1.

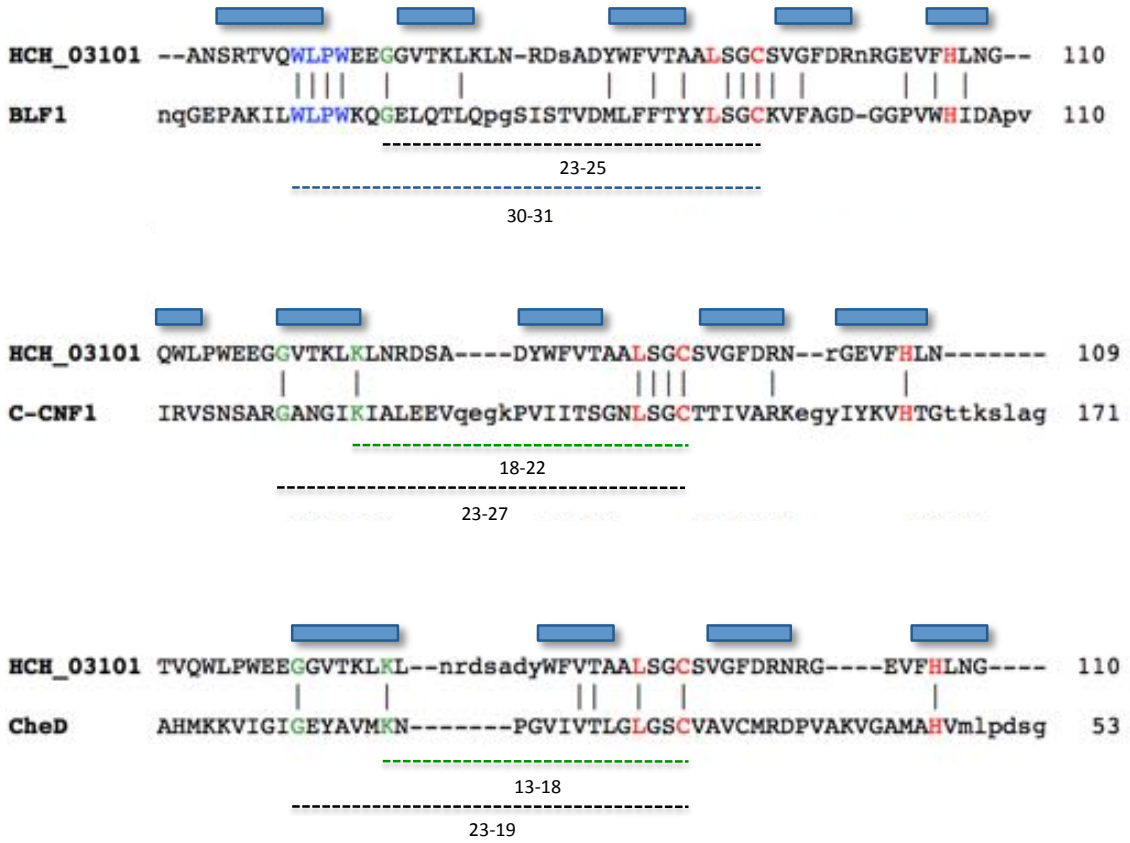


Figure 6.2.7 – Constructing a stronger search motif by incorporating novel sequence conservation close to the active site. This diagram shows structure-based sequence alignments of the active site regions of HCH_03101 with BLF1, C-CNF1 and CheD respectively. The active site dyad and conserved LEU, incorporated in the current search motif, are highlighted red. The previously highlighted WLPW motif (section 6.2.1) conserved between BLF1 and HCH_03101 is coloured blue, whilst the single conserved residues identified through comparison with C-CNF1 and CheD are green. Underneath each alignment is a dotted line denoting the length between the conserved position and the catalytic CYS residue. The WLPW motif is only conserved with BLF1, with no alignment identified in the other examples. The isolated GLY and LYS residues however are far more prevalent with the conserved GLY present in four examples and the LYS shared amongst three. When the distantly related CheD is discounted, the distance between the conserved positions and the catalytic CYS across all three examples is reduced to a fine range suitable for inclusion in a search motif. Diagram produced using Lsqkab alignments.

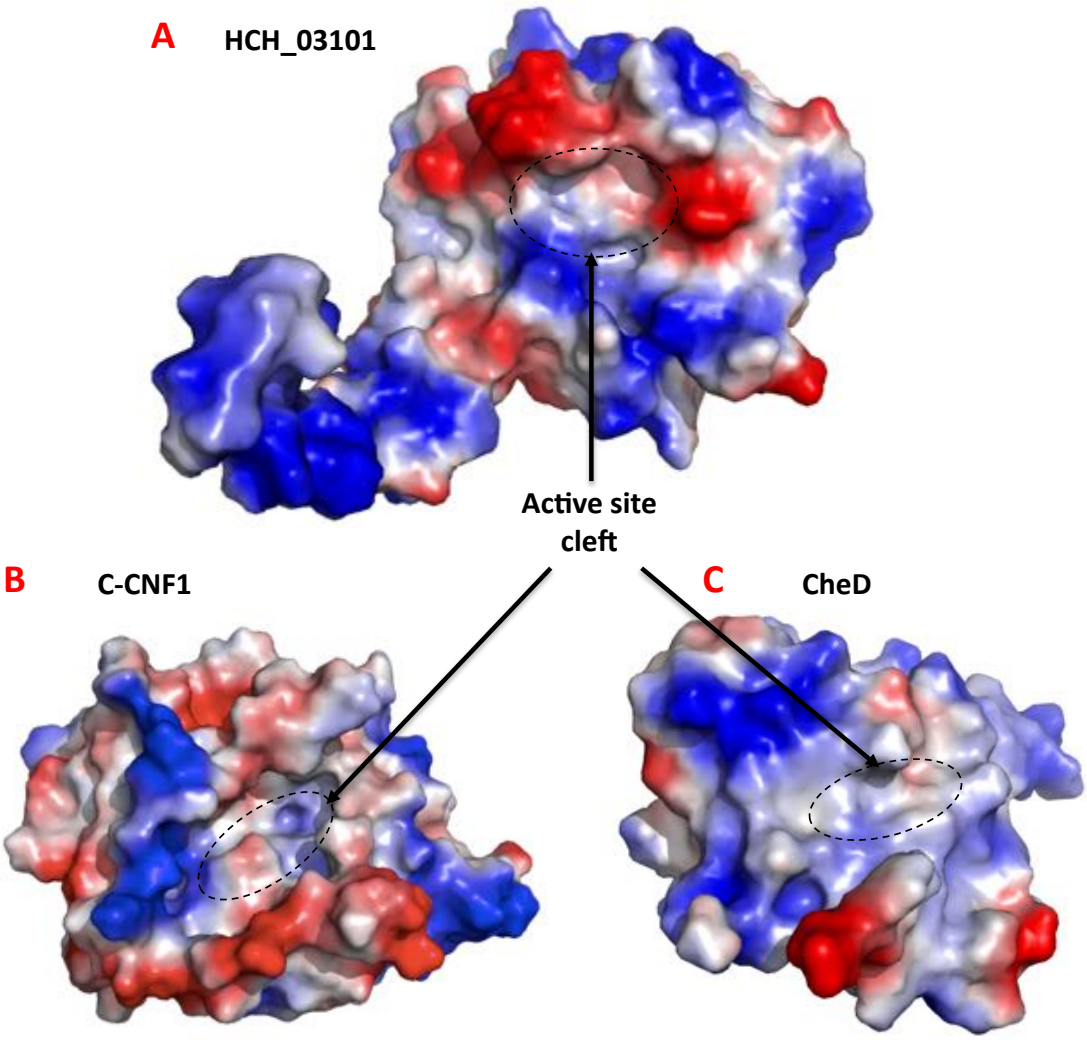


Figure 6.2.8 – Comparison of the active site clefts of HCH_03101 with C-CNF1 and CheD. Panels **A**, **B** and **C** are electrostatic charge diagrams showing the front face of HCH_03101, C-CNF1 and CheD respectively. On each diagram blue and red regions correspond to positive and negative charges respectively, with the active site cleft highlighted by a black dashed circle. The surface charge characteristics of HCH_03101 are not shared with either C-CNF1 or CheD. The active site cleft of HCH_03101 is characterised by a deep negatively charged cleft surrounded by a highly charged ridge, flanked by both positive and negatively charged patches on either side. In contrast C-CNF1 shares the deep cleft, but is positively charged and flanked by either weakly charged or neutral patches whilst. CheD is unlike both HCH_03101 and C-CNF1, in that the active site cleft is located within a neutral valley in the middle of a weakly charged region. Nothing in the above comparison suggests that HCH_03101 will share binding partners with either enzyme. Diagram produced in Pymol.

6.3 – Structural analysis of the C94S active site mutant of HCH_03101

The C94S mutant of HCH_03101 was initially produced for the purpose of carrying out pull-down assays. In order to validate these experiments it was first desirable to determine whether the mutant shares the WT protein's tertiary folding characteristics. The structure of the hypothetical inactive C94S mutant was solved from a differing crystal form, exhibiting unit cell dimensions extended 12 Å along the c axis (section 5.7).

Comparison of the gross structural alignments of WT HCH_03101 with the C94S mutant indicate that the novel crystal form does induce subtle conformational changes in the structure (figure 6.3.1). The β-protrusion in particular is arranged in a different orientation, appearing to fold about a hinge point beyond the initial β-stranded base. There are also minor alterations observed in the connecting loops, particularly the longer examples flanking the active site. Despite these deviations the core secondary structure elements in both structures align extremely closely. Validating the mutants use in the pull-down experiments detailed in chapter 7.

The active site region of the C94S mutant adopts a slightly different conformation compared to the WT (figure 6.3.2). The majority of the peptide backbone in the mutant aligns perfectly. However, close inspection of the mutated position 94 shows that the side chain is oriented facing away from any possible interaction with HIS 107. This is unexpected, as the terminal carbonyl of the mutant SER 94 should still be capable of forming a hydrogen bond with the imidazole ring. This small shift at the active site appears to have a far larger knock-on effect on a surface accessible loop above, which significantly alters the crystal contacts formed on the top face of the enzyme.

Examination of the crystal packing shows that there is no alteration in either the orientation or distances between the crystal contacts surrounding the β-protrusion (figure 6.3.3). In both cases the tip of the protrusion is held in a cleft, formed from residues 155-160 located at the base of an alternate symmetry related β-protrusion (figure 6.3.3B). With the side of the protrusion held via a hydrogen bond between GLU 186 and a symmetry related ARG 6 residue (figure 6.3.3C). However, closer inspection of the active site region reveals that the C94S mutation has a direct effect on the orientation of the loop containing TRP 67. Part of the WLPW motif conserved with BLF1, which is located at the surface of the protein (figure 6.3.4). This modification alters a crystal contact formed between the backbone carbonyl of GLY 70 and the backbone amide of a symmetry related ALA 150, with the C94S mutant displaying an 18 Å gap between equivalent positions. Therefore, the 12 Å difference observed between the two crystal forms is because of alterations in the crystal packing.

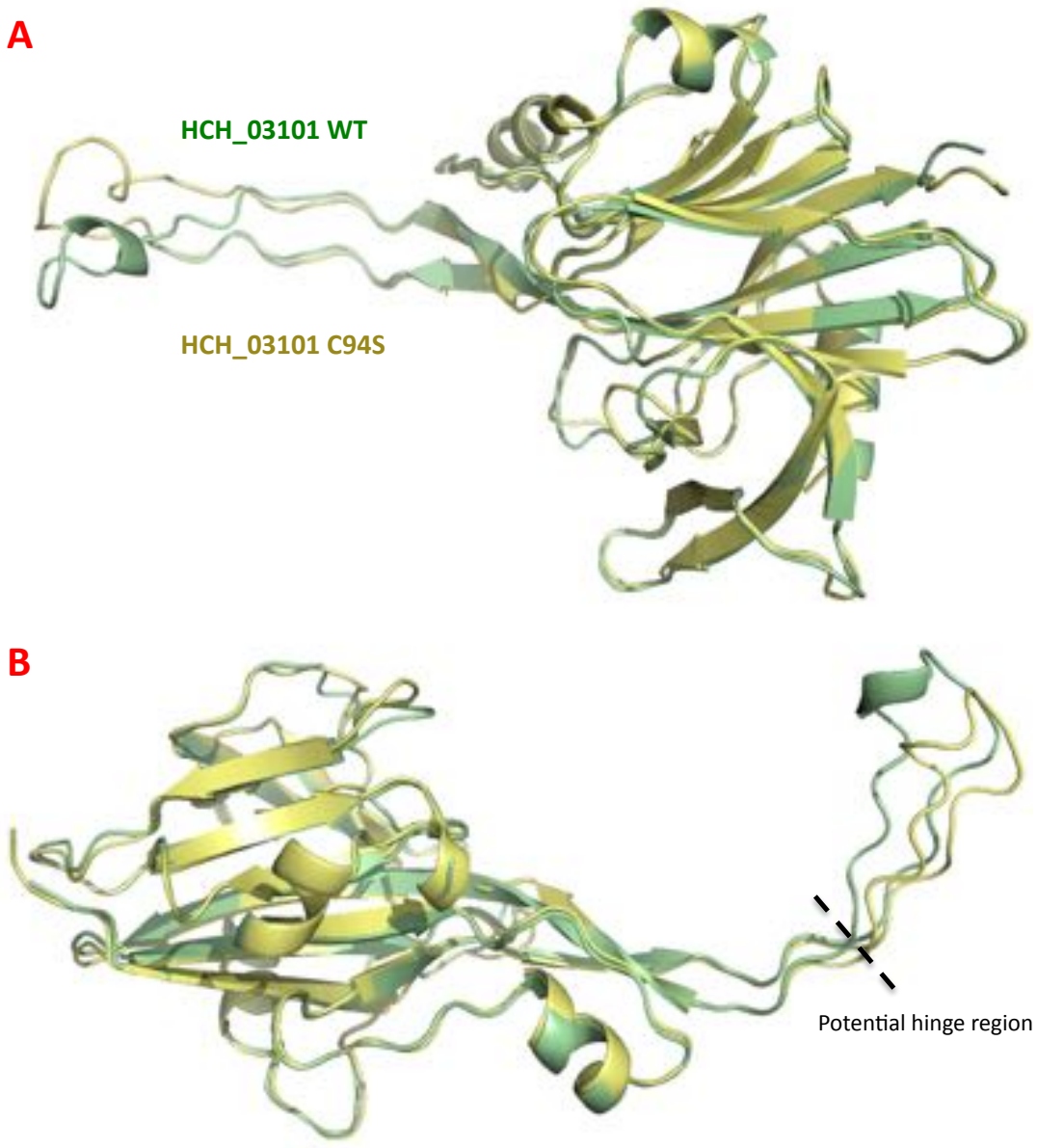


Figure 6.3.1 – Comparing WT HCH_03101 with an active site mutant C94S. Panel **A** shows an alignment of WT HCH_03101 with the C94S mutant looking down on the top face, with panel **B** showing the same alignment but from the side face. The gross secondary structure of both crystal forms remains largely the same. However, there are significant differences observed in both the longer connecting loops, particularly above the active site and past the β-stranded base of the protrusion. The latter is particularly interesting, as the β-protrusion appears to be flexible about a hinge region highlighted in panel **B** with a black dashed line. Diagram produced in Pymol.

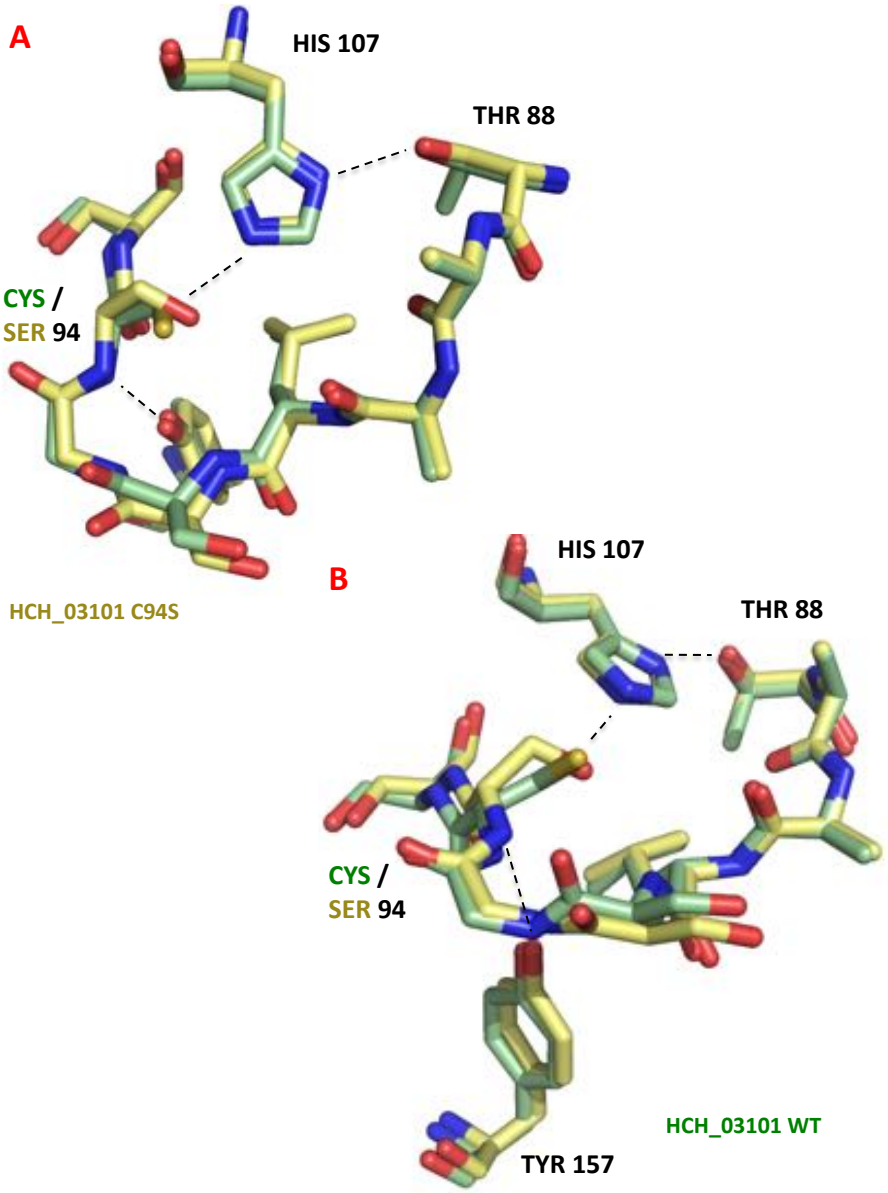


Figure 6.3.2 – Active site comparison between the C94S and WT structures of HCH_03101. Panel **A** shows an alignment of the active site regions of WT (green) and C94S (yellow) HCH_03101 shown from the top face, whereas panel **B** displays the same region but from the solvent accessible face. This alignment shows that the peptide backbone remains unchanged across the two crystal forms. However, the mutation from CYS 94 to SER 94 does appear to have a subtle effect on the active site, with SER 94 oriented facing away from an interaction with HIS 107. Diagram produced in Pymol.

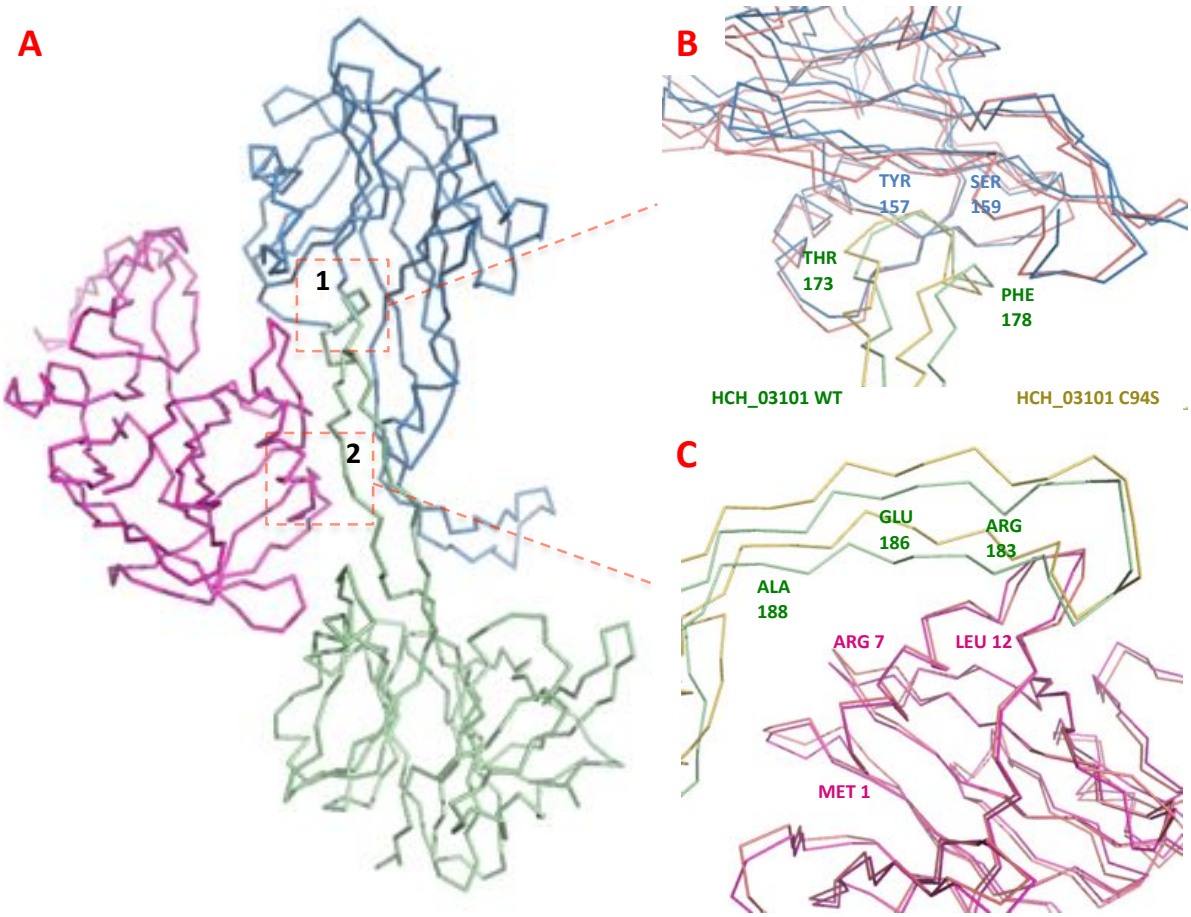


Figure 6.3.3 – The β-protrusion is held by the same crystal contacts between both crystal forms. Panel **A** shows a ribbon representation of WT HCH_03101 with 2 symmetry related subunits interacting with the β-protrusion, the crystal contacts formed are highlighted in red boxes. The following diagrams show both the WT (green) and C94S mutant (yellow) aligned, showing that the contacts made in both crystal forms are the same. Panel **B** zooms into the first crystal contact. Where the tip of the β-protrusion is held in place by residues 156-159, located on β-14 at the base of a symmetry related protrusion, coloured blue and red for the WT and mutant respectively. Panel **C** shifts to the second crystal contact where residues 183-188 in the centre of the protrusion, interact with residues 4-11 located on α1 coloured magenta and red for the WT and mutant respectively. These diagrams show that there is little alteration to the crystal contacts formed at the β-protrusion, certainly none that could account for the 12 Å difference observed between the two crystal forms. Diagram produced in Pymol.

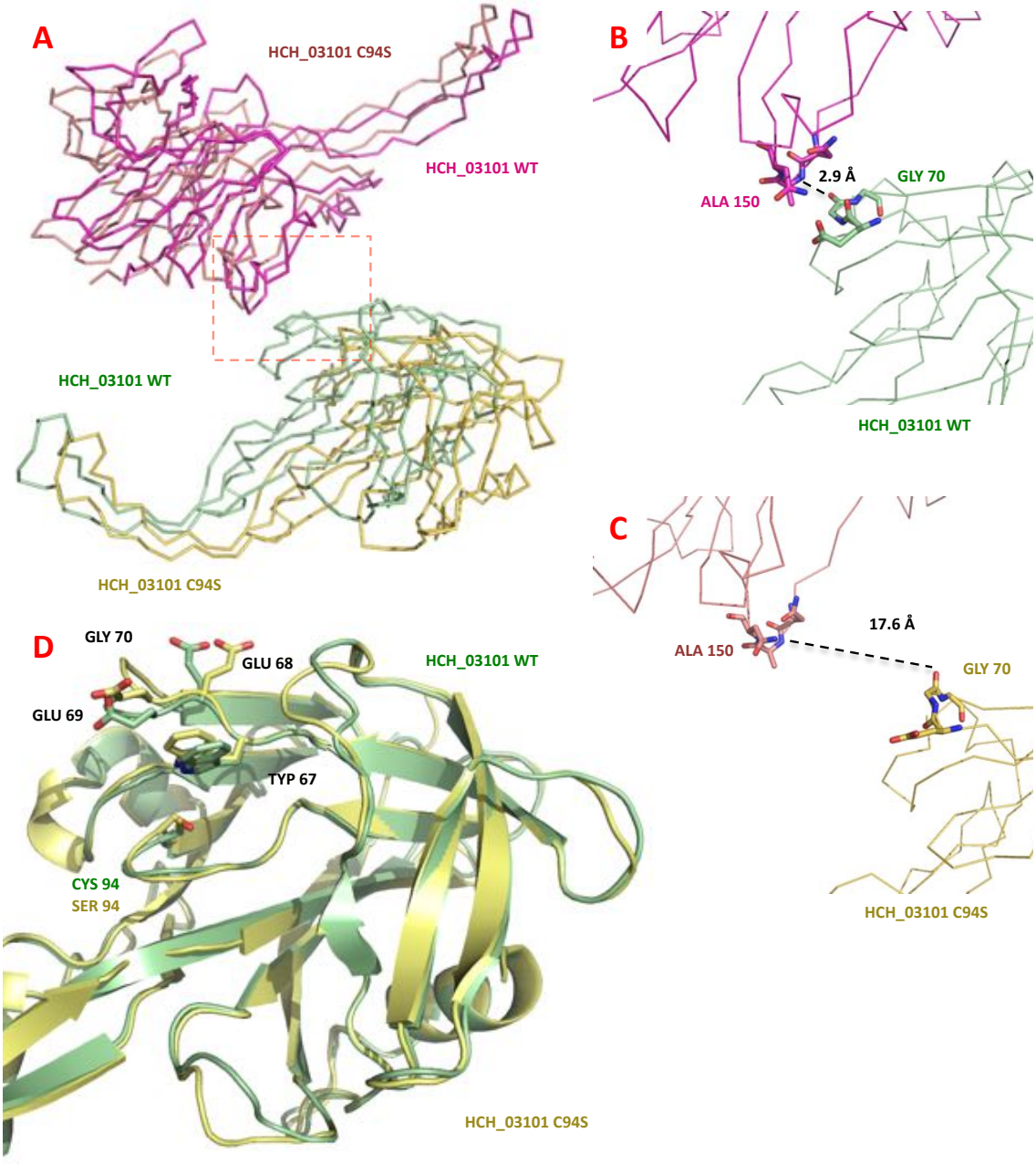


Figure 6.3.4 – The C94S mutation induces alternative crystal packing along the globular body. Panel **A** shows a pair of symmetry related molecules from both the WT (green & magenta) and C94S mutant (yellow & red) of HCH_03101, with the altered crystal contact highlighted in a red box. Panel **B** zooms into this contact in the WT crystal form, showing a clear hydrogen bond formed between the backbone carbonyl of GLY 70 with the backbone amide of the symmetry related ALA 150. Panel **C** displays the same region within the C94S mutant, where a significant 17 Å distance between the two previously hydrogen bonded residues has developed. Panel **D** shows that the altered crystal packing is the result of an alternative conformation exhibited by the loop containing GLY 70. This modification is presently believed to be the result of the C94S mutation, which pushes the aromatic TYR 67 side chain upwards along with the rest of the loop. Diagram produced in Pymol.

6.4 – Examination of the β-protrusion

Having examined HCH_03101 alongside the other Glutamine de-amidase enzymes, it is clear that the β-protrusion is a novel feature. It is unlikely that a loop this long would be conserved without reason. However, what is not clear is if the β-protrusion has a functional role in substrate binding. Sequence and structure comparison of the protrusion using BLAST and Dali-lite respectively, identifies no similar features amongst characterised proteins, leaving few avenues to assign a functional role. However, it is located directly in front of the active site and has been shown to be flexible about a hinge region some 30 Å removed from the cleft, which may suggest a role in substrate binding.

The β-protrusion extends out 44 Å projecting directly out from the β-sandwich from β14 and β16. Examination of the peptide backbone (figure 6.1.5C) reveals evidence to suggest that the protrusion is held in a two stranded β-sheet up till a hinge point (figure 6.3.1). This hinge region then marks the point where the β-protrusion becomes less ordered, with fewer backbone interactions formed. This flexibility is neatly highlighted by the absence of α 15 in the C94S model, with the turning point instead interpreted as a flexible 3-10 loop.

The charge distribution of the β-protrusion is also a significant feature, as the vast majority of the residues present in this region are positively charged. Inspection of the protrusions residue distribution reveals a repeating ARG heavy region followed closely by an AXPAXPA motif (figure 6.4.1). The APAXPA motif is repeated on the 2nd strand of the protrusion, as part of what appears to be a duplication between positions 165 – 171 and 182 - 188. This repeating motif is noteworthy because not only is the sequence motif mirrored across both strands of the β-protrusion, but so is the structure. The two sequence-repeats adopt identical peptide backbone conformations giving rise to a pseudo-2-fold symmetry axis, whilst interacting with one another as part of the two anti-parallel strands of the β-protrusion (figure 6.4.1C). This raises the possibility that the repeat has some functional significance.

6.5 HCH_03101 structural conclusions

HCH_03101 closely resembles all the current members of the Glutamine de-amidase family, with the β-sandwich broadly conserved between all three and the exact route of the β-strands shared with both BLF1 and C-CNF1 the toxin examples.

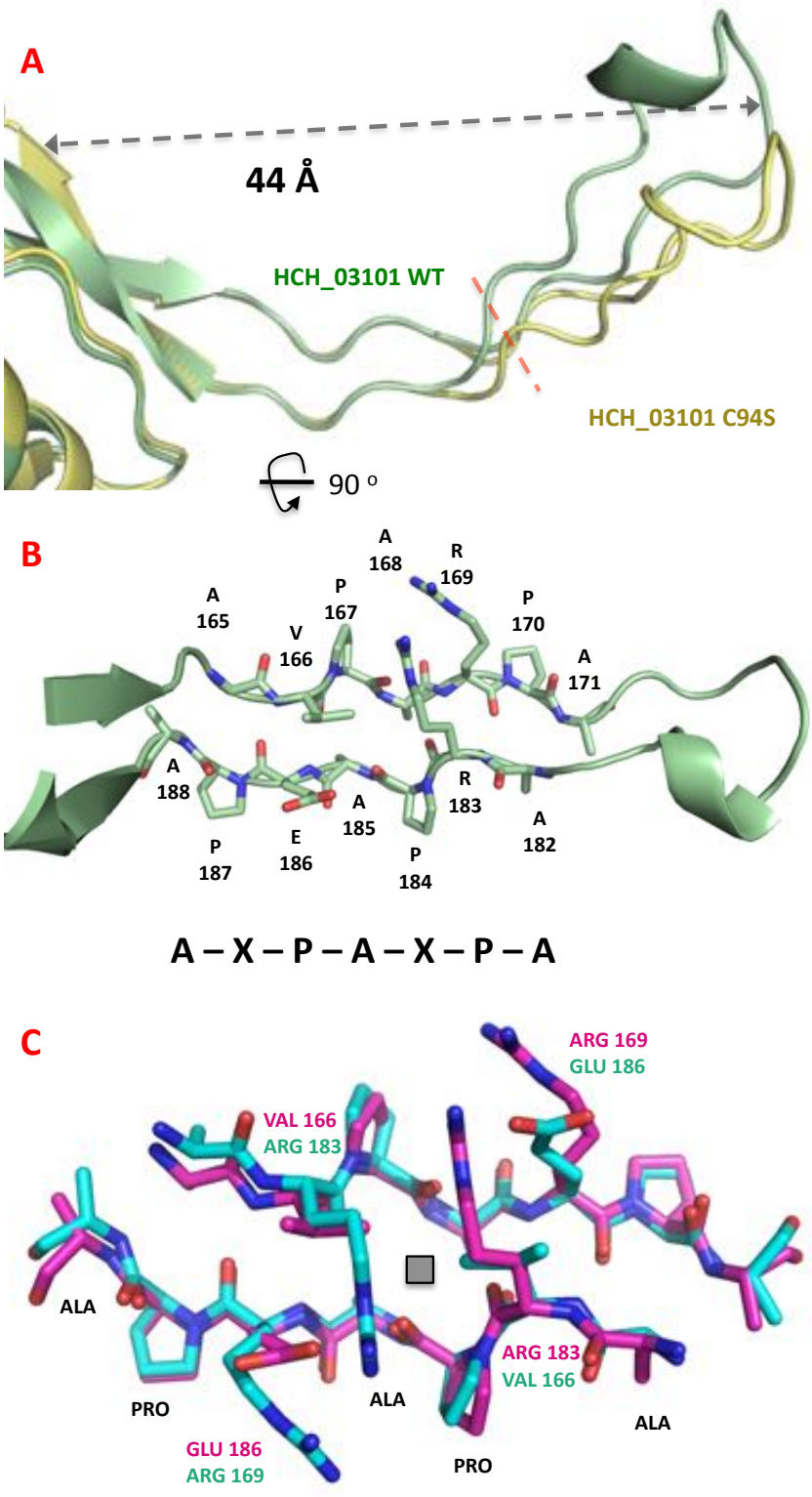


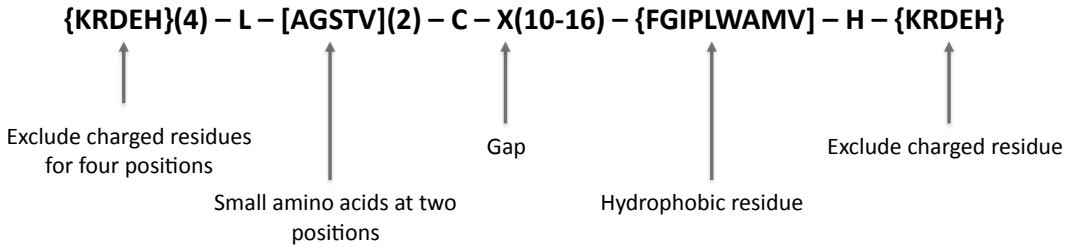
Figure 6.4.1 – The long β protrusion contains a Proline rich repeating motif. Panel **A** shows the β-protrusion region of the WT (green) and C94S (yellow) structures of HCH_03101, highlighting the putative hinge region in red. Panel **B** shows the protrusion from the top face, with the repeating AXPAXPA motif highlighted. Panel **C** is a super-position of the repeating AXPAXPA motif with residues 182-188 aligned on top of 165-171, showing that the motif shared on both strands of the β-protrusion, is not only structurally conserved but also displays a pseudo-d-fold symmetry axis (grey square). Diagram constructed in Pymol.

The active site of HCH_03101 exhibits the characteristic CYS-HIS dyad, with the peptide backbone super-imposing closely with BLF1 and both dyad residues supported through conserved interactions from the side and below. This remarkable level of active site conservation combined with a moderate, 20 % level of sequence conservation combine to suggest that HCH_03101 likely exhibits Glutamine de-amidase activity.

Despite these similarities HCH_03101 displays several novel features not observed in previous Glutamine de-amidase examples. The active site cleft in particular lies in a very different electrostatic environment to both CheD and C-CNF1 with only BLF1 sharing any common characteristics. The final note worthy feature is the large β-protrusion that lies adjacent to the active site and displays a degree of flexibility about a hinge region, which may play a role in substrate binding. Even more curiously, this feature appears to exhibit a mirrored AXPAXPA motif on both strands that shares a pseudo-2-fold symmetry axis. This is not a common structural feature and likely involved in the function of the protein, or is as a result of a recent gene insertion event.

Structural alignments also highlight several conserved primary sequence positions that can be used to strengthen the search motif used to identify future Glutamine de-amidase targets. The original and strengthened search motifs are shown below:

Original:



Strengthened:

G - X(4-5) – K – X(8-15) – {KRDEH}(4) – L – [AGSTV](2) – C – X(10-16) – [FGIPLWAMV] – H – {KRDEH}

Chapter 7: Characterising HCH_03101

Having solved the structure of HCH_03101 it was clear that this protein shares significant sequence and structure conservation with other members of the Glutamine de-amidase super-family. The next aim of this study was to identify the binding partners and functional properties of HCH_03101.

7.1.1 - HCH_03101 does not display toxic activity in J774 macrophage cells

Macrophage killing assays, performed by Dr Lynda Partridge (University of Sheffield), highlight the challenges involved in characterising HCH_03101. In this case the toxicity of HCH_03101 was probed by incubation alongside a BALB/c J774.2 macrophage cell line, held in Dulbecco's modified Eagle medium (DMEM) (4.5g/L Glucose, 2 mM L-Glutamine and 1 mM Sodium pyruvate, supplemented with 10 % (v/v) fetal bovine serum). No J774 cell death was observed, but the assays probed with high concentration HCH_03101 (> 1 mg ml^{-1}) displayed crystal formation showing that the concentrations present may have been significantly reduced.

7.1.2 – HCH_03101 Pull-down assays

HCH_03101 has been shown to purify quickly and cleanly using a Ni-NTA affinity column (chapter 5.3.1). This raises the possibility that HCH_03101 could be probed for its binding partners whilst safely immobilised when bound to a Ni-NTA resin.

7.1.3 - General pull-down methodology

A simple method to probe an immobilised protein's binding partners is the pull-down assay (Chapter 4.4). Pull-down assays start by immobilising a bait protein onto a Ni-NTA aminodiacetic acid resin (Sigma Aldrich). This resin permits manipulation of the bound bait proteins and any complexes formed, through rapid re-suspension in a variety of buffering conditions, intended to first wash away contaminants and then elute any bound proteins. The charged resin loaded with bait protein, is probed via incubation with cell free extract containing many potential binding partners. After this incubation the cell free extract is removed and the non-specific contaminants washed away, through repeated rinsing and resuspension of the resin in low salt buffers. The washing buffers used have their salt concentration gradually increased, which encourages the dissociation of weakly binding proteins and non-specific contaminants. The putative binding partners that form strong

complexes with the bait protein are then eluted through two washing steps. The first is a high salt buffer, which dissociates proteins in complex with the bait protein. The second is an Imidazole wash, that strips away all the proteins and contaminants remaining in complex with either the bait protein or resin.

The eluates from the high salt and Imidazole washes are then run on a SDS-PAGE gel to assess the quantity and size of the potential binding partners. However, identification of significant hits requires suitable controls to differentiate the bands that equate to specific interactions from those that are caused by non-specific contamination. The target bands are then excised from this SDS-PAGE gel and digested through incubation with 200 ng of Trypsin overnight at 37 ^{o}C, producing small peptides. These peptides were then analysed using ESI TOF MS/MS Mass-spectrometry with the peptides identified and matched to putative proteins using the MASCOT server (Matrix Science^{TM}). The Mass-spectrometry data presented in this study was kindly provided by Dr Mark Dickman (University of Sheffield).

7.2 - Pull-down proof of concept with C-CNF1

Pull-down assays have previously been used to determine the substrate of BLF1 (Cruz-Migoni *et al.,* 2011). However, prior to using the technique to analyse novel targets, it was decided that a proof of concept would be desirable. For this purpose an in-active mutant of C-CNF1 was produced, with the catalytic CYS 866 position substituted with a SER residue and a 6xHIS extension inserted at the N-terminus. The rationale behind this experiment is if both BLF1 and C-CNF1 can be characterised using this method, then it is likely that any putative binding partners of HCH_03101 can also be identified this way.

7.2.1 - Production of recombinant C-CNF1 C866S NTD 6xHIS protein

The C-CNF1 gene fragment was amplified from a colony of UT-189 *E. coli* cells, kindly supplied by Professor Ian Roberts (University of Manchester). Both the gene fragment and a sample of pET21a+ expression vector were digested with 5' NcoI and 3' Bam HI restriction sites (chapter 4.1.7), prior to being ligated together with T4 DNA ligase (chapter 4.1.8). Once this construct was confirmed through gene sequencing, conducted by the core genomics group (University of Sheffield), the catalytic CYS 866 position was substituted for a SER residue using the Quick-change II method (chapter 4.1.12). Over-expression trials of this C-CNF1 C866S NTD 6xHIS construct were conducted across 15, 20, 25, 30 and 37 ^{o}C temperature brackets for 4 hours, with C-CNF1 expressing exclusively in the insoluble fraction (figure 7.1.1A). This high level of

insoluble expression suggested that the recombinant protein was either being incorrectly folded or immediately re-directed into inclusion bodies. Attempts to resolve the latter possibility involved a second set of trials, which exchanged the nutrient rich LB medium with M9 minimal medium (chapter 4.2.1). The introduction of M9 medium slows the rate of expression, with sufficient soluble protein obtained after 4 hours at 37 ^{o}C (figure 7.1.1B).

7.2.2 - Purifying C-CNF1 C866S NTD 6xHIS

The tagged C-CNF1 C866S protein was then purified using a 5 ml Ni-HP column (figure 7.1.2). 4 g of cell paste was resuspended in 35 ml of buffer A (50 mM Tris pH 8, 0.5 M NaCl) before the cell walls were disrupted with 3, 20 second rounds of sonication. The soluble fraction was then separated by centrifugation at 60, 000 *g* for 20 minutes in a JLA-25.50 rotor, before being loaded directly onto the Ni-HP 5ml column at 5 ml min^{-1}. The protein was then eluted from the column with a 0-100 % linear gradient of buffer B (50 mM Tris pH 8, 0.5 M NaCl, 0.5 M Imidazole) across 10 column volumes, C-CNF1 eluting at 0.2 M Imidazole. SDS-PAGE analysis of the purification (figure 7.1.2B) shows the protein obtained is > 90 % pure and suitable for pull-down assays without any further chromatography stages.

7.2.3 - C-CNF1 C866S pull-down assays against an *E.coli* expression strain for RhoA-GST

Purified C-CNF1 C866S bait protein was bound on to a Ni-NTA resin as previously described (chapter 4.4.1). The charged resin was then probed through incubation with cell free extract from a BL21 (DE3) strain of *E. coli* expressing a recombinant fusion protein of RhoA with a GST tag, for 2 hours at 4 ^{o}C. Post incubation the resin was washed 4 times in low salt buffer (25 mM HEPES pH 7.5, 0.1 M NaCl) to remove any unbound cell free extract and non-specific contaminants. The tightly bound complexes were then eluted, using first a high salt buffer (25 mM HEPES pH 7.5, 1.2 M NaCl) and then an EDTA buffer (25 mM HEPES pH 7.5, 1 M EDTA) and run out on a SDS-PAGE gel (figure 7.2.3). This pull-down clearly shows that the RhoA-GST fusion protein forms a strong interaction with C-CNF1, with a significant portion of the RhoA administered still associated with the bait protein post salt wash, to be found later in the EDTA wash (red box).

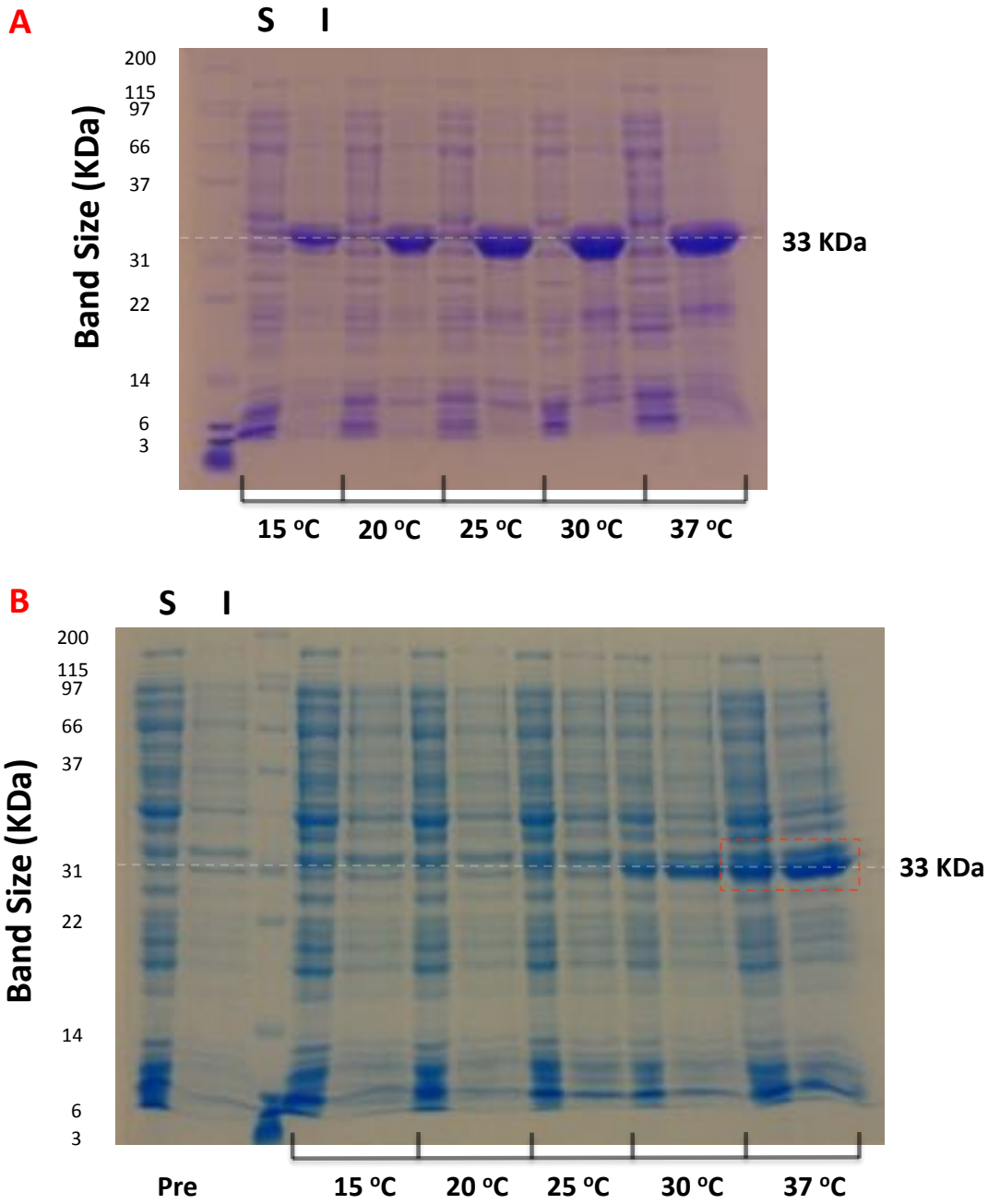


Figure 7.2.1 – Over-expression trials for C-CNF1 C866S 6xHIS. Panel **A** shows an over-expression trial for C-CNF1 C886S NTD 6xHIS, undertaken in LB medium at 15, 20, 25, 30 and 37 ^{o}C temperature brackets, at the 4 hour time point. It shows that C-CNF1 expresses exclusively in the insoluble fraction regardless of the incubation temperature and that the level of expression is high. Panel **B** shows another over-expression trial, where the nutrient rich LB growth medium has been exchanged for M9 minimal media. This alteration to the culturing conditions slows down the rate of expression, with a sizable portion of soluble protein produced at 37 ^{o}C after 4 hours (red box).

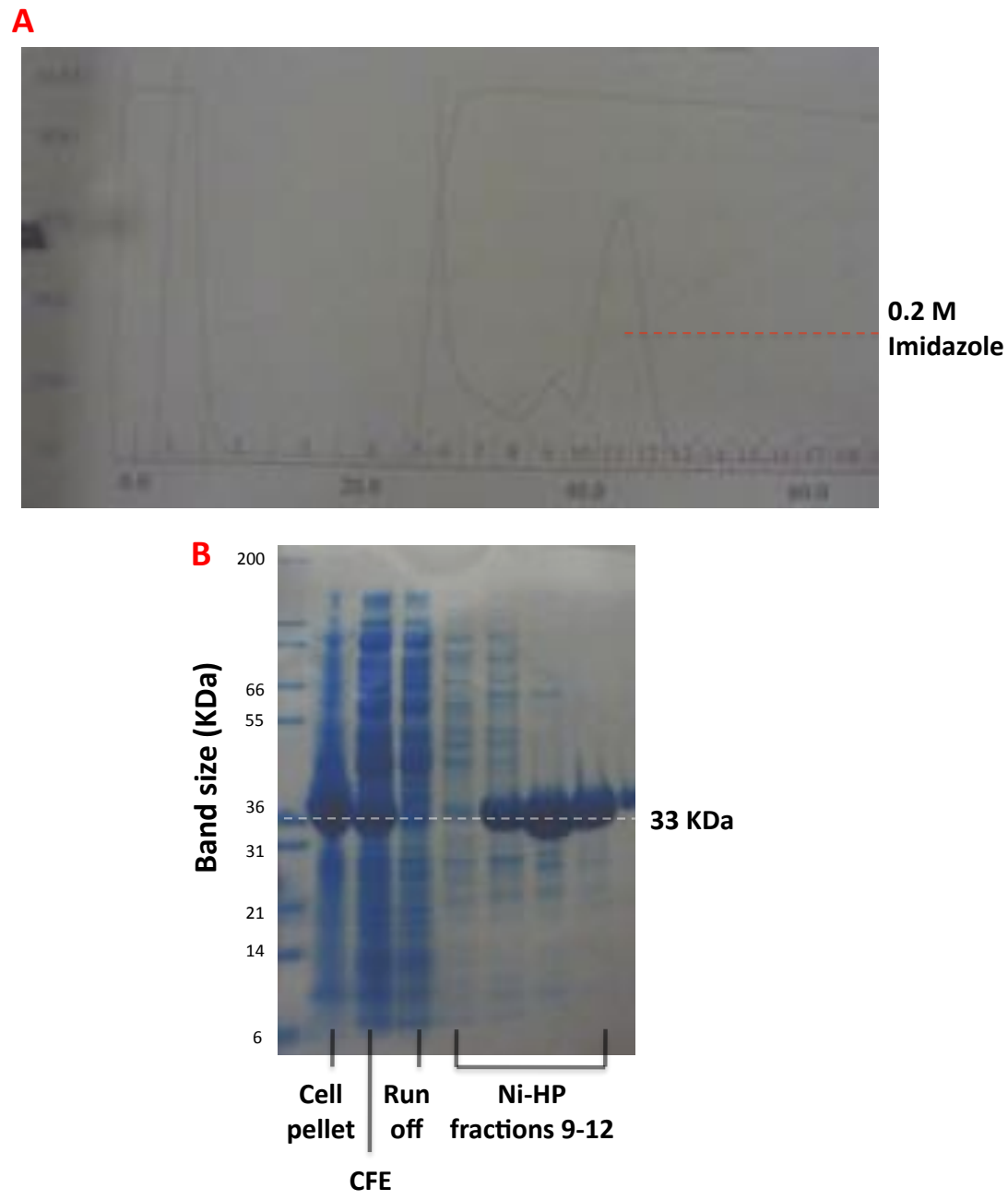
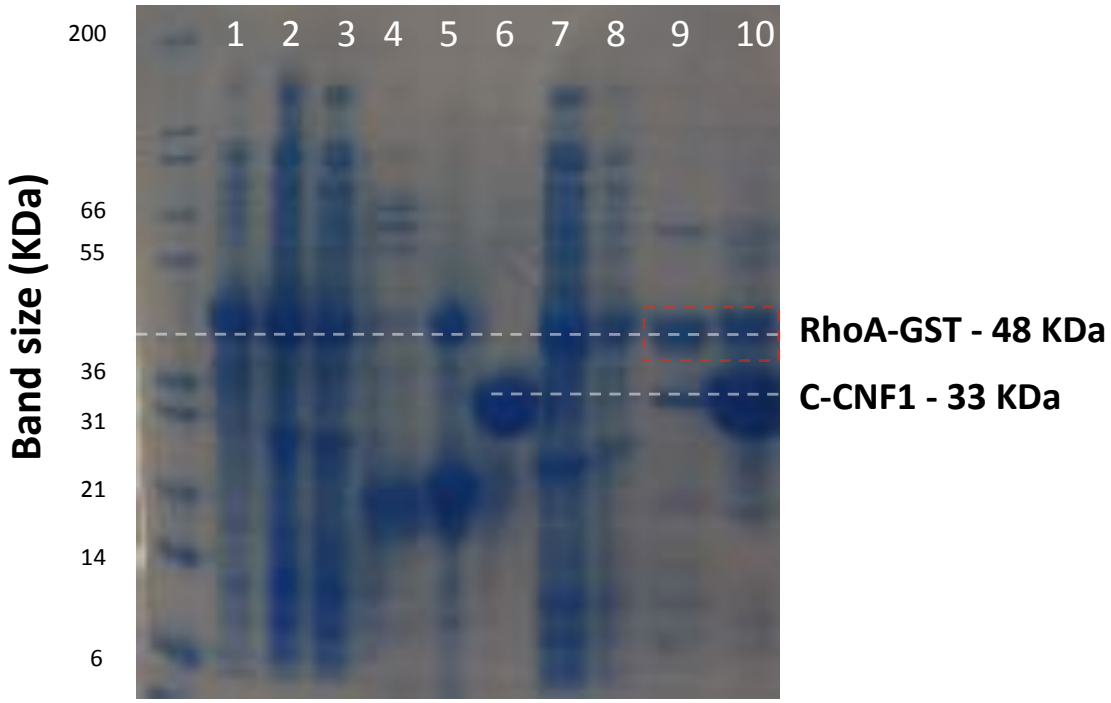


Figure 7.2.2 – C-CNF1 C866S 6xHIS purification. Panel **A** shows the purification trace of a Ni-HP 5 ml column, the y axis shows UV absorbance and the x axis eluted volume. There are two peaks at 37 and 42 ml, corresponding to contaminants and C-CNF1 C886S respectively. The column was run on a linear gradient from buffer A (50 mM Tris pH 8, 0.5 M NaCl) to 100 % buffer B (50 mM Tris pH 8, 0.5 M NaCl, 0.5 M Imidazole), with C-CNF1 eluted at 0.2 M Imidazole. Panel **B** is an SDS-PAGE analysis of the above purification. Lane 1, Mark 12 MW marker; lane 2, 20 µg C-CNF1 C886S cell pellet; lane 3, 20 µg C-CNF1 C886S cell free extract; lane 4, 20 µg Ni-HP run off; lanes 5-8, 13 µl Ni-HP fractions 9-12. The above gel shows that the C-CNF1 C886S protein is > 90 % pure from a single step.



1. **RhoA-GST pellet**
2. **RhoA-GST CFE**
3. **RhoA-GST CFE 24 hrs**
4. **RhoA cleaved**
5. **RhoA combined**
6. **Pure C-CNF1**
7. **Pull-down unbound**
8. **0.1 M NaCl wash**
9. **1.2 M NaCl wash**
10. **EDTA wash**

Figure 7.2.3 – RhoA pull down with C-CNF1 C886S 6xHIS. The C-CNF1 C866S NTD 6xHIS bait was probed using cell free extract from a BL21 strain of *E. coli* expressing a GST-tagged fusion protein of RhoA. Lane 1, 20 µg RhoA-GST cell pellet; lane 2, 20 µg RhoA-GST cell free extract; lane 3, 20 µg RhoA-GST cell free extract after 24 hours; lane 4, Thrombin treated RhoA-GST; lane 5, 20 µg purified RhoA and RhoA-GST; lane 6, 20 µg pure C-CNF1 C886S; lane 7, 20 µg unbound material; lane 8, 15 µl 0.1 M NaCl wash; lane 9, 15 µl 1.2 M NaCl wash; lane 10, 15 µl 0.5M EDTA wash. The RhoA-GST band at 48 KDa, when cleaved with thrombin (lane 4) yields a band the correct size for RhoA at 22 KDa, validating the construct. The Ni-NTA resin was charged with purified C-CNF1 (lane 6) then incubated with the RhoA-GST cell free extract (lane 2) for 2 hours at 4 ^{o}C. Post incubation the resin was washed with low salt buffer (25 mM HEPES pH 7.5, 0.1 M NaCl), removing the non-specific contaminants. The putative binding partners were eluted, first with a high salt buffer (25 mM HEPES pH 7.5, 1.2 M NaCl) (lane 9) and then with an EDTA wash (25 mM HEPES pH 7.5, 0.5 M EDTA) (lane 10). The presence of RhoA-GST in both the high salt and Imidazole washes shows that one of the components of the fusion protein forms a strong complex with C-CNF1.

Importantly this experiment shows that the pull-down methodology is successful for both characterised Glutamine de-amidase toxins. Therefore, any binding partners identified in this fashion for HCH_03101 may well be significant.

7.3 – HCH_03101 pull-down assay with BALB/c J774.2 macrophages

HCH_03101 has been shown to share substantial structural similarity to BLF1 as well as sharing limited sequence similarity. Thus despite differences in the regions flanking their respective active sites (figure 6.2.4), it remains a possibility that HCH_03101 is a functional homologue of BLF1. To test this hypothesis the same macrophage cell line used to characterise BLF1 and its interaction with eIF4a, BALB/c J774.2 (J774), was also used to probe HCH_03101.

7.3.1 – Control selection for the HCH_03101 – J774 macrophage pull-down

The following pull-downs were conducted as previously described (chapter 4.4), with four different bait proteins. These four proteins were YloQ (negative control), BLF1 C94S (positive control), HCH_03101 WT (active bait) and HCH_03101 C94S (inactive bait). YloQ was chosen because it does not exhibit any specific interactions with the J774 proteome. Therefore, any bands exhibited in the YloQ assay would be the result of non-specific interactions and could be safely discounted if encountered in the HCH_03101 pull-down. BLF1 C94S was chosen as a positive control, with an established banding pattern from previous pull-down assays. HCH_03101 may also share a substrate with BLF1, adding comparative value to this control. Both the WT and C94S mutant of HCH_03101 were tested to maximise the chances of isolating a binding partner. The putative inactive C94S mutant is also a negative control, to aid in determining the binding site of HCH_03101.

7.3.2 – J774 macrophage pull-down methodology

A quantity of 1.3 x 10^8 J774 macrophage cells, kindly supplied by Dr Lynda Partridge (University of Sheffield), were cultured in 50 ml flasks of DMEM medium. The J774 cells were then centrifuged at 5000 *g* for 5 minutes and resuspended in 2 ml of breakage buffer (25 mM HEPES pH 7.5, 0.1 M NaCl, 1 % TRITON X-100, 10 % Glycerol, 1 mM PMSF and Protease inhibitor). These re-suspended cells were then broken by passage through a fine syringe needle 5 times, prior to centrifugation at 20, 000 *g* for 10 minutes to produce the clarified cell free extract. This cell free extract had a measured protein concentration of 12 mg ml^{-1}, determined by Bradford assay (chapter 4.2.4). Dr Svetlana Sedelnikova (University of Sheffield) kindly supplied

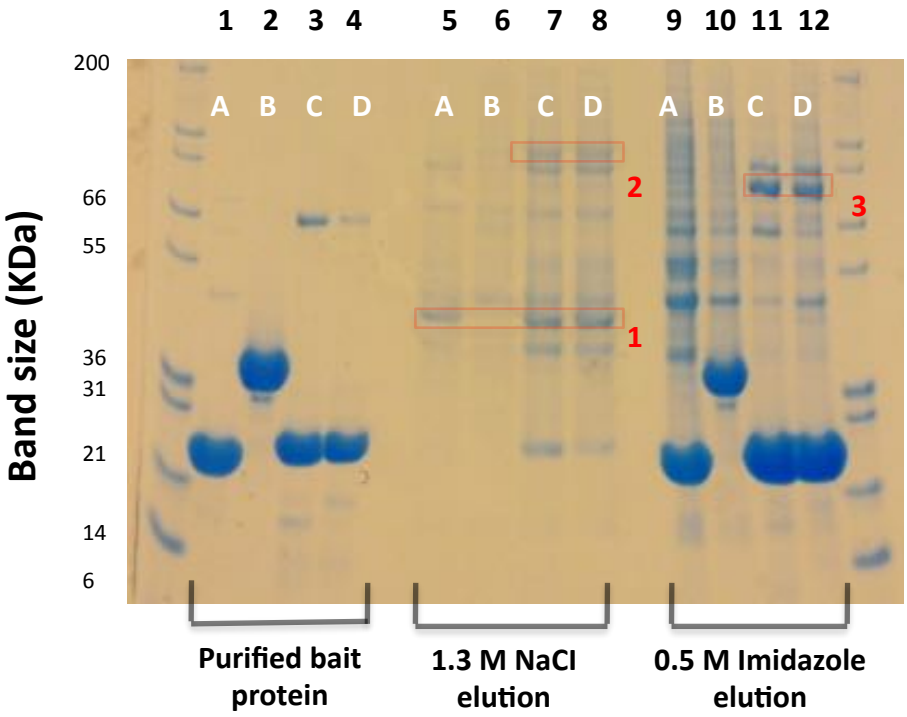
the YloQ and BLF1 C94S samples, whilst the WT and C94S HCH_03101 samples were purified freshly (chapter 5.3.1). 20 µl of Ni-NTA resin was charged with 2 mg of each of the bait proteins and equilibrated in breakage buffer. 500 µl (6 mg) of the J774 cell free extract was then added to this charged resin and incubated for 2 hours at 4 ^{o}C. This cell free extract was then removed and the resin washed with 3 rounds of buffer A (25 mM HEPES pH 7.5, 0.3 M NaCl, 1 % TRITON X-100), followed by a further 3 rounds of buffer A minus the detergent to remove low affinity contaminants. The binding partners are then eluted with 50 µl of high salt elution buffer (25 mM HEPES pH 7.5, 1.3 M NaCl), incubated with the resin under agitation for ten minutes at 4 ^{o}C. A second wash step finally strips the resin by introducing 100 µl of Imidazole buffer (50 mM Tris pH 8, 0.5 M Imidazole). To identify the binding partners pulled down 15 µl of the high salt elution and 20 µg of the Imidazole elution where then run on a SDS-PAGE gel alongside the purified bait protein.

7.3.3 – HCH_03101 shares a binding partner the same size as eIF4a with BLF1

The SDS-PAGE analysis of the proteins pulled down (figure 7.3.1) shows several potential binding partners for HCH_03101 (red boxes). Band 1 is shared with BLF1 and is approximately 40 KDa large, roughly the same size as eIF4a, strongly suggesting that HCH_03101 also binds the initiation factor. However HCH-03101 also exhibits two unique bands labelled 2 and 3, which are not pulled down by BLF1, at 100 and 80 KDa respectively.

7.3.4 – MS/MS Mass-spectroscopy shows that HCH_03101 binds to eIF4a

The most intriguing band from the J774 macrophage pull-down is band 1, which at 46 KDa is the same molecular weight as eIF4a and is shared between HCH_03101 and BLF1. Peptides analysed from this band, using MASCOT (Figure 7.3.2) searching the SWISSPROT database, show that in both BLF1 and HCH_03101 band 1 corresponds to eIF4a. However unlike BLF1, HCH_03101 also pulls down several other proteins involved in translation, at a lower level as judged by the number of unique peptides matched. Such as the Eukaryotic peptide chain release factor, Eukaryotic elongation factor 1α and Eukaryotic translation initiation factor 3. However, in comparison with the HCH_03101 pull-down of eIF4a, which displays high prot-scores, the additional translation machinery hits have been pulled down with far fewer unique peptides and correspondingly lower prot-scores. This is relevant because the prot-score provides a qualitative measure, based on the number of unique peptides matched, of how likely it is that the identified protein is present in the sample.



A = BLF1 C94S, B= YloQ, C = HCH_03101, D = HCH_03101 C94S

Figure 7.3.1 – J774 Macrophage pull-down assay. The SDS-PAGE gel above is for a pull-down assay run with 4 different bait proteins against J774 macrophage cell free extract. The four bait proteins were: A, BLF1 C94S; B, YloQ; C, HCH_03101 WT and D, HCH_03101 C94S. The pull-down assay is divided into 3 sections the first is the purified bait protein prior to charging the Ni-NTA resin. The second is a 1.3 M NaCl salt wash intended to elute proteins in complex with the bait, with the final stage a 0.5 M Imidazole wash that strips the Ni-NTA beads. Lane1, Mark 12 MW marker; lanes 1-4, 15 µg bait protein; lanes 5-8, 15 µl high salt elution; lanes 9-12, 20 µg Imidazole elution. There are 3 bands of interest highlighted in red boxes. Band 1 is present in both BLF1 and HCH_03101 and is 46 KDa, the same MW as eIF4a. However, bands 2 and 3 at 100 and 80 KDa respectively, appear unique to HCH_03101 and could represent this enzymes specific binding partner.

A **Top BLF1 C94S hits (gel band 1)**

gi|1903220	type II intermediate filament of hair keratin [Homo sapiens]
gi|4504935	keratin, type II cuticular Hb5 [Homo sapiens]
gi|530400232	PREDICTED: keratin, type II cuticular Hb3 isoform X1 [Homo sapiens]
gi|530412192	PREDICTED: keratin, type I cuticular Ha1 isoform X1 [Homo sapiens]
gi|14917115	keratin, type I cuticular Ha1 [Homo sapiens]
gi|10337581	keratin, type I cuticular Ha3-II [Homo sapiens]
gi|45752211	keratin hHa3-I [Homo sapiens]
gi|224036248	Chain A, Crystal Structure Of The Eif4a-Pdcd4 Complex
gi|119581133	keratin, hair, acidic, 4 [Homo sapiens]
gi|485388	eukaryotic initiation factor 4AII [Homo sapiens]
gi|22798968	type II hair keratin 2 [Homo sapiens]
gi|119581140	keratin, hair, acidic, 5, isoform CRA_a [Homo sapiens]
gi|119581139	keratin, hair, acidic, 2 [Homo sapiens]
gi|1181994	hair type I acidic keratin [Homo sapiens]
gi|7661920	eukaryotic initiation factor 4A-III [Homo sapiens]
gi|4504927	keratin, type I cuticular Ha6 [Homo sapiens]
gi|12311759	type I hair keratin 8 [Homo sapiens]
gi|762885	Plakoglobin [Homo sapiens]
gi|7161763	type II hair keratin 1 [Homo sapiens]
gi|61966777	V-set and immunoglobulin domain-containing protein 8 precursor [Homo sapiens]
gi|93279940	Chain A, Structure Of The Dead Domain Of Human Eukaryotic Initiation Factor 4a, Eif4a
gi|609412460	Chain A, Crystal Structure Of Human Lnkh2b-h2a.z-anp32e
gi|1147813	desmoplakin I [Homo sapiens]
gi|223582	histone H4
gi|12655474	keratin associated protein 9.8 [Homo sapiens]
gi|31791022	keratin-associated protein 19-5 [Homo sapiens]
gi|51859376	H3 histone, family 3A [Homo sapiens]
gi|187761884	RecName: Full=Tubulin beta-8 chain-like protein LOC260334 [Homo sapiens]
gi|32111	unnamed protein product [Homo sapiens]
gi|193784993	unnamed protein product [Homo sapiens]
gi|189054178	unnamed protein product [Homo sapiens]
gi|578830527	PREDICTED: keratin, type I cytoskeletal 24 isoform X1 [Homo sapiens]
gi|12329925	keratin associated protein [Homo sapiens]
gi|530410154	PREDICTED: myosin phosphatase Rho-interacting protein isoform X1 [Homo sapiens]

B **Top HCH_03101 WT hits (gel band 1)**

IF4A1_HUMAN	Eukaryotic initiation factor 4A-I OS=Homo sapiens GN=EIF4A1 PE=1 SV=1
IF4A2_HUMAN	Eukaryotic initiation factor 4A-II OS=Homo sapiens GN=EIF4A2 PE=1 SV=2
IF4A3_HUMAN	Eukaryotic initiation factor 4A-III OS=Homo sapiens GN=EIF4A3 PE=1 SV=4
K2C1_HUMAN	Keratin, type II cytoskeletal 1 OS=Homo sapiens GN=KRT1 PE=1 SV=6
K1C10_HUMAN	Keratin, type I cytoskeletal 10 OS=Homo sapiens GN=KRT10 PE=1 SV=6
ACTB_HUMAN	Actin, cytoplasmic 1 OS=Homo sapiens GN=ACTB PE=1 SV=1
K1C9_HUMAN	Keratin, type I cytoskeletal 9 OS=Homo sapiens GN=KRT9 PE=1 SV=3
TBA1B_HUMAN	Tubulin alpha-1B chain OS=Homo sapiens GN=TUBA1B PE=1 SV=1
HAT1_HUMAN	Histone acetyltransferase type B catalytic subunit OS=Homo sapiens GN=HAT1 PE=1 SV=1
COR1A_HUMAN	Coronin-1A OS=Homo sapiens GN=CORO1A PE=1 SV=4
K22E_HUMAN	Keratin, type II cytoskeletal 2 epidermal OS=Homo sapiens GN=KRT2 PE=1 SV=2
RBBP7_HUMAN	Histone-binding protein RBBP7 OS=Homo sapiens GN=RBBP7 PE=1 SV=1
ERF1_HUMAN	Eukaryotic peptide chain release factor subunit 1 OS=Homo sapiens GN=ETF1 PE=1 SV=3
UBA3_HUMAN	NEDD8-activating enzyme E1 catalytic subunit OS=Homo sapiens GN=UBA3 PE=1 SV=2
ACTBL_HUMAN	Beta-actin-like protein 2 OS=Homo sapiens GN=ACTBL2 PE=1 SV=2
K1H1_HUMAN	Keratin, type I cuticular Ha1 OS=Homo sapiens GN=KRT31 PE=2 SV=3
NP1L1_HUMAN	Nucleosome assembly protein 1-like 1 OS=Homo sapiens GN=NAP1L1 PE=1 SV=1
TBB4B_HUMAN	Tubulin beta-4B chain OS=Homo sapiens GN=TUBB4B PE=1 SV=1
PDIA6_HUMAN	Protein disulfide-isomerase A6 OS=Homo sapiens GN=PDIA6 PE=1 SV=1
ALBU_HUMAN	Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2
SH3G1_HUMAN	Endophilin-A2 OS=Homo sapiens GN=SH3GL1 PE=1 SV=1
HNRPF_HUMAN	Heterogeneous nuclear ribonucleoprotein F OS=Homo sapiens GN=HNRNPF PE=1 SV=3
DX39A_HUMAN	ATP-dependent RNA helicase DDX39A OS=Homo sapiens GN=DDX39A PE=1 SV=2
ARP3_HUMAN	Actin-related protein 3 OS=Homo sapiens GN=ACTR3 PE=1 SV=3
PLEC_HUMAN	Plectin OS=Homo sapiens GN=PLEC PE=1 SV=3
EIF3F_HUMAN	Eukaryotic translation initiation factor 3 subunit F OS=Homo sapiens GN=EIF3F PE=1 SV=1
KIF15_HUMAN	Kinesin-like protein KIF15 OS=Homo sapiens GN=KIF15 PE=1 SV=1
EIF3E_HUMAN	Eukaryotic translation initiation factor 3 subunit E OS=Homo sapiens GN=EIF3E PE=1 SV=1
CCD18_HUMAN	Coiled-coil domain-containing protein 18 OS=Homo sapiens GN=CCDC18 PE=2 SV=1
EF1A1_HUMAN	Elongation factor 1-alpha 1 OS=Homo sapiens GN=EEF1A1 PE=1 SV=1
NIBL1_HUMAN	Niban-like protein 1 OS=Homo sapiens GN=FAM129B PE=1 SV=3

Figure 7.3.2 – MS-MS Mass-spec analysis of band 1 from the J774 pull-down assay. Panel **A** shows the MASCOT output from MS-MS Mass-spectroscopy on band 1 from the BLF1 C94S pull-down, ordered by prot-score. Outside of the heavy Keratin contamination observed, the top hits are eIF4a-I, alongside its two isomers eIF4a II and III (red). Panel **B** shows the MASCOT output for band 1 from the HCH_03101 WT pull-down, ordered by prot-score. The top three hits are isoforms of eIF4a (red) shared with BLF1, but there are also several further components of the translation machinery pulled back (blue) not present in the BLF1 pull-down. Such as Eukaryotic peptide chain release factor, eIF3 and elongation factor 1-α.

Therefore, the elongation factor (figure 7.3.4A) with a prot-score of 47 based on a single unique peptide may not be a reliable identification. However, the peptide release factor has been assigned from 4 unique peptides and gives a prot-score of 214, which strongly suggest it is present within the band. Therefore, HCH_03101 has been shown to interact with more than one protein family, a feature not observed in BLF1, which merits further investigation.

The remaining potential binding partners located in bands 2 and 3 are specific to HCH_03101 and correspond to α-actinin and β-tubulin respectively. Both of these proteins are common contaminants, that have previously been identified in BLF1 pull-downs. Therefore, neither band represents a significant binding partner for HCH_03101.

7.3.5 – HCH_03101 displays Glutamine de-amidase activity

The MASCOT server identifies peptides based on their measured mass and can be configured to take into account de-amidation and oxidisation events. Closer inspection of the peptides uncovered for eIF4a-I show that BLF1 C94S and both HCH_03101 proteins both pull-down a peptide containing GLN 339, the substrate-binding site of BLF1 (figures 7.3.3B and 7.3.4B). The BLF1 C94S bait used in this pull-down is inactive and as a result the peptides identified at this position are not de-amidated (figure 7.3.3B). Whereas the WT HCH_03101 pull-down (figure 7.3.3D) shows two variants of the same peptide, the first is a high scoring single de-amidation and the second a low scoring double de-amidation event. Given that de-amidation has been detected in a peptide containing the GLN residue targeted by BLF1, it is likely that HCH_03101 not only displays Glutamine de-amidase activity but also targets the same substrate as BLF1. However, this result is tempered by the negative control, HCH_03101 C94S, which also de-amidates this particular peptide. This is possibly due to low-level WT HCH_03101 contamination present in the mutant C94S sample, introduced during purification and further work is underway to determine whether the C94S mutant of HCH_03101 is an active Glutamine de-amidase.

7.4 - HCH_03101 pull-down assay with *Tetraselmis. suecica* Algae

The J774 macrophages are a useful cell line for analysing the potential interactions HCH_03101 may form with Eukaryotic proteins. *H. chejuensis* the organism that produces HCH_03101, is a recently discovered marine bacterium, with no evidence to suggest it is a human pathogen. Therefore it is presently unknown what organism HCH_03101 is targeted against.

A **Top BLF1 C94S hits (gel band 1)**

gi|4503529 **Mass:** 46125 **Score:** 1521 **Matches:** 67(26) **Sequences:** 30(17)
eukaryotic initiation factor 4A-I isoform 1 [Homo sapiens]
gi|485388 **Mass:** 46365 **Score:** 858 **Matches:** 41(16) **Sequences:** 19(9)
eukaryotic initiation factor 4AII [Homo sapiens]
gi|7661920 **Mass:** 46841 **Score:** 434 **Matches:** 20(5) **Sequences:** 12(4)
eukaryotic initiation factor 4A-III [Homo sapiens]

B **Previously identified BLF1 binding site**

Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide
1073.0850	2144.1554	2143.1273	1.0281	0	(53)	0.01	1		R.GIDVQQVSLVINYDLPTNR.E
715.7610	2144.2612	2143.1273	1.1338	0	69	0.00019	1		R.GIDVQQVSLVINYDLPTNR.E

C **Top HCH_03101 WT hits (gel band 1)**

IF4A1_HUMAN **Mass:** 46125 **Score:** 1820 **Matches:** 130(66) **Sequences:** 32(24)
Eukaryotic initiation factor 4A-I OS=Homo sapiens GN=EIF4A1 PE=1 SV=1
IF4A2_HUMAN **Mass:** 46373 **Score:** 1128 **Matches:** 75(36) **Sequences:** 23(15)
Eukaryotic initiation factor 4A-II OS=Homo sapiens GN=EIF4A2 PE=1 SV=2
IF4A3_HUMAN **Mass:** 46841 **Score:** 978 **Matches:** 37(22) **Sequences:** 18(15)
Eukaryotic initiation factor 4A-III OS=Homo sapiens GN=EIF4A3 PE=1 SV=4
ERF1_HUMAN **Mass:** 49000 **Score:** 214 **Matches:** 6(4) **Sequences:** 5(4)
Eukaryotic peptide chain release factor subunit 1 OS=Homo sapiens GN=ETF1 PE=1 SV=3
EIF3F_HUMAN **Mass:** 37540 **Score:** 57 **Matches:** 4(1) **Sequences:** 2(1)
Eukaryotic translation initiation factor 3 subunit F OS=Homo sapiens GN=EIF3F PE=1 SV=1
EF1A1_HUMAN **Mass:** 50109 **Score:** 44 **Matches:** 1(1) **Sequences:** 1(1)
Elongation factor 1-alpha 1 OS=Homo sapiens GN=EEF1A1 PE=1 SV=1

D **HCH_03101 WT De-amidates eIF4a at the same site as BLF1**

Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Unique	Peptide
715.6820	2144.0242	2144.1113	-0.0872	0	(33)	0.059	1		R.GIDVQQVSLVINYDLPTNR.E + Deamidated (NQ)
1073.1050	2144.1954	2144.1113	0.0841	0	101	8.2e-009	1		R.GIDVQQVSLVINYDLPTNR.E + Deamidated (NQ)
715.7440	2144.2102	2144.1113	0.0988	0	(71)	8.4e-006	1		R.GIDVQQVSLVINYDLPTNR.E + Deamidated (NQ)
715.9740	2144.9002	2145.0954	-0.1952	0	(27)	0.25	1		R.GIDVQQVSLVINYDLPTNR.E + 2 Deamidated (NQ)
716.1100	2145.3082	2145.0954	0.2128	0	(27)	0.79	1		R.GIDVQQVSLVINYDLPTNR.E + 2 Deamidated (NQ)

Figure 7.3.3 – MS/MS Mass-spectroscopy shows that HCH_03101 binds to and de-amidates eIF4a. Panel **A** shows the proteins pulled-down from J774 cells by BLF1 C94S (figure 7.3.1 – lane 5), all three are isoforms of eIF4a. Panel **B** is the peptide that corresponds to the substrate-binding site of BLF1, pulled down with BLF1 C94S no de-amidation is detected. Panel **C** shows the proteins pulled down from J774 cells by WT HCH_03101 (figure 7.3.1 – lane 7), which include all three isoforms of eIF4a along with 3 other components of the translation machinery. Panel **D** shows the peptides matched with eIF4a-I from the HCH_03101 WT pull-down, which correspond to the BLF1 substrate binding site. This peptide in the HCH_03101 WT pull-downs has been de-amidated, with a high scoring match made with a single de-amidation site and a much lower ranking match with a double de-amidation. Curiously there was no peptide identified that matched the mass of the unmodified peptide seen in BLF1 C94S, suggesting that all the available eIF4a had been modified by HCH_03101.

However, one possibility is that it may target the Dinoflagellate genus, a branch of red algae. This is because *H. chejuensis* has previously been shown to secrete a pigment molecule that causes cell lysis in these species (Jeong *et al.,* 2005). As a result the next batch of pull-down experiments were attempted with Algae cell free extract.

7.4.1 – *T. suecica* is an un-sequenced algae species, un-related to the Dinoflagellate genus

Working with Algae presents a multitude of challenges, first amongst these is that many algal species are difficult to culture, moreover there is a shortage of fully sequenced organisms. The only species available locally for these pull-downs was the green Algae *Tetraselmis. suecica*, strain CCAP 66/4, kindly provided by Dr Jim Gilmour (University of Sheffield). However, there is presently no genome sequence for *T. suecica*, which limits the certainty with which hits can be identified. Nevertheless, sequence comparison between human eIF4a and the corresponding protein in *Chlamydomonas. reinhardtii*, a model organism for green Algae, shows 66 % sequence similarity. This means that if HCH_03101 pulls down an algal eIF4a, in all likelihood the peptides analysed by MS-MS Mass-spectroscopy, whilst different, will contain conserved motifs.

7.4.2 – Production of *T. suecica* cell free extract

A quantity of 1 g of *T. suecica* cell pellet was resuspended in 15 ml of breakage buffer and disrupted through 3, 20 second rounds of sonication prior to centrifugation at 60, 000 *g* for 20 minutes in a JLA 25-50 rotor. The cell free extract had a measured concentration significantly lower than anticipated, at 2 mg ml^{-1}. Therefore, the pull-down experiment proceeded as previously described (section 7.3.2) with one alteration; 2 ml (4 mg) of algal cell free extract was incubated with the charged resin instead of 500 µl.

7.4.3 – The *T. suecica* banding pattern is different than observed with J774 macrophages

SDS-PAGE analysis of the *T. suecica* pull-down (figure 7.4.1) shows that when probed with algal cell free extract BLF1 and HCH_03101 behave differently. There are 4 prominent bands eluted from the pull-down that warrant further inspection. Band 1 is a strong band at 46 KDa, the expected size of eIF4a, which is eluted in the salt wash but only present in the two HCH-03101 pull-downs. This band appears to extend into the Imidazole wash, with a pair of bands (band 3) observed at approximately 46 KDa that are present in the BLF1 C94S and both HCH_03101.

The remaining two bands are exclusive to HCH_03101 and are present at approximately 32 KDa (band 2) and 80 KDa (band 4).

7.4.4 – MS/MS Mass-spectrometry is far less conclusive with un-sequenced organisms

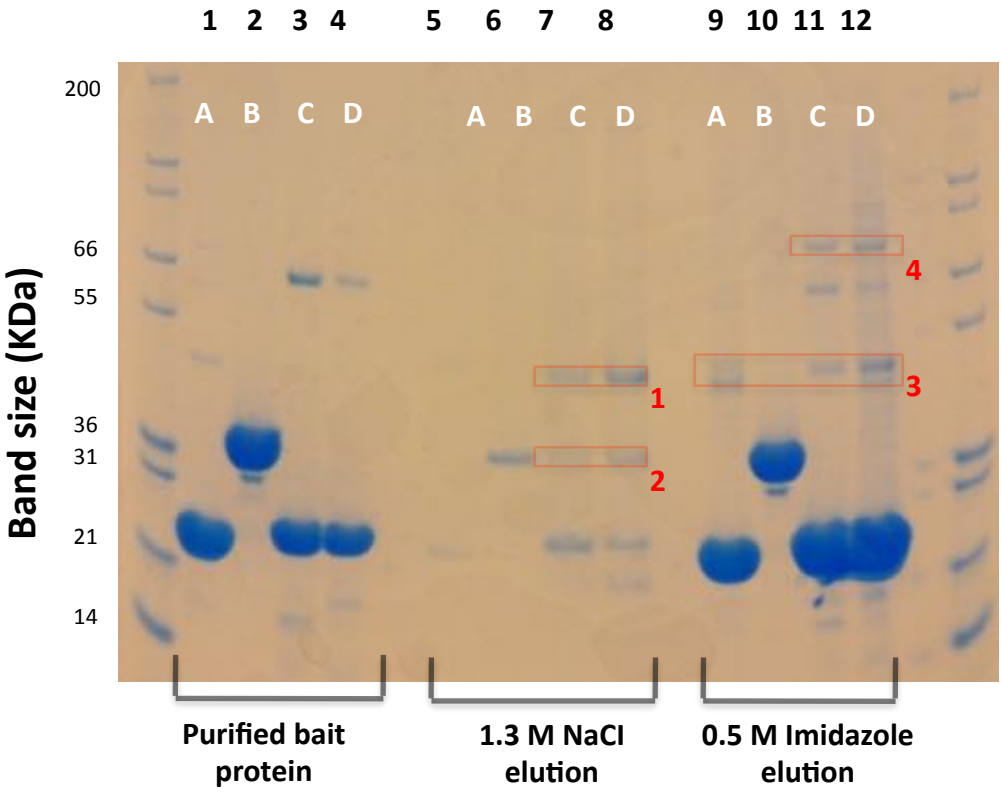
The main challenge with analysing the MS/MS Mass-spectroscopy data from the *T. suecica* peptide fragments is the lack of a genome sequence. This is because MASCOT matches the peptide mass with a database of potential peptide sequences, in this case the NCBI non-redundant database. Therefore, because *T. suecica* has not been sequenced the only significant matches that can be made will be with proteins that are strongly conserved in the probe organism. As a result little emphasis has been placed on either the prot-scores or the number of peptides identified for each hit.

7.4.5 Both BLF1 and HCH_03101 pull-down a selection of translation machinery components

Band 1 from the salt elution is approximately 46 KDa large and exclusive to HCH_03101. MASCOT identifies a high confidence match with the Elongation factor super-family of proteins involved in translation, particularly Elongation factor 1-α-like protein from *Tetraselmis. tetrathele* (figure 7.4.2A). Band 2, also exclusive to HCH_03101, continues the trend pulling down the ribosome biogenesis GTPase RsgA, from *B. subtilis*. Band 3 is shared between BLF1 C94S and both HCH_03101 samples, with band 3 from the BLF1 C94S pull-down also identifying both elongation factors and RsgA (figure 7.4.3A), whilst not pulling down any peptides matching eIF4a. The last band to be examined by MS/MS Mass-spectroscopy was band 4, which is 90 KDa large and exclusively pulled down by HCH_03101. MASCOT identifies only a single non-contaminant match at this position, the Heat shock protein 90 (HS90) super-family. Whilst this family could undoubtedly be a potential toxin target, its relative abundance in most cell types, upwards of 6 % total cellular protein (Crevel *et al.,* 2001) points towards this representing a non-specific interaction.

7.4.6 – HCH_03101 pulls down an Algal eIF4a whereas BLF1 does not

The proteins identified in band 3 from the C94S HCH_03101 pull-down (figure 7.4.3B), are elongation factor 1-α, ribosome biogenesis GTPase RsgA and a selection of Eukaryotic initiation factors including eIF4a. The strongest match with a prot-score of 1049 is an elongation factor 1-α-like protein from *T. tetrathele*.



A = BLF1 C94S, B= YloQ, C = HCH_03101, D = HCH_03101 C94S

Figure 7.4.1 – *Tetraselmis. Suecica* pull-down assay. The cell free extract probe was sourced from *T. suecica* is an un-sequenced species of Algae. The four bait proteins used were BLF1 C94S, YloQ, HCH_03101 WT and HCH_03101 C94S respectively. The figure above shows an SDS-PAGE gel of the pull-down: Lane1, Mark 12 MW marker; lanes 1-4, 15 µg bait protein; lanes 5-8, 15 µl high salt elution; lanes 9-12, 20 µg Imidazole elution. Analysis of the above gel shows that there are 4 interesting bands, highlighted in red. The first two are unique to HCH_03101, with BLF1 not pulling down any binding partners within the salt elution. However, there are additional bands of interest located in the Imidazole wash. Band 3 is shared between both HCH_03101 proteins and BLF1 C94S, and is the same size as eIF4a at 46 KDa. However, on close inspection band 3 is shown to be composed of two bands close together, for MS/MS Mass-spectrometry they have been treated as a single band. Finally there is a 4th band unique to both HCH_03101 proteins with a much higher molecular weight, located exclusively in the Imidazole wash at approximately 90 KDa.

A **Top hits HCH_03101 C94S (Gel band 1)**

gi|148524171 **Mass:** 48986 **Score:** 707 **Matches:** 23(4) **Sequences:** 13(2)
elongation factor-1 alpha-like protein [Tetraselmis tetrathele]

gi|224593227 **Mass:** 35176 **Score:** 464 **Matches:** 15(0) **Sequences:** 10(0)
translation elongation factor-like protein [Tetraselmis striata]

gi|16078641 **Mass:** 33776 **Score:** 330 **Matches:** 5(2) **Sequences:** 5(2)
ribosome biogenesis GTPase RsgA [Bacillus subtilis subsp. subtilis str. 168]

gi|193811886 **Mass:** 48057 **Score:** 158 **Matches:** 4(0) **Sequences:** 3(0)
translation elongation factor-like protein [Raphidiophrys contractilis]

gi|74272649 **Mass:** 50834 **Score:** 139 **Matches:** 3(0) **Sequences:** 3(0)
elongation factor alpha-like protein [Chlamydomonas incerta]

B **Top hits HCH_03101 C94S (Gel band 2)**

gi|284177802 **Mass:** 33777 **Score:** 454 **Matches:** 17(4) **Sequences:** 7(3)
glyceraldehyde-3-phosphate dehydrogenase [Ulva sp. EE2]

gi|16078641 **Mass:** 33776 **Score:** 380 **Matches:** 6(3) **Sequences:** 6(3)
ribosome biogenesis GTPase RsgA [Bacillus subtilis subsp. subtilis str. 168]

gi|515860668 **Mass:** 36450 **Score:** 312 **Matches:** 14(3) **Sequences:** 5(2)
glyceraldehyde-3-phosphate dehydrogenase [Leptolyngbya boryana]

gi|87330881 **Mass:** 41239 **Score:** 125 **Matches:** 5(0) **Sequences:** 3(0)
glyceraldehyde-3-phosphate dehydrogenase subunit A [Scherffelia dubia]

C **Top hits HCH_03101 C94S (Gel band 4)**

gi|15215642 **Mass:** 60807 **Score:** 929 **Matches:** 36(15) **Sequences:** 16(7)
AT5g56010/MDA7_5 [Arabidopsis thaliana]

gi|568215067 **Mass:** 80366 **Score:** 807 **Matches:** 31(15) **Sequences:** 12(7)
Hsp90-2-like [Solanum tuberosum]

gi|439981295 **Mass:** 79711 **Score:** 687 **Matches:** 23(7) **Sequences:** 11(6)
heat shock protein 90 [Salicornia europaea]

gi|16078641 **Mass:** 33776 **Score:** 309 **Matches:** 7(2) **Sequences:** 6(2)
ribosome biogenesis GTPase RsgA [Bacillus subtilis subsp. subtilis str. 168]

Figure 7.4.2 – MS-MS Mass-spec analysis of *T. suecica* pull-down bands 1, 2 and 4. Panel **A** is the top MASCOT hits from band 1, pulled down by HCH_03101. Despite sharing a similar molecular weight with eIF4a all the identified binding partners are alternative translation proteins. Panel **B** is the top MASCOT hits from band 2, pulled down by HCH_03101. The peptides identified here are predominantly dehydrogenase enzymes. However a single component involved at the ribosome is also identified, called RsgA. Panel **C** is the top MASCOT hits from the high molecular weight band 4, pulled down by HCH_03101. Without exception all the specific peptides identified were from the heat shock 90 protein super-family.

A **Top hits BLF1 C94S (Gel band 3)**

gi|148524171 **Mass:** 48986 **Score:** 1317 **Matches:** 48(13) **Sequences:** 24(7)
elongation factor-1 alpha-like protein [Tetraselmis tetrathele]

gi|16078641 **Mass:** 33776 **Score:** 487 **Matches:** 12(4) **Sequences:** 8(3)
ribosome biogenesis GTPase RsgA [Bacillus subtilis subsp. subtilis str. 168]

gi|193811886 **Mass:** 48057 **Score:** 271 **Matches:** 7(2) **Sequences:** 5(2)
translation elongation factor-like protein [Raphidiophrys contractilis]

gi|307931164 **Mass:** 37149 **Score:** 438 **Matches:** 9(3) **Sequences:** 7(3)
translation elongation factor Tu [Tetraselmis chuii]

gi|307931162 **Mass:** 38388 **Score:** 183 **Matches:** 3(2) **Sequences:** 3(2)
translation elongation factor Tu [Pterosperma cristatum]

B **Top hits HCH_03101 C94S (Gel band 3)**

gi|148524171 **Mass:** 48986 **Score:** 1049 **Matches:** 30(8) **Sequences:** 17(7)
elongation factor-1 alpha-like protein [Tetraselmis tetrathele]

gi|552818355 **Mass:** 46611 **Score:** 649 **Matches:** 27(9) **Sequences:** 12(4)
eukaryotic initiation factor 4A (ATP-dependent RNA helicase eIF4A) [Chlorella variabilis]

gi|16078641 **Mass:** 33776 **Score:** 410 **Matches:** 9(3) **Sequences:** 7(3)
ribosome biogenesis GTPase RsgA [Bacillus subtilis subsp. subtilis str. 168]

gi|307931164 **Mass:** 37149 **Score:** 363 **Matches:** 9(2) **Sequences:** 6(2)
translation elongation factor Tu [Tetraselmis chuii]

gi|545358516 **Mass:** 46683 **Score:** 347 **Matches:** 10(3) **Sequences:** 6(3)
eukaryotic initiation factor 4A-like protein [Coccomyxa subellipsoidea C-169]

gi|386758301 **Mass:** 33761 **Score:** 271 **Matches:** 6(2) **Sequences:** 5(2)
ribosome-associated GTPase [Bacillus sp. JS]

gi|198430288 **Mass:** 48770 **Score:** 218 **Matches:** 10(6) **Sequences:** 3(2)
PREDICTED: eukaryotic initiation factor 4A-II [Ciona intestinalis]

gi|11602713 **Mass:** 46306 **Score:** 173 **Matches:** 4(1) **Sequences:** 4(1)
translation initiation factor 4A-like protein [Echinococcus multilocularis]

Figure 7.4.3 - MS-MS Mass-spec analysis of *T. suecica* pull-down band 3. Panel **A** shows the top MASCOT hits for band 3, pulled down by BLF1. The hits from this band correlate with band 1 from the HCH_03101 pull-down, with all the identified proteins belonging to the Eukaryotic elongation factor family. Panel **B** shows the MASCOT hits from band 3, pulled down by HCH_03101. HCH_03101 once again interacts with elongation factors. However, unlike BLF1, HCH_03101 also pulls back eIF4a homologues from several species; the *C. variabilis* hit in particular matches with a high number of unique peptides.

Strong matches are also seen for eIF4a homologues from *Chlorella. variabilis* and *Coccomyxa. subellipsoidea*, which are both species of green algae.

Therefore, HCH_03101 is capable of interacting with algal variants of eIF4a, whereas BLF1 is not. Sequence alignments show that the eIF4a proteins identified in *C. variabilis* and *C. subellipsoidea* share 77 and 71 % sequence similarity with human eIF4a, with both proteins exhibiting strong conservation of the BLF1 binding site. At present it is unclear why HCH_03010 is capable of interacting with these algal eIF4a homologues when BLF1 is not. However, this difference in binding affinity may suggest that HCH_03101 is better suited to targeting eIF4a from these species, or alternatively that it binds its substrates in a different fashion to BLF1 possibly due to the sequence difference observed between the initiation factors. Unfortunately outside of identifying eIF4a from algal sources, the MS/MS Mass-spectrometry data sheds little light on the larger question of whether HCH_03101 can de-amidate algal eIF4a. This is because none of the identified peptides correspond to expected de-amidation site.

7.5 - HCH_03101 pull-down assay with *Dictylostelium. discoideum*

The final organism to be probed against HCH_03101 using the pull-down method was *Dictylostelium. discoideum* a soil dwelling amoeba. This organism was chosen as it exhibits genetic similarity with both animal and plant species, as shown by an EF-1α analysis placing the phylum to which the Dictylostelium belongs, in the animal-fungal glade (Baldauf and Doolittle, 1997). Therefore, with the natural host of *H. chejuensis* still unclear this model organism represents an ideal model to identify potential binding partners.

7.5.1 – Preparation of *D. discoideum* cell free extract

3 g of *D. discoideum* cell paste was kindly supplied by Dr Don Watts (University of Sheffield). This cell paste was resuspended in 8 ml of breakage buffer and broken through 3, 4 second rounds of sonication. Then centrifuged at 20, 000 *g* for 10 minutes to produce a clarified cell free extract. This cell free extract had a measured concentration of 12 mg ml^{-1}. Therefore, 400 µl of cell free extract (4 mg) was incubated with the resin charged with bait protein (section 4.4.1).

7.5.2 – The *D. discoideum* pull-down displays no bands in the high salt elution

The SDS-PAGE analysis of the *D. discoideum* pull-down (figure 7.5.1) shows that none of the previously established bait proteins pull-down anything in the salt elution. There are also no bands at 46 KDa in the Imidazole wash, the anticipated weight of both the Eukaryotic elongation and initiation factors previously pulled down. However, there are novel bands at lower molecular weights in the Imidazole at 20 KDa and 6 KDa, highlighted in red. However, at this point in time these bands have not been analysed through MS/MS Mass-spectrometry.

7.6 – Recombinant eIF4a

Having identified both the functional activity and binding partner of HCH_03101 more stringent assays were required to confirm its functional properties. These additional experiments required a stable source of recombinant eIF4a. Human eIF4a was chosen for two reasons, firstly there was a stock of expression pellets left over from the characterisation of BLF1 and secondly HCH_03101 had been shown to have catalytic activity with this homologue of eIF4a.

7.6.1 – The initial eIF4a over-expressions yielded extremely low levels of protein

The pre-made eIF4a expression construct, kindly provided by Dr Guillaume Hautbergue (University of Sheffield), contained the full-length eIF4a gene nicked with 3' NdeI and 5' Bam HI restriction sites, which were ligated into a pET9 expression vector. However, the expression of recombinant eIF4a from this construct was extremely poor, with no protein band visible at 46 KDa on an SDS-PAGE gel. Despite this set back, attempts were made to purify the trace levels of eIF4a produced. These initial attempts revolved around maximising the possible yield, by minimising the chromatography stages needed to produce a homogenous sample. 6 g of cell paste was resuspended in 40 ml and broken through 3, 20 second rounds of sonication. The eIF4a was then purified using a single affinity chromatography stage, a 2 ml Ni-NTA column (Cube Bioscience^{TM}) charged with BLF1 C94S. The BLF1 affinity matrix was then washed for 2 column volumes with a low Imidazole buffer (50 mM Tris pH 8, 50 mM Imidazole) to remove contaminants. The eIF4a was then eluted with a linear gradient of 0-100 % buffer B (50 mM Tris pH 8, 0.5 M Imidazole) across 10 column volumes, with eIF4a eluted at approximately 0.2 M Imidazole (figure 7.6.1A). However, from the 6 g of cell paste broken this method only yields 0.2 mg of pure eIF4a, which was not a sufficient quantity for the planned experiments.

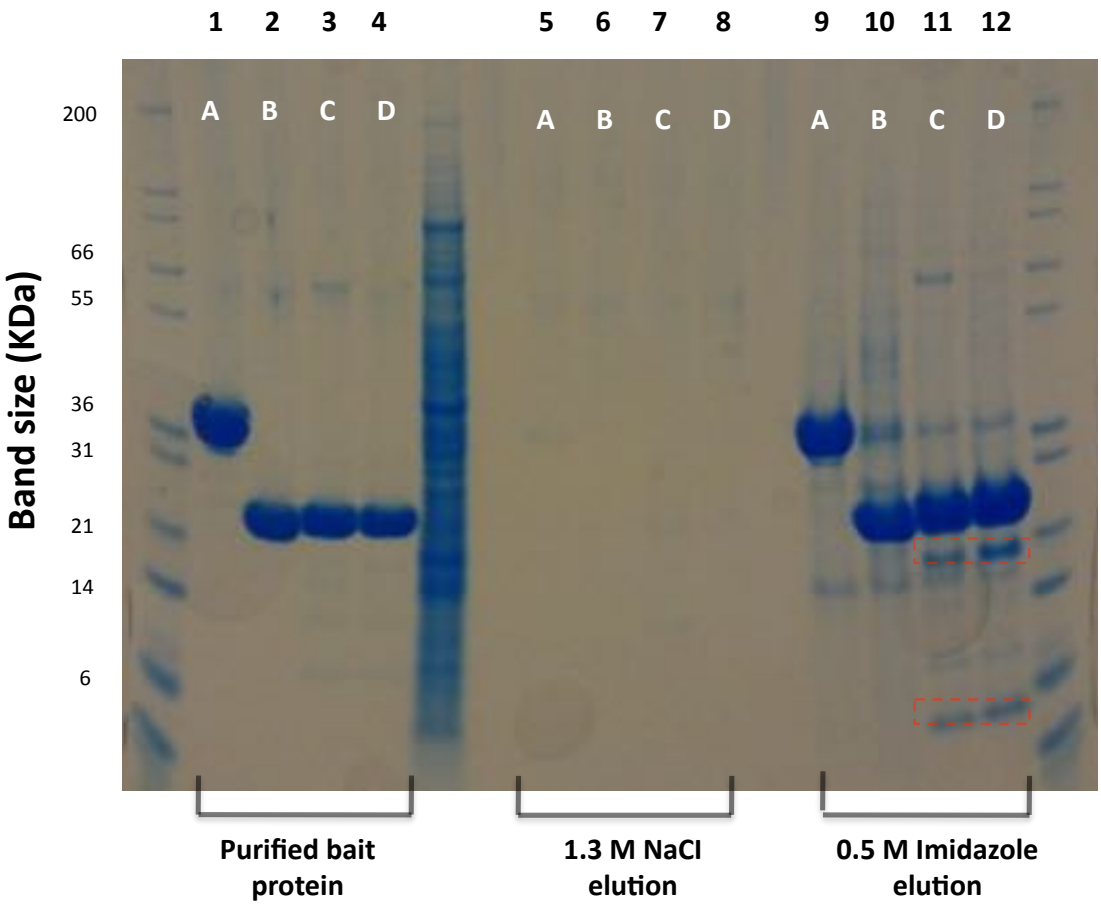


Figure 7.5.1 – *Dictylostelium. Discoideum* pull-down assay. The Eukaryotic amoeba *D. discoideum* was chosen as the third probe organism to test for the binding partners of HCH_03101. The four bait proteins used were YloQ, BLF1 C94S, HCH_03101 WT and HCH_03101 C94S respectively. The pull-down assay is divided into 3 sections the first is the purified bait protein. The second is a 1.3 M NaCl salt wash intended to elute proteins in complex with the bait, with the final stage a 0.5 M Imidazole wash that strips the Ni-NTA beads clean. The SDS-PAGE gel above shows: Lane1, Mark 12 MW marker; lanes 1-4, 15 µg bait protein; lanes 5-8, 15 µl high salt elution; lanes 9-12, 20 µg Imidazole elution. Neither BLF1 nor HCH_03101 appears to bind with any of the previously identified proteins when probed with *D. discoideum* cell free extract. The SDS-PAGE analysis of the high salt elution is clear, with no identifiable bands to analyse across any of the four bait proteins. However there are two bands identified in the Imidazole wash, highlighted in red, which are approximately 18 and 5 KDa large. Neither band corresponds to a similar sized match made in either of the previous pull-downs.

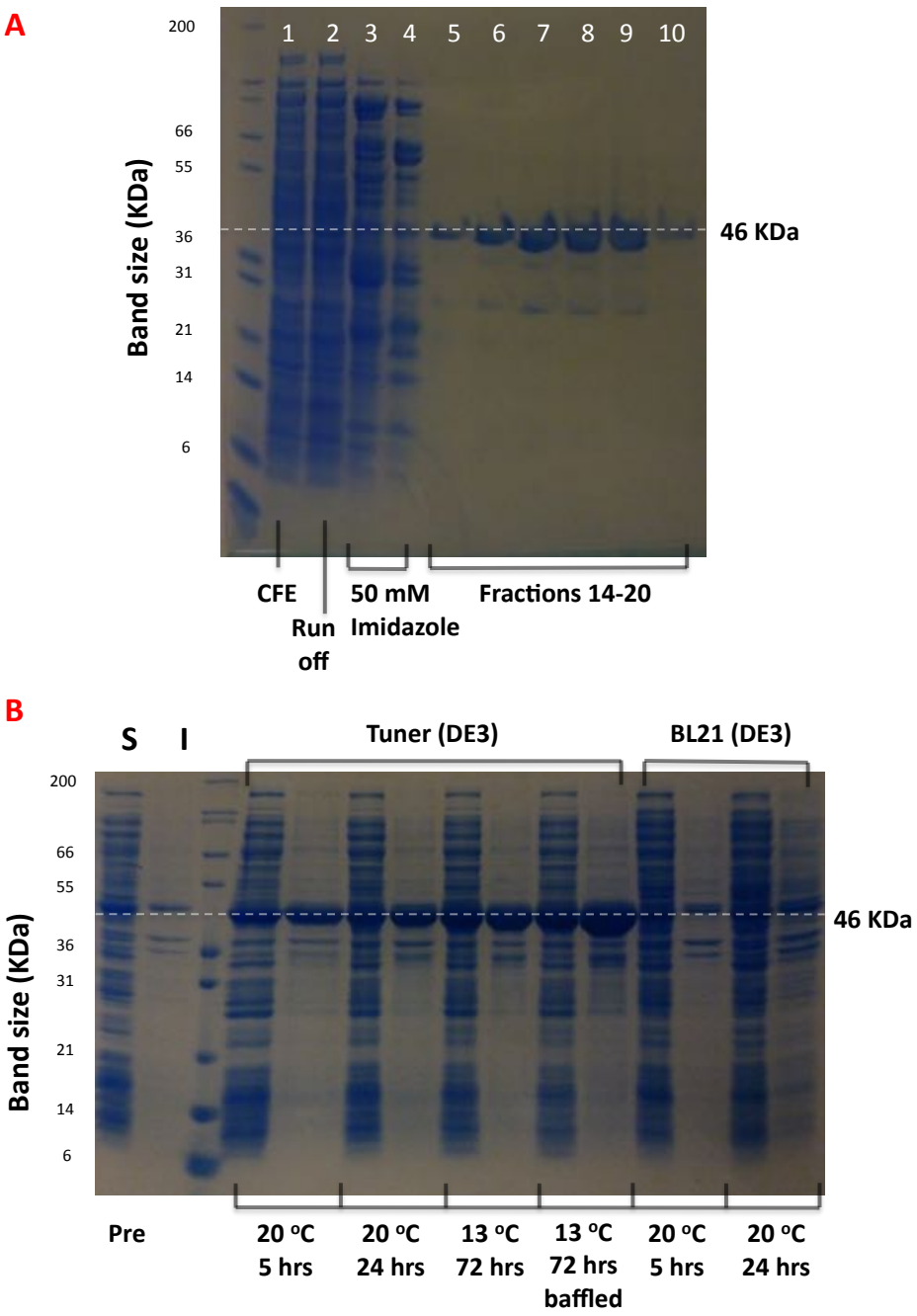


Figure 7.6.1 – BLF1 affinity column purification of eIF4a and over-expression trials of a higher yield eiF4a expression construct. Panel **A** shows an SDS-PAGE analysis of the affinity purification of low yield eIF4a, using a 2 ml Cube Ni-NTA column charged with 6xHIS tagged BLF1 C94S. The column was loaded slowly at 1 ml min^{-1}. Lane 1, 20 µg eIF4a CFE; lane 2, 20 µg BLF1 affinity column run off; lanes 3 and 4, 20 µg 50 mM Imidazole wash; lanes 5-10, 15 µl eluted fractions. Panel **B** shows an SDS-PAGE gel of the over-expression trials undertaken with the eIF4a construct transformed into both BL21 (DE3) and Tuner (DE3) expression cell lines. When expressed in BL21 cells eIF4a is not present. However, in the same conditions when the experiment is repeated using Tuner cells, eIF4a expression is abundant in the soluble fraction. With the best eIF4a expression obtained when incubated at 13 ^{o}C for 72 hours in baffled flasks.

7.6.2 – Transfer of the eIF4a construct into Tuner (DE3) cells yields improved expression

A better protocol for expressing soluble eIF4a was required. To achieve this the plasmid was extracted from 0.1 g of cell pellet using a modified version of the plasmid extraction method (chapter 4.1.10), with the buffer volumes doubled. The extracted plasmid was then sequenced by the core genomics group (University of Sheffield) and transformed into both BL21 (DE3) and Tuner (DE3) strains of *E. coli* for expression trials. Expression trials conducted using the Tuner (DE3) cell line, reveal that this construct is capable of high levels of eIF4a production, with the best expression obtained when incubated in baffled flasks at 13 ^{o}C for 72 hours (figure 7.7.1B). Whereas the same construct expressed in the BL21 (DE3) cell line under the same conditions yields no detectable expression.

7.6.3 – Purification of eIF4a

With improved levels of over-expression, the fresh batch of eIF4a had to be purified using large-scale chromatography methods. eIF4a has a predicted (PROTPARAM) iso-electric point of 5.32, indicating that it is best suited for purification using anion exchange chromatography. A three-step protocol was developed to obtain high purity eIF4a in sufficient quantities to allow co-crystallisation trials to proceed (figure 7.6.2). 4 g of cell paste was resuspended in 30 ml of buffer A (50 mM Tris pH 8) then disrupted through 3, 20 second rounds of sonication. The cell free extract was then separated from the insoluble material by centrifugation at 60, 000 *g* for 20 minutes. The cell free extract was then loaded onto a 5 ml DEAE FF column (GE Healthcare^{TM}), a weak anion exchange matrix, at 5 ml min^{-1}. The eIF4a was then eluted with a linear gradient moving from 0 – 100 % buffer B (50 mM Tris pH 8, 0.5 M NaCl) across 15 column volumes, with eIF4a eluting at 0.29 M NaCl. The protein was spread evenly across 4 fractions (10 ml) and prior to moving onto the next chromatography stage these where concentrated to 5 ml and buffer exchanged into buffer A. The next stage was a 5ml Resource Q column, which was loaded at 2 ml min^{-1} and run on a linear gradient from 0 - 100 % buffer B once again across 15 column volumes, with eIF4a eluted at 0.25 M NaCl. The fractions from this stage were then pooled and concentrated down to a 1 ml volume, prior to being loaded at 1 ml min^{-1} onto a Sepharose 16/60 Gel-filtration column (GE Healthcare^{TM}) which was run for 120 ml at a rate of 1 ml min^{-1}. eIF4a elutes from this column after 82 ml which equates to a monomeric molecular weight of 46 KDa.

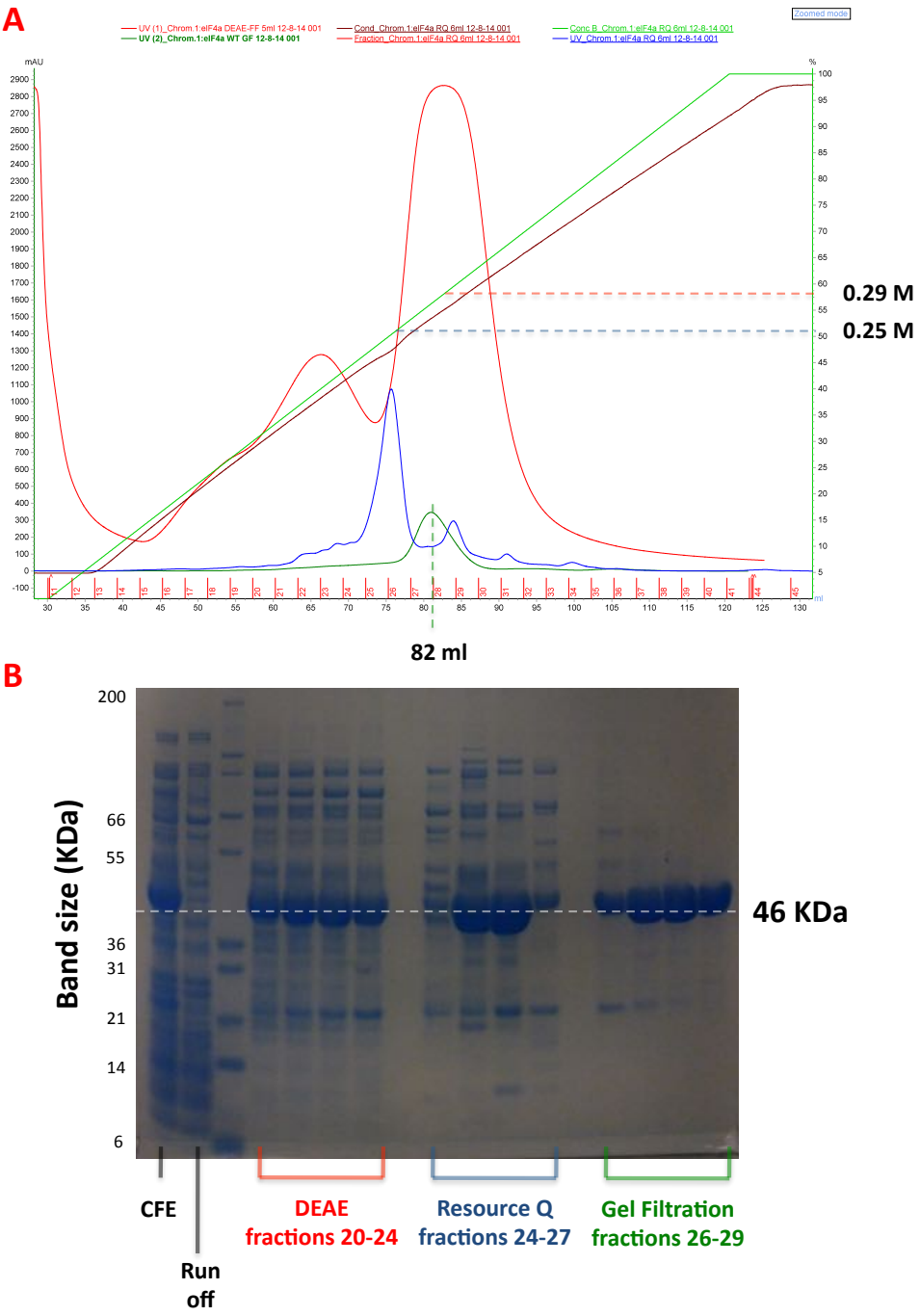


Figure 7.6.2 – Purification strategy for WT eIF4a. Panel **A** shows the purification traces for all three chromatography stages: DEAE FF (red), Resource Q (blue) and Gel-filtration (green). In the case of both anion exchange columns they were run with a linear gradient of 0-100 % B (50 mM Tris pH 8, 0.5 M NaCl) across 15 column volumes. With the eIF4a eluting with 0.29 and 0.25 M NaCl from the DEAE FF and Resource Q columns respectively. The final stage was passage through a Sepharose 16/60 Gel-filtration column, with eIF4a eluted after 82 ml, which corresponds to a monomeric molecular weight of 46 KDa. Panel **B** shows an SDS-PAGE analysis of the above purification protocol, with the final pooled Gel-filtration fractions > 95 % pure and suitable for future crystallisation experiments.

The SDS-PAGE analysis of the above purification (figure 7.6.2B) shows that the eIF4a left at the end is extremely pure. However, in future purifications the Resource Q column could be exchanged for a HiTrap Q-HP 5ml column (GE Healthcare^{TM}), a midrange anion exchange resin. This is because the purity gains made with the Resource Q stage are vastly outweighed by the loss in yield, with only 18 mg being returned from the 70 mg of protein loaded.

7.7 – eIF4a binding assays

The previously described pull-down assays have shown that eIF4a binds with both BLF1 and HCH_03101, but up till this point the low yield of protein obtainable from the recombinant expression strain of eIF4a has prevented in depth studies of the interactions formed between these proteins.

7.7.1 – Complex formation with eIF4a does not improve HCH_03101 solubility

A quick first experiment to attempt with the recombinant eIF4a was to attempt complex formation with HCH_03101. The first attempt at mixing the two proteins was at high concentration, 100 µl of a 20 mg ml^{-1} solution of purified eIF4a (20 mM Tris pH 8) was mixed with 200 µl of an 8 mg ml^{-1} solution of purified HCH_03101 (20 mM Tris pH 8, 100 mM NaCl). Almost immediately the mixture precipitates, with the protein concentration measured at 4 mg ml^{-1} after 10 minutes centrifugation at 20, 000 *g*. Currently the best method of combining these two proteins is to slowly add a moderately concentrated eIf4a solution (5 mg ml^{-1}) into a low concentration HCH_03101 (1 mg ml^{-1}) solution, before concentrating them together. However, while it is possible to then concentrate these proteins alongside one another, any complex formation that maybe occurring has no positive effect upon the solubility issues prevalent when working with HCH_03101. Therefore, if the intention is to examine HCH_03101 interacting with its substrate any potential experiment has to be planned with limiting time and immobilisation constraints in mind.

Given the limited time available (July 2014) two experiments were selected. The first was a comparison of gel-filtration elution profiles, between the constituent proteins and the putative complexes formed (section 7.7) . The second experiment was to attempt co-crystallisation of the complex (chapter 8). These experiments were conducted back to back and with both BLF1 and HCH_03101 for comparative value. Of particular note was the compromised separation observed during size exclusion chromatography using a superdex 200 16/60 column, likely

caused by pressure compression. Consequently, the following observations pertaining to BLF1 and HCH_03101s affinity to eIF4a only offer a guideline to their relative specificity.

7.7.2 – BLF1 C94S purification

The Gel-filtration and co-crystallisation trials are the first experiments covered in this thesis that have been conducted with BLF1. There were ample frozen supplies of WT BLF1 but the C94S mutant had to be purified as required. The BLF1 C94S construct expresses well making up approximately 10 % of the total protein content. The purification protocol for BLF1 is established (Cruz-Migoni *et al.,* 2011) and utilises two chromatography stages. 2 g of cell pellet was resuspended in 20 ml of buffer A (50 mM Tris pH 8) and broken through 3, 20 second rounds of sonication. The cell debris was then separated by centrifugation at 60, 000 *g* for 20 minutes in a JLA 25-50 rotor. The cell free extract was then loaded onto a DEAE FF 5 ml column at 5 ml min^{-1} with the protein eluted through a 0-100 % linear gradient of buffer B (50 mM Tris pH 8, 0.5 M NaCl) across 15 column volumes, BLF1 C94S eluting at 0.17 M NaCl (figure 7.7.1A). The DEAE fractions were then pooled and concentrated to 1 ml before being loaded onto the Sepharose 16/60 Gel-filtration column and run for 120 ml at 1 ml min^{-1}, BLF1 C94S is eluted at 88 ml which equates to a monomeric molecular weight of 22 KDa (figure 7.7.1B). SDS-PAGE page analysis of this purification shows that the BLF1 C94S produced is 90 % pure in the best two fractions 26 and 27 (figure 7.7.1C). Future purifications may benefit from an additional anion chromatography stage, possibly a Q-sepharose column.

7.7.3 – WT BLF1 forms a loose complex with eIF4a

For this experiment eIF4a was purified freshly (section 7.6.3) and during this purification a control trace was recorded for eIF4a on its own (figure 7.7.2A – blue). The WT BLF1 sample was from a frozen stock, but the C94S mutant was purified freshly (section 7.7.2) and acts as a second control (figure 7.7.2A – pink). 1 mg of BLF1 WT was mixed with 2 mg of eIF4a in approximately 1:1 molar ratios at low < 1 mg ml^{-1} concentrations in a 50 mM Tris pH 7.5, 0.1 M NaCl buffer. This mixture was then incubated at 20 ^{o}C for 1 hour prior to concentration down to a 1 ml volume for loading onto the Sepharose 16/60 Gel-filtration column. The Gel-filtration column was then run for 120 ml at 1 ml min^{-1} in a 50 mM Tris pH 8, 0.5 M NaCl buffer, with a single peak eluted at 89 ml (figure 7.7.2A – purple). The gel filtration trace shows no second peak corresponding to BLF1 on its own, however the single peak observed elutes at exactly the same volume as eIF4a on its own. 13 µl of eluate from the peak fractions (24-30), were run out

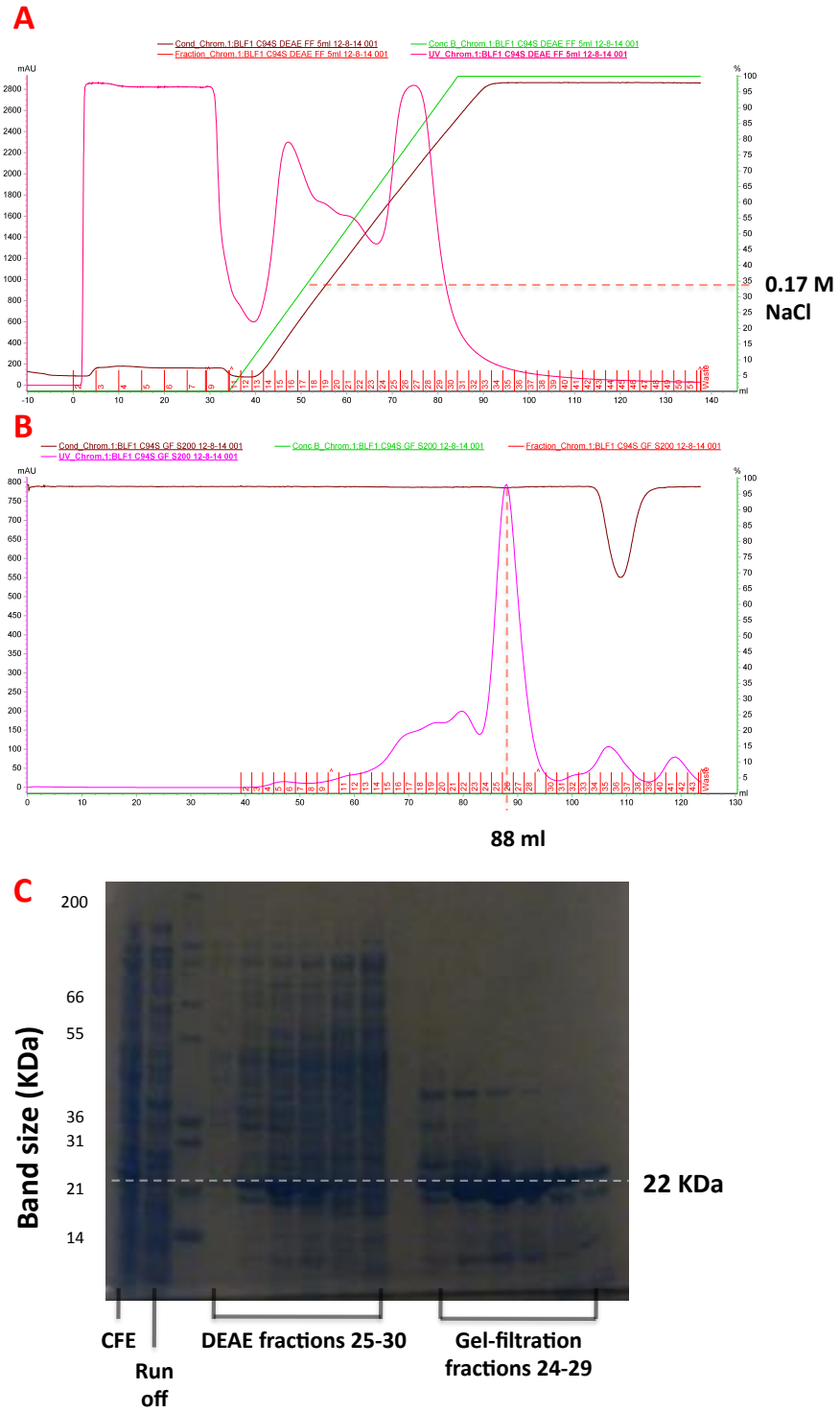


Figure 7.7.1 – Purification of BLF1 C94S. Panels **A** and **B** are the purification traces for the DEAE FF 5ml and Gel-filtration columns respectively. The DEAE FF 5 ml column was run on a linear gradient from 0-100 % buffer B (50 mM Tris pH 8, 0.5 M NaCl) across 15 column volumes, with BLF1 C94S eluting at 0.17 M NaCl. The Sepharose 16/60 size exclusion column was run for 120 ml, with BLF1 C94S eluted at 88 ml corresponding to a monomeric molecular weight of 22 KDa. Panel **C** is an SDS-PAGE gel of the above purification scheme showing that the BLF1 C94S produced in the best two fractions, 26 and 27, is approximately 90 % pure.

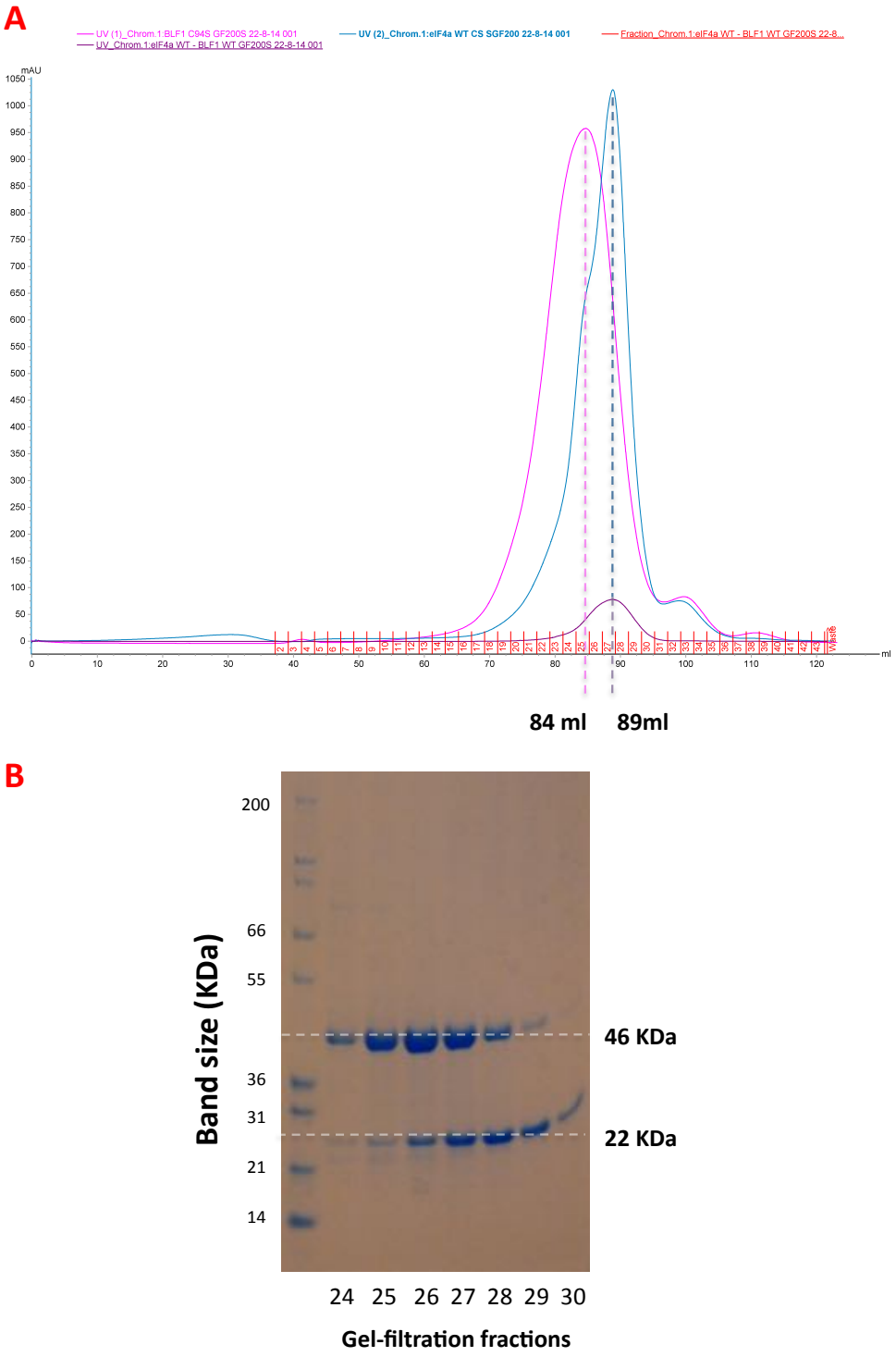


Figure 7.7.2 – Gel-filtration analysis of the complex formation between WT BLF1 and eIF4a. Panel **A** shows the elution profiles of eIF4a (blue), BLF1 C94S (pink) and eIF4a in complex with BLF1 WT (purple) run through the same Sepharose 16/60 Gel-filtration column. Independently eIF4a elutes at 89 ml and BLF1 C94S elutes at 84 ml, when eIF4a is combined with WT BLF1 and incubated for 1 hour at 20 ^{o}C the resulting complex elutes as a single peak at 89 ml. Panel **B** shows an SDS-PAGE gel of fractions 24-30, which correspond to the eluted eIF4a-BLF1 WT peak. The distribution of the two component proteins within these fractions is not balanced with only a single fraction (27) appearing to contain a 1:1 stoichiometric mixture.

on a SDS-PAGE gel, confirming that both eIF4a and BLF1 are present within the single observed peak. However, closer inspection reveals that the interaction is transient, under the 0.5 M NaCl buffering conditions the Gel-filtration was run in, with BLF1 appearing to elute later then the eIF4a.

7.7.4 – C94S BLF1 appears to form a tighter complex with eIF4a

BLF1 C94S was purified freshly for this experiment (section 7.7.2) with the Gel-filtration trace retained as a control (figure 7.7.3A – pink) along with the previous eIF4a trace (figure 7.7.3A – blue). 1 mg of BLF1 C94S was incubated alongside 2 mg of eIF4a at low (1 mg ml^{-1}) concentration, in incubation buffer (50 mM Tris pH 7.5, 0.1 M NaCl) at 20 ^{o}C for 1 hour. Then concentrated down to a 1 ml volume before being loading onto the Sepharose 16/60 Gel-filtration column. The Gel-filtration column was run across 120 ml at a rate of 1 ml min^{-1} in buffer B (50 mM Tris pH 8, 0.5 M NaCl), with a single peak eluted at 86 ml (figure 7.7.3A – purple). The comparative gel-filtration analysis shows that the BLF1 C94S – eIF4a complex is eluted 3 ml earlier than the WT BLF1 equivalent at 89 ml. Inspection of the SDS-PAGE gel for the peak, fractions 23-29, reveals that more of the fractions share a 1:1 stoichiometric mixture of the two components in complex than observed with WT BLF1.

7.7.5 – HCH_03101 C94S appears to bind eIF4a as tightly as BLF1 C94S

HIS tagged HCH_03101 C94S was purified freshly (chapter 5.3.1) using a Ni-NTA column, but due to solubility imposed time constraints it was not run independently down the Gel-filtration column. The controls for this experiment are eIF4a on its own (figure 7.7.4A – blue) and the complex formed between eIF4a and BLF1 C94S (figure 7.7.4A – purple). The pooled fractions from the Ni-NTA column were concentrated from 12 ml down to 2 ml, then diluted in 15 ml of incubation buffer (50 mM Tris pH 7.5, 0.1 M NaCl) to reduce the salt and Imidazole present during incubation. The HCH_03101 sample was calculated to contain 14 mg of protein so 30 mg of eIF4a were gradually added to the sample at low concentration. Unlike the BLF1 samples that were incubated for 1 hour at 20 ^{o}C, HCH_03101 C94S and eIF4a were incubated for 3 hours at 20 ^{o}C, which was the time the sample required to concentrate ready for loading onto the Gel-filtration column. The incubated complex was then loaded and run on a Sepharose 16/60 Gel-filtration column for 120 ml at 1 ml min^{-1}, with a single peak eluted at 86 ml (figure 7.7.4A – green). Examination of both the trace and SDS-PAGE gel for the eluted fractions (figure 7.7.4B) shows that HCH_03101 C94S interacts with eIF4a as strongly as its BLF1 counterpart, with both eluting after 86 ml.

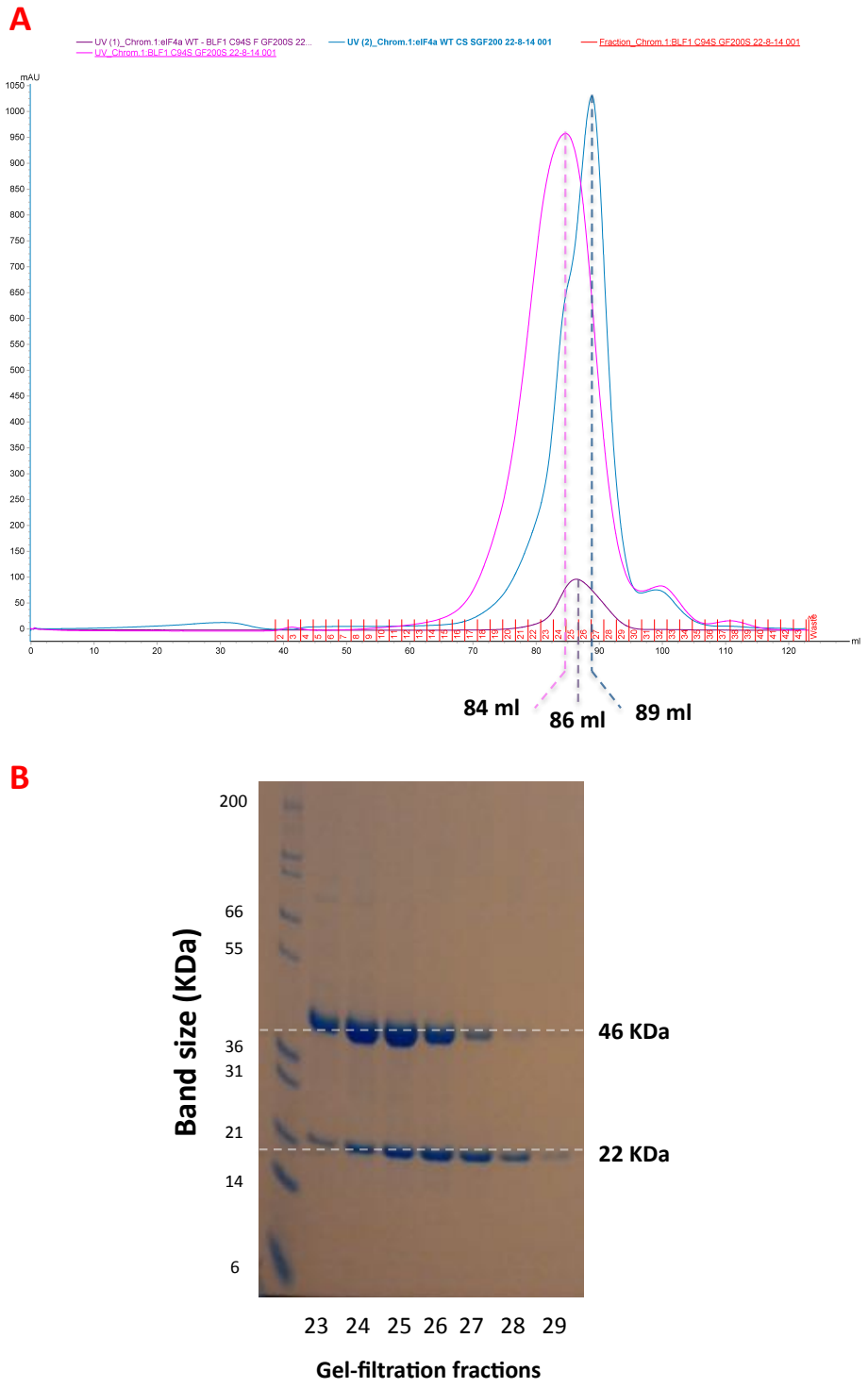


Figure 7.7.3 – Gel-filtration analysis of complex formation between BLF1 C94S and eIF4a. Panel **A** shows the purification traces of eIF4a (blue), BLF1 C94S (pink) and eIF4a in complex with BLF1 C94S (purple) run through the same Sepharose 16/60 Gel-filtration column. BLF1 C94S on its own elutes at 84 ml and eIF4a independently elutes at 89 ml. The complex formed between the two elutes at 86 ml in the middle of these two peaks. Panel **B** shows an SDS-PAGE analysis of the eluted fractions, corresponding to the eIF4a-BLF1 C94S peak. The distribution of the two component proteins within these fractions is relatively balanced with 3 fractions (24-26) appearing to contain a 1:1 stoichiometric mixture of both eIF4a and BLF1 C94S.

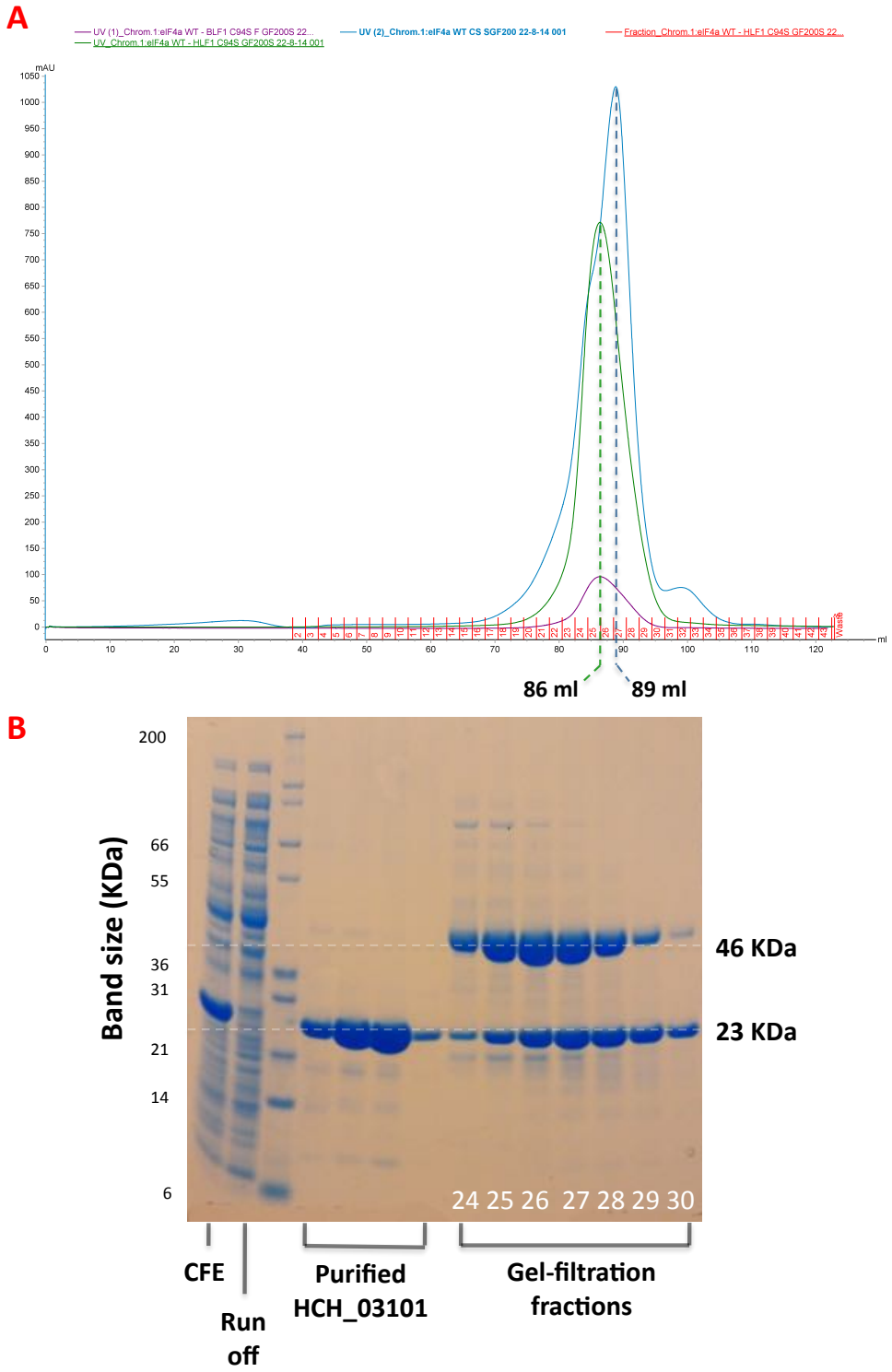


Figure 7.7.4 - Gel-filtration analysis of complex formation between HCH_03101 C94S and eIF4a. Panel **A** shows the elution traces of eIF4a (blue), eIF4a – BLF1 C94S (purple) and eIF4a in complex with HCH_03101 C94S (green), which have been run through the same Sepharose 16/60 Gel-filtration column. The eIF4a – HCH_03101 C94S complex elutes at the same point as the eIF4a – BLF1 C94S complex before it at 86 ml. Panel **B** is an SDS-PAGE gel showing fractions 24-30, which correspond to the eIF4a-HCH_03101 C94S peak. This peak contains 3 fractions (26-28) with approximately equivalent molar ratios of the component proteins, eIF4a and HCH_03101.

7.8 – Co-crystallisation trials for the eIF4a - BLF1 / HCH_03101 complex

The complex samples examined in the previous Gel-filtration experiments were then immediately concentrated into crystallisation (20 mM Tris pH 7.5, 0.1 M NaCl buffer). The two BLF1-eIF4a samples were both concentrated to 20 mg ml^{-1}, but it was only possible to concentrate the HCH_03101 C94S-eIF4a complex to 8 mg ml^{-1}. All three samples were then placed in sparse matrix crystallisation trials across the PACT, JCSG+, Classics, MPD and Proplex screens. For each screen 4 robot trays were laid down between two varying drop sizes, 400 and 700 nl large, with one of each droplet size incubated at 8 ^{o}C or 17 ^{o}C. The methodology used above failed to produce co-crystals of either WT BLF1 or C94S BLF1 or HCH_03101 in complex with eIF4a. However efforts undertaken with a four fold excess of BLF1 C94S at the incubation stage yielded well diffracting crystals discussed in depth in chapter 8.

7.9 – Conclusions

This chapter has detailed a variety of pull-down assays, encompassing cell free extract probes from a diverse range of Eukaryotic cell types. Pull-downs conducted with J774 macrophage cells, show that HCH_03101 can bind to and de-amidate eIF4a, a trait shared with the Glutamine de-amidase toxin BLF1. Binding assays conducted with recombinant human eIF4a, comparing complex formation with both BLF1 C94S and HCH_03101 C94S show that HCH_03101 forms a tight complex with eIF4a closely matching the complex formed with BLF1. However, when probed against algal cell free extract HCH_03101 pulls down an algal homologue of eIF4a where BLF1 cannot. This is surprising because the identified algal homologues of eIF4a display strong conservation at the BLF1 binding site. This suggests that HCH_03101 interacts with eIF4a in a different fashion to BLF1, which might be expected given the differences observed flanking their respective active site clefts.

Presently not enough is known about *H. chejuensis*, the microorganism that produces HCH_03101, to assign a target organism. However, successful pull-down experiments from both macrophage and algal cell free extracts point towards HCH_03101 targeting a Eukaryotic species. Therefore, HCH_03101 would appear to conform to all the traits, structural and functional, typically associated with the toxin members of the Glutamine de-amidase super-family. Furthermore, the lack of any signalling peptides, a property shared with BLF1, indicates that HCH_03101 may share the same delivery mechanism, which suggests that *H. chejuensis* may be an intracellular pathogen.

7.10 – Future work

There are multiple avenues for following on from the work presented in this thesis, which are listed below in order of both urgency and personal interest.

7.10.1 – BLF1 or HCH_03101 co-crystallisations in complex with eIF4a

eIF4a has been shown to form a tight complex with both BLF1 and HCH_03101 (section 7.7), particularly the C94S active mutants of each respective protein. Therefore, the co-crystallisation experiments currently underway (section 7.8) are of particular importance. This is because at the present there are no structures showing any of the Glutamine de-amidase toxins in complex with their substrates, as a result how these toxins bind to or effect catalysis in their substrates is currently poorly understood. A high-resolution structure of the complex could potentially aid in assigning catalytic roles to several conserved portions of the active site. The conformation of the active site dyad upon binding to eIF4a would be particularly interesting to observe, as would any change in conformation of the β-protrusion in HCH_03101.

7.10.2 – HCH_03101 activity assays

In order to fully characterise HCH_03101 as a functioning toxin it is going to be necessary to conduct a suite of activity assays that quantify the rate of de-amidation. There are presently two different assays that could be attempted to ascertain some basic rate measurements. The first utilises Mass-spectroscopy to quantify the ratio of native and de-amidated substrate eIF4a, at selected time points. The second measures the helicase activity of eIF4a and can in-directly confer a measurement of the enzymes activity.

7.10.3 – Characterisation of *H. chejuensis* as an intracellular organism

The discovery of HCH_03101 raises the possibility that *H. chejuensis* is an intracellular pathogen. Whilst outside my personal area of expertise it would be interesting to conduct experiments with the aim of ascertaining both the pathogenesis and lifestyle of *H. chejuensis*. These experiments could range from macrophage killing assays, through to visualising the bacteria gaining access to the host cell using electron microscopy.

7.10.4 – Binding assays with the assortment of targets not pursued from the pull-downs

The pull downs described earlier in this chapter, particularly the *T. suecica* algal pull-down showed that HCH_03101 pulls down several proteins besides eIF4a. Recombinant expression constructs are required for the elongation factor 1-α, ribosome biogenesis GTPase RsgA and Hsp90 hits, so that assays can be conducted to show whether complex formation or de-amidation occurs when these proteins are incubated with HCH_03101.

7.10.5 – Structure determination of the previously identified Glutamine de-amidase candidates

In chapter 5 the cloning and over-expression trials for 4 alternative Glutamine de-amidase candidates, which did not progress through to structure determination, were described. The reason for this was that none of these enzymes expressed in the soluble fraction during preliminary expression trials. However, they all warrant further investigation. Therefore, if suitable expression conditions could be identified then the subsequent experiments would follow a similar pattern to those described to characterise HCH_03101. Starting with structure determination before moving onto functional characterisation through pull-down assays.

7.10.6 – Identification of novel Glutamine de-amidase enzymes with an improved search primary sequence search motif

Analysis of the structure of HCH_03101 highlighted two residues that are conserved across multiple Glutamine de-amidase enzymes, which were close to the active site. Both of these positions could be used to strengthen the current search motif, to aid in the discovery of novel Glutamine de-amidase enzymes. The updated motif is shown below with the new additions coloured red:

G - X(4-5) – K – X(8-15) – {KRDEH}(4) – L – [AGSTV](2) – C – X(10-16) – [FGIPLWAMV] – H – {KRDEH}

Chapter 8 – Structure determination of BLF1 in complex with its substrate eIF4a

The following chapter details structural studies undertaken on a stable complex formed between an inactive C94S mutant of the glutamine de-amidase toxin BLF1, with a modified construct of its substrate eIF4a. The WT eIF4a construct referred to throughout this chapter incorporates a short NTD truncation removing the first 20 residues thought to be disordered.

8.1 – Production of a stable BLF1 C94S – eIF4a complex

In chapter 7.7 and 7.8 attempts to form a complex between BLF1 C94S and WT eIF4a were shown with an approximately 1:1 stoichiometric mixture observed on SDS-PAGE gels after incubation and co-purification via size exclusion chromatography, whilst in the presence of high 0.5M NaCl (figure 7.7.3B). However, despite extensive attempts to crystallise this observed, presumably stable complex, none of the crystals produced contained eIF4a.

As previously discussed in section 7.7, the superdex 200 16/60 gel filtration column used for the complex formation studies exhibited compromised separation characteristics. However, further co-purifications undertaken with an undamaged column, show that an inactive C94S mutant of BLF1 forms an extremely stable complex with eIF4a (figure 8.8.1A). A 1:4 molar mixture of pure homogenous WT eIF4a with mutant BLF1 C94S was incubated at 4 ^{o}C for 45 minutes in 50 mM TRIS pH 7.5 then concentrated to 10 mg ml^{-1}, of the eIF4a component determined by Bradford assay (BLF1 does not react with Bradford reagent), then run on a Superdex 200 16/60 Gel-filtration column in a low salt buffer (50 mM TRIS pH 8, 0.1 M NaCl). An SDS-PAGE gel was then run (figure 8.8.1B), with the fractions containing the correct ratio of the two proteins pooled and buffer exchanged into a minimal buffer (10 mM TRIS pH 7.5) for crystallisation.

8.2 – Crystallisation and structure solution of the BLF1 C94S – eIF4a complex

8.2.1 – Crystallisation and data collection of the BLF1 C94S – eIF4a complex

The pooled fractions from the co-purification were further concentrated to 12 mg ml^{-1} and trialled in several sparse matrix screens (JCSG+, PACT, MPD, Proplex and Classics) with two differing drop sizes (400 and 700 nl) with each screen and drop size incubated at both 8 and 17 ^{o}C. There were several different crystal forms tested at Diamond (2/10/14) on beamline I04. However, only one contained both protein components the rest were BLF1 in isolation. The complex crystals were grown in the Proplex screen,

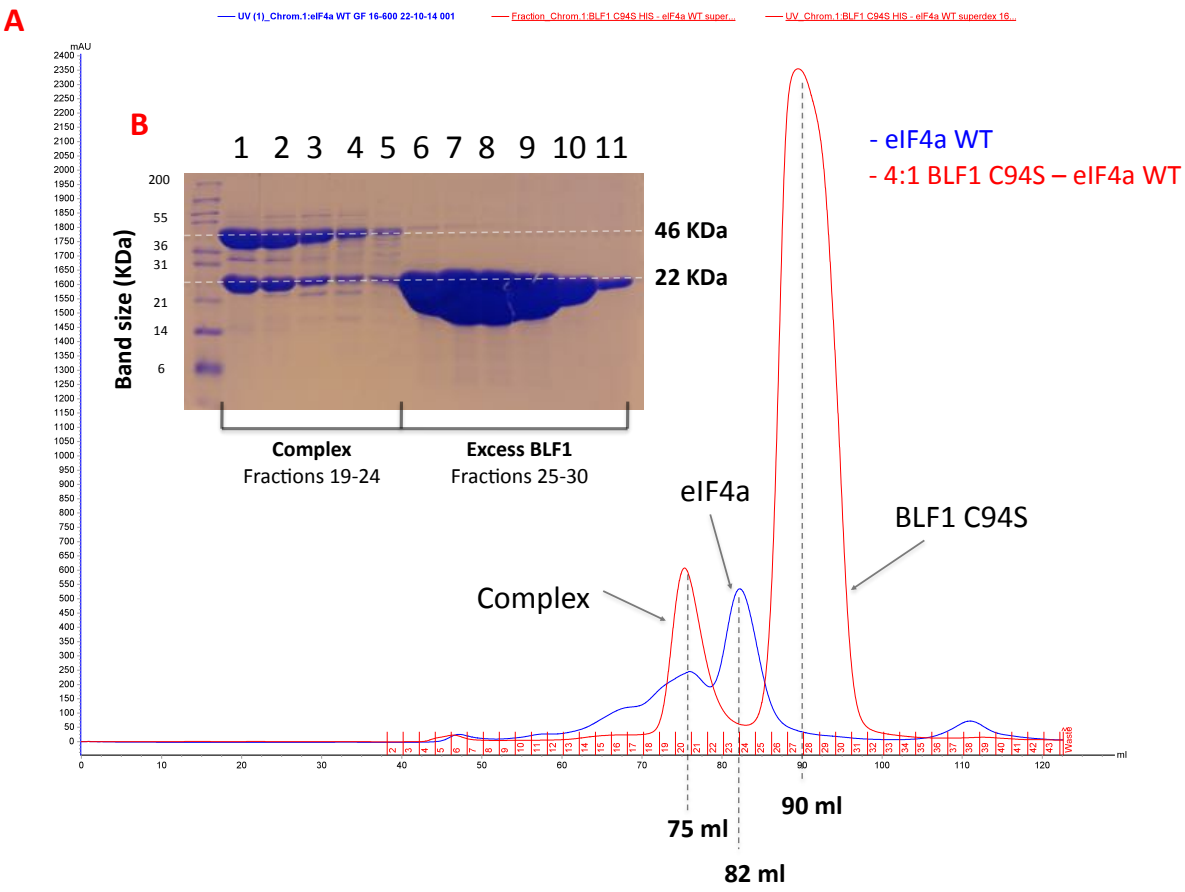
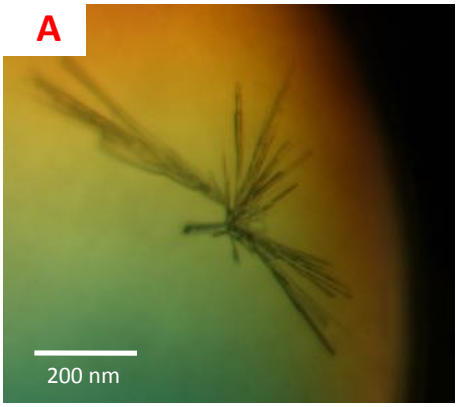


Figure 8.1.1 – An inactive mutant BLF1 C94S forms a stable complex with WT eIF4a. Panel **A** shows the elution profile of eIF4a (blue) and a mixture containing a 4:1 molar excess of BLF1 C94S with eIF4a (red), from a Superdex 200 16/60 size exclusion column. eIF4a is a 46 KDa protein that elutes exclusively as a monomeric species at 82 ml. However, there is no peak observed at 82 ml when eIF4a is incubated with BLF1. Instead two clear peaks are observed at 75 and 90 ml, corresponding to the 68 KDa BLF1 C94S – eIF4a complex and the excess monomeric 22 KDa BLF1 C94S respectively. Panel **B** shows an SDS-PAGE analysis of the fractions collected from the above size exclusion column, lanes 1-5 correspond to the first peak eluted at 75 ml with lanes 6-11 matching the second peak at 90 ml. In peak 1 (75 ml) there are bands observed with masses at both 22 and 46 KDa, which remain in an approximately 1:1 stoichiometric ratio across the entire peak from high to low concentration. Peak 2 (90 ml) on the other hand is comprised exclusively from the excess BLF1 C94S, with no eIF4a present. Curiously there is no peak at 82 ml post incubation at the expected mass of unbound eIF4a, which suggests that all the eIF4a is present in complex with BLF1 C94S.

A



Proplex H10 – 50 mM MgCl2, 0.1 M MES pH 6.5, 10 % (v/v) 2-propanol, 5 % (w/v) PEG 4000

B

BLF1 C94S – eIF4a WT - Native			
	Overall	**Low**	**High**
High resolution limit (Å)	2.52	11.27	2.52
Low resolution limit (Å)	63.07	63.07	2.59
Completeness	99.2	94.9	98.7
Multiplicity	3.8	3.3	3.6
I / sigma (I)	8.1	30.2	1.7
R_{pim} (I)	0.096	0.022	0.441
Wilson B factor (Å²)	8.4		
Total observations	76,647	803	5285
Total unique observations	20,378	244	1457
Unit cell dimensions			
A, B, C (Å)	135.98	50.36	95.74
α, β, γ (°)	90.0	111.9	90.0
Space group	C2		

Figure 8.2.1 – Crystallisation condition and processing statistics for BLF1 C94S – eIF4a co-crystals. Panel **A** details the crystals obtained in the sitting drop sparse matrix screen Proplex, condition H10 (50 mM $MgCl_2$, 0.1 M MES pH 6.5, 10 % (v/v) 2-propanol, 5 % (w/v) PEG 4000). This condition displayed crystals with a fine rod morphology extending from a single nucleation event. The crystals were cryo-protected through a 10 second incubation in a separate drop containing mother liquor plus 30 % (v/v) ethylene glycol, then plunged rapidly into LN_2. Panel **B** details a dataset incorporating 200 ^{o} of rotation in 0.1 ^{o} oscillations, collected on beam line I04 at the Diamond light source. The crystals are monoclinic, displaying a C2 space group with unit cell dimensions of 135.9, 50.4 and 95.7 (A, B and C), with resolution reaching an upper limit of 2.5 Å. The only unusual measurement is the low Wilson B Factor of 8.4, with strong signal to noise and merging statistics indicating that the data collected are suitable for structure determination.

condition H10 (50 mM $MgCl_2$, 0.1 M MES pH 6.5, 10 % (v/v) 2-propanol, 5 % (w/v) PEG 4000), at 17 ^{o}C in the larger 700 nl drop size (figure 8.2.1A). These crystals were frozen in a cryo-protectant containing the mother liquor plus 30 % (v/v) ethylene-glycol. Initial tests revealed that the crystals exhibited a C2 space group with unit cell dimensions of 135.9, 50.4 and 95.7 (A, B and C). A dataset incorporating 200 ^{o} of rotation in 0.1 ^{o} oscillations was collected to a resolution of 2.5 Å, encompassing the entire until cell with a measured multiplicity of 3.8 (figure 8.2.1B).

8.2.2 – Structure solution of the BLF1 C94S – eIF4a complex

Having collected a dataset that measured the whole unit cell, a molecular replacement phasing strategy was employed. However, initial attempts to solve the structure with either BLF1 or eIF4a in isolation proved unsuccessful. With BLF1 yielding a correct solution but poor electron density features, which was expected as it only accounts for 1/3 of the unit cell contents. Conversely, it was not possible to match full-length eIF4a search ensembles, indicating that complex formation with BLF1 C94S induces substantial conformational changes in eIF4a.

The eventual molecular replacement strategy involved two stages, each incorporating two search ensembles. First an incomplete partial solution was produced, by searching against BLF1 and the NTD of eIF4a (residues 63-231). This partial solution was then passed forward to a second round of molecular replacement with the CTD of eIF4a (residues 240-389) used as the second ensemble. However, as the two-domain structure of eIF4a is the result of probable gene duplication, the NTD and CTD exhibit extremely similar tertiary folds. To guard against incorrect orientations of the two domains both the NTD and CTD search ensembles were heavily truncated at both the N and C terminus to allow for accurate determination of a correct solution. The molecular replacement solution determined using the above ensembles yields a high quality electron density map with ample novel density features, corresponding to the truncated regions not factored into search ensemble (figure 8.2.2).

8.3 – Structure of the BLF1 C94S – eIF4a complex

8.3.1 – Model building and validation

The initial map revealed that the overall structure of BLF1 was largely unchanged from the search ensemble (PDB ID: 3TU8). However, there were significant differences observed in the eIF4a portion of the structure. Consequently, a further 100 residues located between the NTD

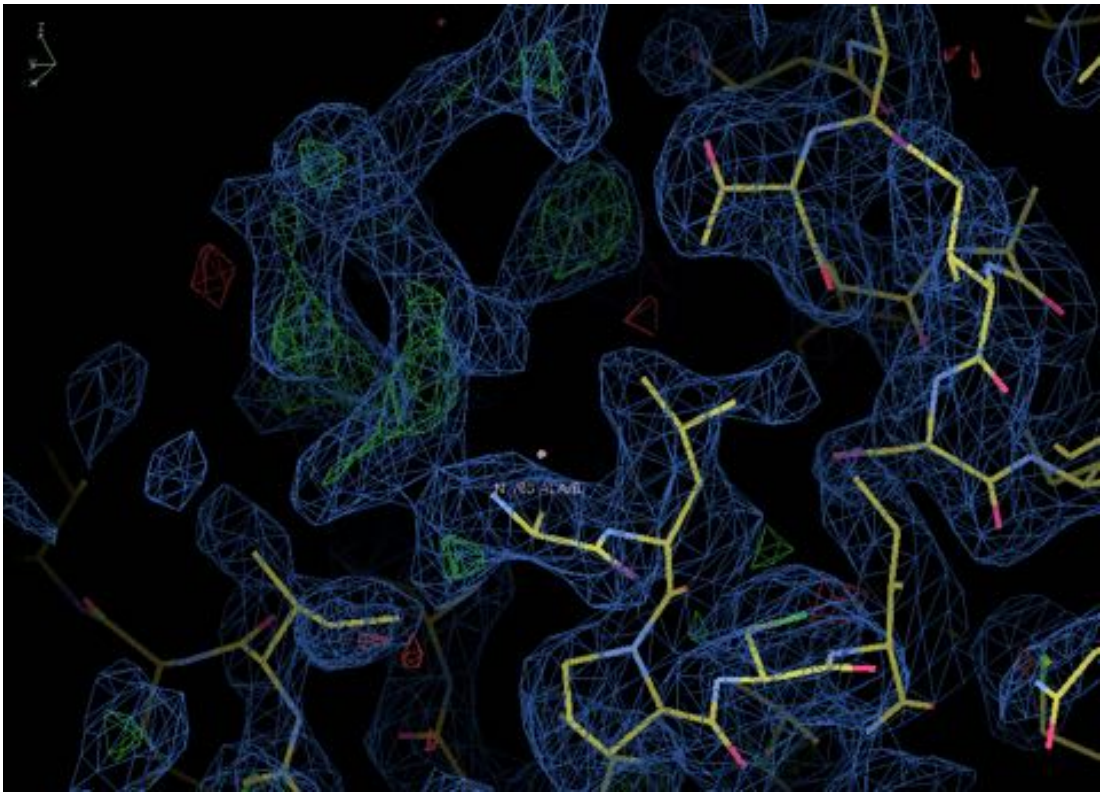


Figure 8.2.2 – The Initial electron density map produced post molecular replacement, displays difference density features not accounted for in the search ensemble. The diagram above shows the eIF4a NTD search ensemble (yellow) truncated at ALA 63, overlaid with the initial electron density map calculated directly after molecular replacement contoured to 1 σ. The map at this point has not been refined. However, there is substantial difference density (green) accounting for the N termini residues truncated from the search model, with ample evidence to assign sequence register. Thus validating the orientation of the two domains of eIF4a within the unit cell.

and CTD domains of eIF4a were deleted and the remaining model was refined first using 5 cycles of rigid body refinement in Phenix refine (Adams *et al.,* 2010) and then 10 cycles of maximum likelihood refinement in Refmac (Murshudov *et al.,* 1997), to an R_{work} of 0.3 and R_{free} of 0.35. The resultant electron density map was then used to model the remaining residues, with the final model incorporating 585 of the 597 expected positions. All refinement past the initial structure solution stage was conducted using Phenix refine (Adams *et al.,* 2010) in an iterative fashion up till an R_{work} of 0.24 and R_{free} of 0.29. The refinement and validation statistics are detailed in table 8.3.1.

8.3.2 – BLF1 interfaces with both the N and C terminal domains of eIF4a

The global structure of the BLF1 C94S – eIF4a complex (figure 8.3.1), shows that eIF4a is held in a closed conformation with BLF1 sitting directly in between the N and C terminal domains of eIF4a. However, the relative orientation of the N and C terminal domains, relative to one another, when bound to BLF1 is significantly altered from either of the previously observed conformations of eIF4a (figure 8.3.2): closed (substrate bound) or open (*apo*). The difference observed between the classic open and closed modes of eIF4a are relatively subtle, with a closure between domains of approximately 20 ^{o} about a central hinge region. Given that molecular replacement was not successful with the full-length eIF4a, in either their open or closed state, substantial deviation from previously reported structures was expected. However, when bound to BLF1 C94S the CTD domain is rotated approximately 70 ^{o} counter-clockwise in relation to the fixed superimposed NTD, with a closure between the domains more closely resembling the open form. Curiously, despite significant modifications having been made to eIF4a, there does not appear to be any conformational changes observed in BLF1 upon complex formation.

8.3.3 BLF1 C94S binds to Glutamine 339 in human eIF4a

Examination of the active site of BLF1 (figure 8.3.3) reveals that the toxin binds to the sidechain hydroxyl of GLN 339 through hydrogen bonds with the backbone amide moieties of the conserved SER and GLY residues present in the central LSGC motif, forming an oxyanion hole. This oxyanion hole is responsible for holding the substrate sidechain in place, adjacent to the nucleophilic CYS located at position 94 (in the WT enzyme). There also appears to be an unexpected electrostatic interaction with TRP 66 (BLF1), which lays directly above GLN 339 (eIF4a) in the active site and is conserved between BLF1 and HCH_03101.

Model	HCH_03101 CTD WT
Resolution (Å)	63.07 – 2.52
Total unique observations	20,378
Protein molecules per asymmetric unit	1
Number of atoms	4735
Number of waters	97
Number of residues / modeled residues	597/585
Truncated residues (Cα)	N/A
Ramachandran favoured (%)	96 %
Ramachandran outliers (%)	0.87 %
Poor rotamers (%)	0.4 %
RMSD bond length (Å)	0.003
RMSD angle (°)	0.69
Average B-factors (Å^2)	
Protein (Å^2)	46.0
Waters (Å^2)	44.3
R_{work}	0.24
R_{free}	0.29
MolProbity score	1.66 (99 %)
MolProbity Clash score	6.49 (98 %)

Table 8.3.1 – Refinement and model validation statistics for the BLF1 C94S – eIF4a complex. The finalised model of BLF1 C94S in complex with WT eIF4a accounts for 585 / 597 (98 %) of the expected residues present across both protein components, with all the positions unaccounted for located in a single loop in the NTD domain of eIF4a. The model was refined in an iterative fashion using Phenix refine (Adams *et al.,* 2010) to an R_{work} of 0.24 and R_{free} of 0.29. The completed model was then validated using Molprobity (Chen *et al.,* 2010), with Molprobity and Clash scores placing the model within the top 98 % of structures deposited in the PDB within a 2.5 Å (± 0.25 Å) resolution range. The only outlying statistics are the 0.87 % Ramachandran outliers, which account for 5 residues (GLY 51, ASP 307, LEU 332, ALA 333 and ILE 336) all clustered in the same region close to where BLF1 binds its substrate glutamine.

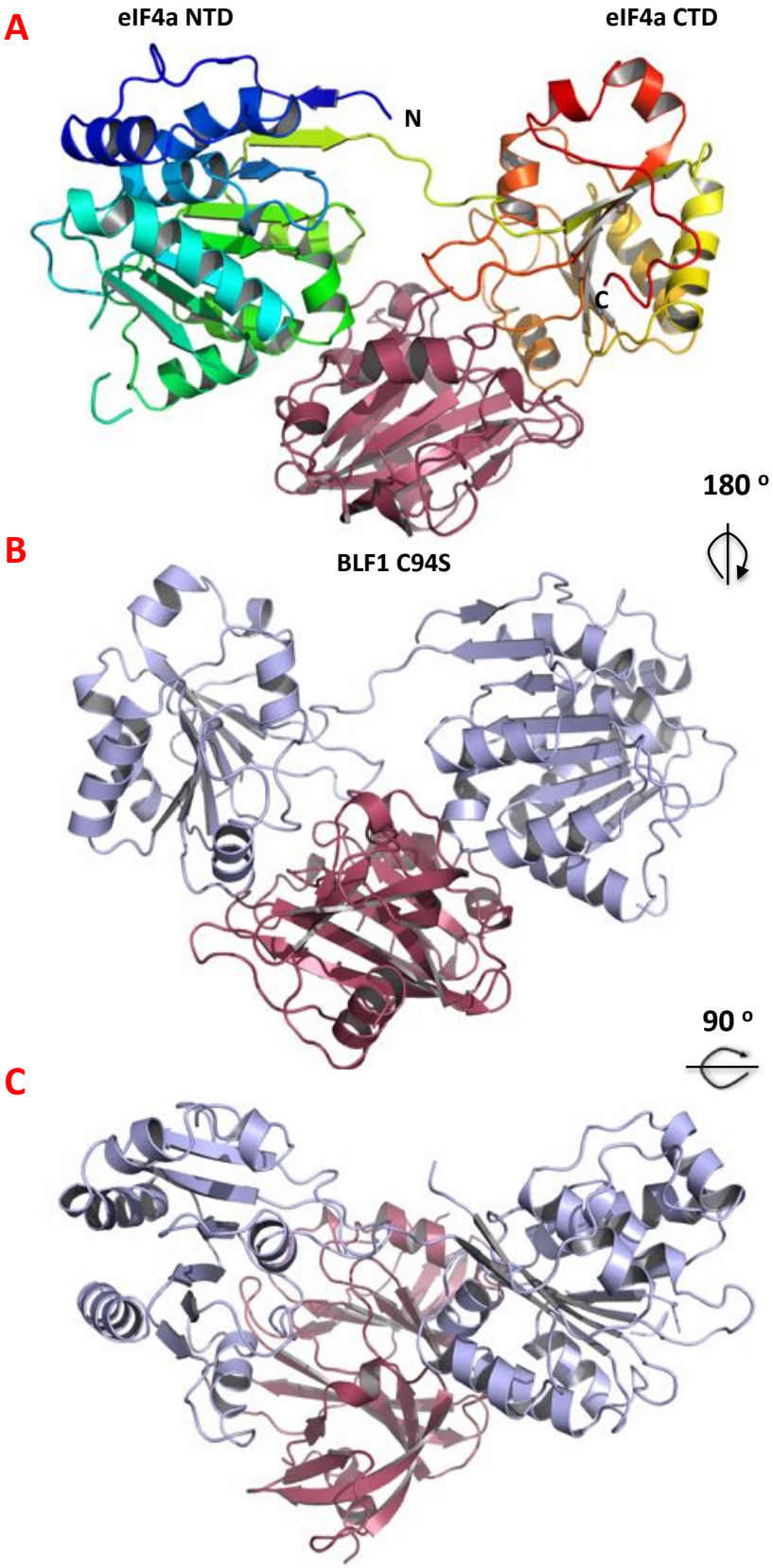


Figure 8.3.1 – Gross structure of the BLF1 C94S – eIF4a complex. Panel **A** shows the complex from a side face, with eIF4a in rainbow colours tracking progression from the N terminus (blue) to the C terminus (red). eIF4a is a two domain protein, each composed of a single β-sheet flanked by α-helices. Panels **B** and **C** show the relative position of BLF1 C94S (red) in relation to eIF4a (blue) rotated 180 ^{o} and looking down from the top face respectively.

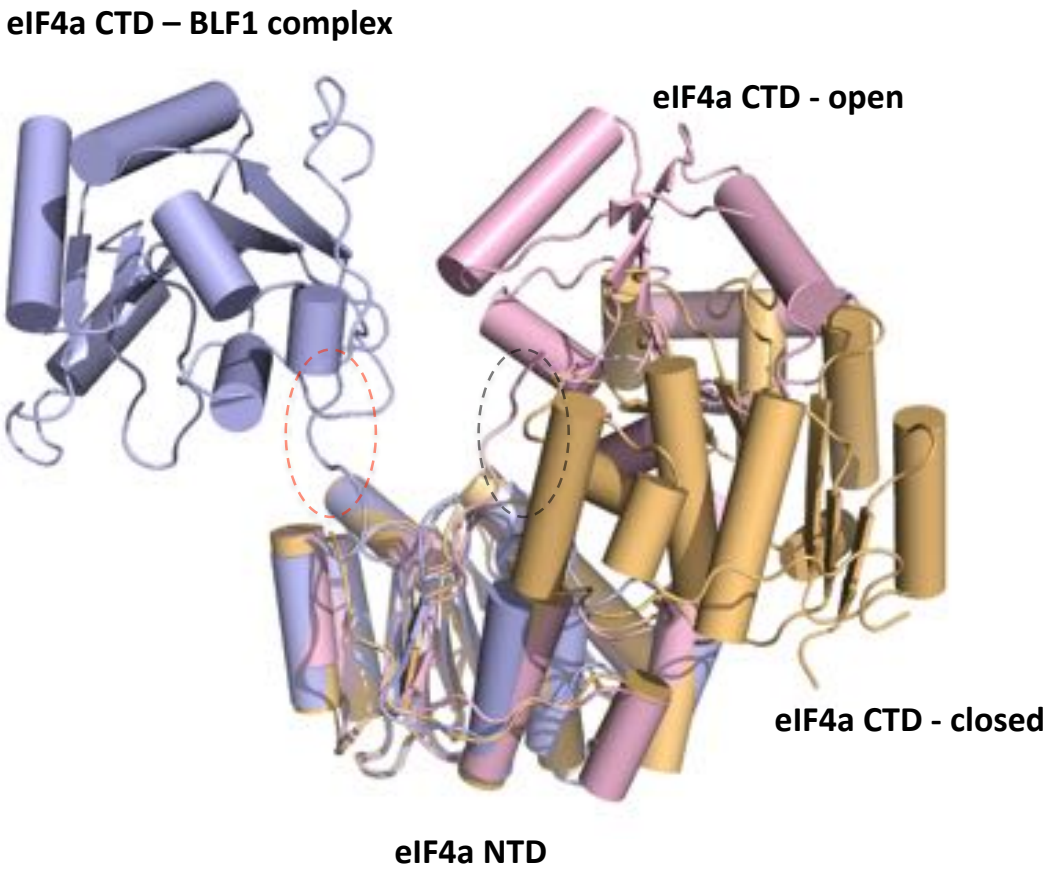


Figure 8.3.2 – Interaction with BLF1 induces significant domain reorientation in eIF4a. The above diagram superimposes three structurally characterised orientations of eIF4a, superimposed on their NTD domains. The *apo* form (pink) adopts an open conformation and the closed conformation (orange) is bound to Tumour suppressor programmed cell death protein 4 (PDCD4), which is not shown. These two conformations display a closure of the gap between the NTD and CTD domains of eIF4a, about a central hinge (black dashed circle). Conversely, upon complex formation with BLF1 C94S (blue) eIF4a adopts a more drastic conformational change; with the hinge region (red dashed circle) now rotated by approximately 70 ^{o} counter clockwise, whilst held in a similar conformation to the open form.

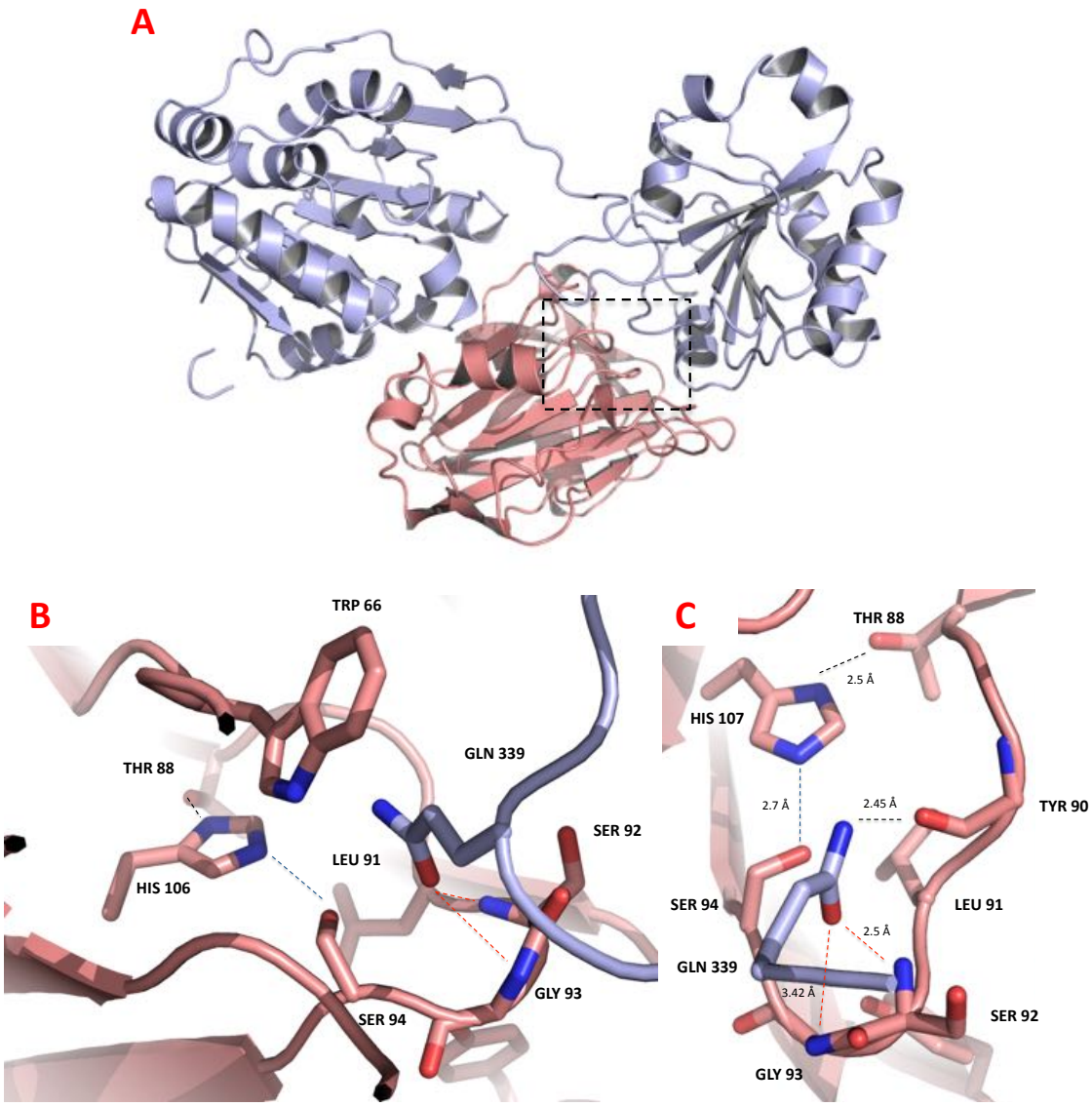


Figure 8.3.3 – BLF1 co-ordinates its substrate GLN 339 through formation of an oxyanion hole, prior to nucleophlic attack by its essential CYS 94 residue. Panel **A** shows the gross structure of the BLF1 C94S – eIF4a complex, with a black dashed box indicating the location of the active site in BLF1. Panels **B** and **C** zoom into the active site of BLF1, where GLN 339 the site-specific substrate of BLF1 is present. Within the active site, BLF1 interacts with eIF4a via hydrogen bonds between the side-chain hydroxyl moeity and the backbone amide moieties of SER 92 and GLY 93 (red), forming an oxyanion hole. This interaction in turn holds GLN 339 in place close to the catalytic nucleophile CYS 94 (mutated to SER), which in turn is interacting with HIS 107 (blue), to form a dyad. Additionally the substrate is held in place by electrostatic interactions with TRP 66, which is likely involved in orienting the terminal amine of GLN 339 within reach of the catalytic CYS nucleophile.

MS-MS mass spectroscopy analysis of the complex formed prior to crystallisation and co-purified as described in section 8.1 was sampled from an SDS-PAGE gel, indicating that the C94S mutant of BLF1 is inactive, with no de-amidated peptides detected. Therefore, the residue modelled at position 339 in the BLF1 C94S - eIF4a structure remains an unmodified glutamine.

8.3.4 The co-ordinantion of GLN 339 is reminiscent of the Papain

Papain has previously been used as a model for the mechanism of glutamine de-amidation because it causes the cleavage of peptide bonds, which both involve the breakage of C-N scissile bonds. This new structure reveals that the catalytic similarity between these two protein families is extensive, despite sharing no structural resemblance, with both forming a characteristic oxyanion hole interaction with their substrates (previously shown in figure 1.5.2). When combined with their shared CYS – HIS catalytic dyads, it is likely that the Glutamine de-amidase enzymes will follow a similar catalytic mechanism modelled extensively in papain (section 1.5). First forming a tetrahedral acyl-enzyme intermediate, with dissociation of the enzyme driven by hydrolysis with the release of a NH_3^+ leaving group, as presented in figure 1.5.3.

This is interesting for two reasons; first until recently the catalytic mechanism of the glutamine de-amidase enzymes was poorly understood. Second, the cysteine proteases particularly papain, are extremely well characterised with several inhibitors available to block their activity. If any of the current protease inhibitors were capable of inhibiting BLF1, or could be used as springboards towards the design of novel synthetic compounds against the toxin, this information could become extremely valuable in developing front line drugs for the treatment of Melioidosis. Furthermore, information regarding how substrate glutamine residues are coordinated within the active site, allows for the first significant models of CNF1 interacting with its small GTPase substrates RhoA, Rac and cdc42.

8.4 – Utilising the BLF1 C94S – eIF4a complex structure to characterise HCH_03101

8.4.1 – Docking HCH_03101 in the place of BLF1 reveals a possible role for the β-protrusion

In chapter 7 it was shown that HCH_03101 is capable of de-amidating the same glutamine position as BLF1 in human eIF4a. Given the strong structural conservation observed between

BLF1 and HCH_03101 at their active sites it is possible to superimpose the *H. chejuensis* toxin in the place of BLF1 about this region (figure 8.4.1).

At the present it has not been possible to produce a HCH_03101 C94S - eIF4a co-crystal. Therefore, modelling the likely interactions between eIF4a and HCH_03101 is the best avenue towards exploring the role of this proteins unusually long β-protrusion. Conservative docking, with no modifications made to the HCH_03101 model (detailed in chapter 6) reveals that the β-protrusion, which extends 44 Å out from the globular body of the protein, supports the CTD domain of eIF4a. Investigating the surface charge of eIF4a on the face interacting with the β-protrusion (figure 8.4.2A-B) reveals that the entire region is negatively charged. Conversely, the β-protrusion is predominantly positively charged, indicating that the interaction seen in the docked model is likely of physiological relevance. Furthermore, the β-protrusion exhibits a curious 2-fold pseudo-symmetry, with two prominent ARG residues facing in the same direction on the top (eIF4a interface) surface (figure 6.4.1). There are several negatively charged side-chains within prospective hydrogen bonding distance of the modelled ARG positions (figure 8.4.2C-D); most prominently ARG 172 could interact with either ASP 261 or 265, whereas ARG 181 is potentially interacting with ASP 265, GLU 268 or THR 269.

8.4.2 – HCH_03101 shares no sequence conservation with BLF1 in the eIF4a interface region

Surprisingly the conserved residues between BLF1 and HCH_03101 (figure 8.4.3) are not clustered around the binding interface with eIF4a but instead are buried deep with in the core of the toxin. This would suggest that the presumably aquatic species that *H. chejuensis* infects, produces a homologue of eIF4a that is significantly different to the human example studied.

8.5 – Conclusions and future work

Structural solution of BLF1 in complex with eIF4a represents a significant landmark in the study of glutamine de-amidase enzymes, as it is the first time any toxin from this small but significant super-family has been structurally examined with its native substrate. The presented structure not only validates the functional characterisation previously undertaken (Cruz-Migoni *et al.,* 2011) but also provides substantial evidence towards understanding the catalytic mechanism

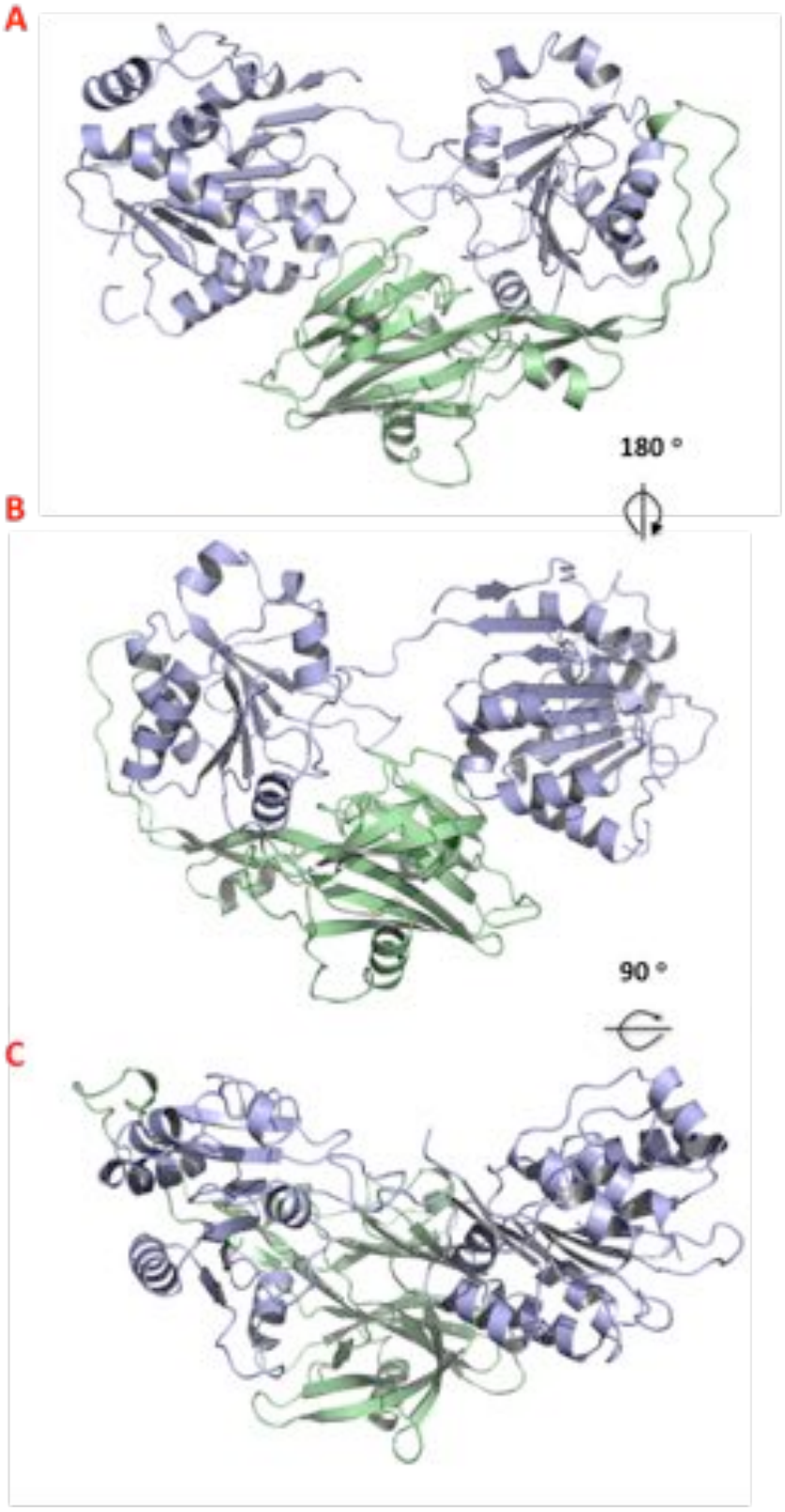


Figure 8.4.1 – When HCH_03101 is docked in the place of BLF1 C94S the β-protrusion cradles the CTD domain of eIF4a. Panel **A** shows HCH_03101 docked in the same position as BLF1, based on the superposition of conserved catalytic residues. Panels **B** and **C** highlight the likely role of the long β-protrusion of HCH_03101, which cradles the CTD of eIF4a with large portions of the protrusion within hydrogen bonding distance of eIF4a.

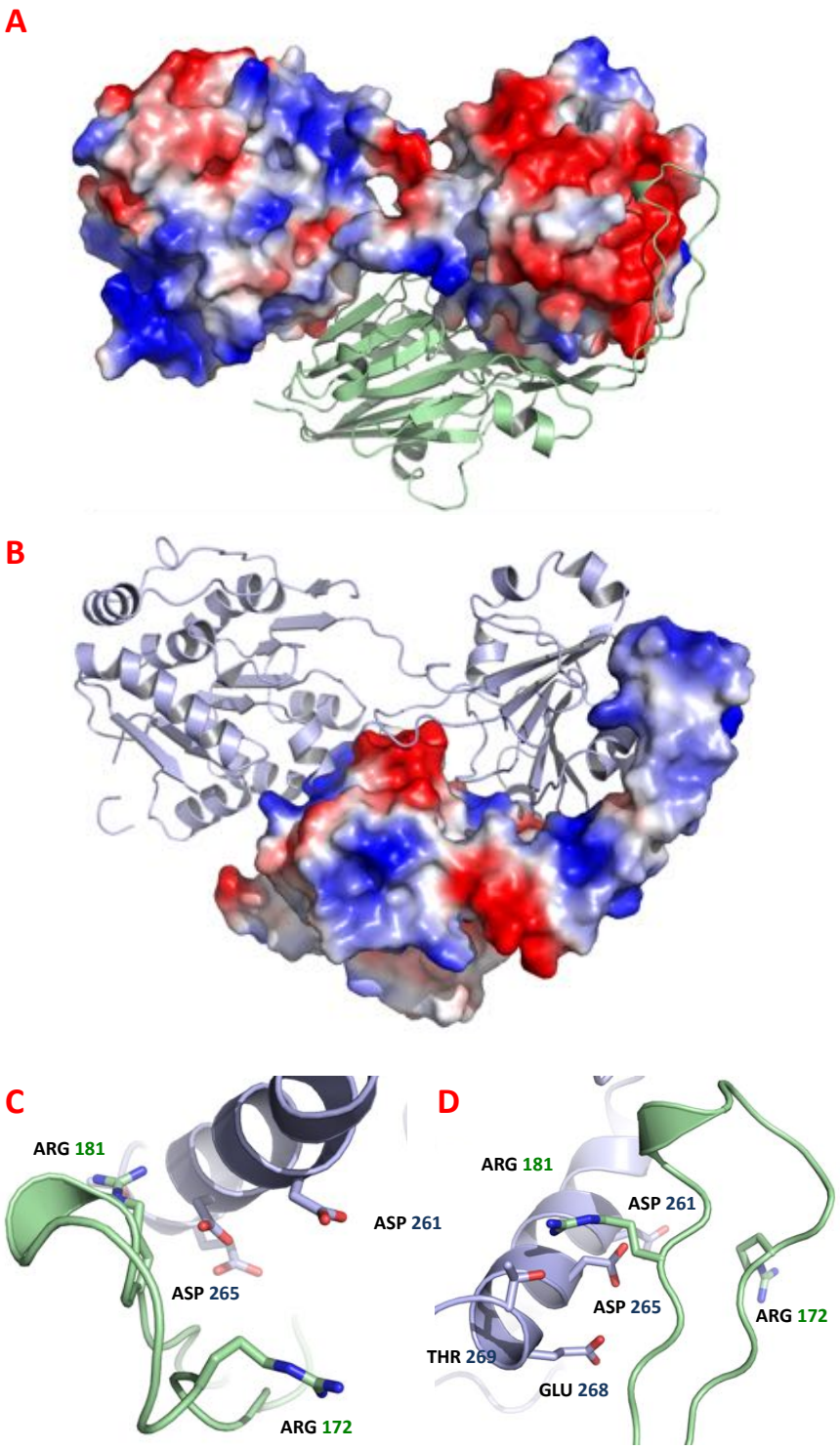


Figure 8.4.2 – The modelled orientation of the β-protrusion lays close to a large region of negatively charged residues in eIF4a. Panels **A** and **B** are mesh overlays showing the electrostatic surface charges of eIF4a and HCH_03101 respectively, with positive charges shown in red and negative charges in blue. The outermost face of the CTD domain of eIF4a, which is cradled by the β-protrusion is almost entirely negatively charged, in stark contrast to the predominantly positively charged toxin protuberance. Panels **C** and **D** highlight the possible eIF4a side chains responsible for specific interactions with prominent ARG residues on the β-protrusion.

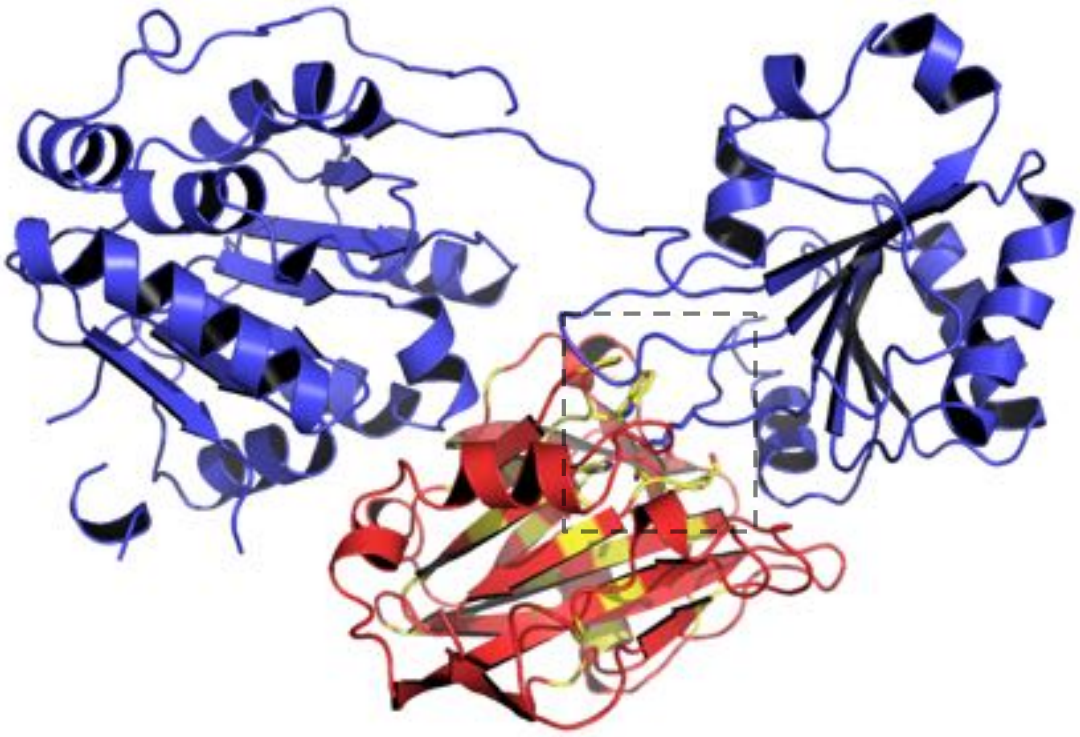


Figure 8.4.3 – BLF1 and HCH_03101 share little sequence conservation in the eIF4a interface region. The above diagram shows eIF4a (blue) in complex with BLF1 (red) with the sequence conservation shared with HCH_03101 highlighted on the toxin in yellow. Outside of the conserved catalytic residues (black dashed box) there is no sequence conservation observed in the flanking eIF4a interface regions. Indeed, the majority of the conservation is located in the central β-sandwich region, which contrasts with the total lack of conservation identified in this region between BLF1 and C-CNF1. Consequently, it is probable that the interaction between these toxins and their substrates is either mediated by multiple less specific electrostatic interactions, or that HCH_03101 targets a divergent homologue of eIF4a.

of these enzymes. The mechanistic link to the papain like family of cysteine proteases permits further probing of their catalytic mechanism, including several well-established assays. The structure is also built to the highest resolution data recorded (PDB) for the full length human eIF4a.

The methodologies described above could also be utilised to examine C-CNF1 in complex with its small GTPase substrates RhoA, Rac and cdc42. This is particularly plausible, as C-CNF1 has been shown to form a stable complex with recombinant RhoA by pull-down assays (section 7.2). With structural information regarding the binding of GLN 339 into the active site of BLF1 available, it is also possible to design a range of short oligo-peptides to probe the minimal binding site required for substrate recognition. Finally, whilst attempts to co-crystallise HCH_03101 in the presence of human eIF4a have proved unsuccessful thus far, the docking reveals that these two proteins are not inherently unsuitable for structural studies. Furthermore, whilst the target organism of *H. chejuensis* remains unknown this model system offers the only readily available alternative for study.

- Adams, P. D., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D-Biological Crystallography 66 213-221.
- Aldhous, P. (2005). Melioidosis? Never heard of it. Nature 434(7034) 692-693.
- Altschul, S. F., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17) 3389-3402.
- Amamoto, T., et al. (1984). EFFECT OF E-64, THIOL PROTEASE INHIBITOR, ON THE SECONDARY ANTI-SRBC RESPONSE INVITRO. Microbiology and Immunology 28(1) 85-97.
- Arad, D., R. Langridge and P. A. Kollman (1990). A SIMULATION OF THE SULFUR ATTACK IN THE CATALYTIC PATHWAY OF PAPAIN USING MOLECULAR MECHANICS AND SEMIEMPIRICAL QUANTUM-MECHANICS. Journal of the American Chemical Society 112(2) 491-502.
- Baker, D. and A. Sali (2001). Protein structure prediction and structural genomics. Science 294(5540) 93-96.
- Baldauf, S. L. and W. F. Doolittle (1997). Origin and evolution of the slime molds (Mycetozoa). Proceedings of the National Academy of Sciences of the United States of America 94(22) 12007-12012.
- Bartlett, J. M. S. and D. Stirling (2003). A short history of the polymerase chain reaction. Methods in molecular biology (Clifton, N.J.) 226 3-6.
- Baumann, P., et al. (1980). RE-EVALUATION OF THE TAXONOMY OF VIBRIO, BENECKEA, AND PHOTOBACTERIUM - ABOLITION OF THE GENUS BENECKEA. Current Microbiology 4(3) 127-132.
- Benner, S. A. and D. Gerloff (1991). PATTERNS OF DIVERGENCE IN HOMOLOGOUS PROTEINS AS INDICATORS OF SECONDARY AND TERTIARY STRUCTURE - A PREDICTION OF THE STRUCTURE OF THE CATALYTIC DOMAIN OF PROTEIN-KINASES. Advances in Enzyme Regulation, Vol 31 31 121-181.
- Blum, G., et al. (1995). GENE CLUSTERS ENCODING THE CYTOTOXIC NECROTIZING FACTOR TYPE-1, PRS-FIMBRIAE AND ALPHA-HEMOLYSIN FORM THE PATHOGENICITY ISLAND-II OF THE UROPATHOGENIC ESCHERICHIA-COLI STRAIN-J96. Fems Microbiology Letters 126(2) 189-195.
- Boquet, P. (2001). The cytotoxic necrotizing factor 1 (CNF1) from Escherichia coli. Toxicon 39(11) 1673-1680.
- Bragg, W. H. and W. L. Bragg (1913). The reflection of X-rays by crystals. Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character 88(605) 428-438.

- Bragg, W. L. (1913). The structure of some crystals as indicated by their diffraction of x-rays. Proceedings of the Royal Society of London Series a-Containing Papers of a Mathematical and Physical Character 89(610) 248-277.
- Brocklehurst, K., et al. (1988). CONSEQUENCES OF MOLECULAR RECOGNITION IN THE S1-S2 INTERSUBSITE REGION OF PAPAIN FOR CATALYTIC-SITE CHEMISTRY - CHANGE IN PH-DEPENDENCE CHARACTERISTICS AND GENERATION OF AN INVERSE SOLVENT KINETIC ISOTOPE EFFECT BY INTRODUCTION OF A P1-P2 AMIDE BOND INTO A 2-PROTONIC-STATE REACTIVITY PROBEI. Biochemical Journal 250(3) 761-772.
- Brunger, A. T. (1993). ASSESSMENT OF PHASE ACCURACY BY CROSS VALIDATION - THE FREE R-VALUE - METHODS AND APPLICATIONS. Acta Crystallographica Section D-Biological Crystallography 49 24-36.
- Buetow, L., et al. (2001). Structure of the Rho-activating domain of Escherichia coli cytotoxic necrotizing factor 1. Nature Structural Biology 8(7) 584-588.
- Buetow, L. and P. Ghosh (2003). Structural elements required for deamidation of RhoA by cytotoxic necrotizing factor 1. Biochemistry 42(44) 12784-12791.
- Caprioli, A., et al. (1983). PARTIAL-PURIFICATION AND CHARACTERIZATION OF AN ESCHERICHIA-COLI TOXIC FACTOR THAT INDUCES MORPHOLOGICAL CELL ALTERATIONS. Infection and Immunity 39(3) 1300-1306.
- Chao, X. J., et al. (2006). A receptor-modifying deamidase in complex with a signaling phosphatase reveals reciprocal regulation. Cell 124(3) 561-571.
- Chen, V. B., et al. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallographica Section D-Biological Crystallography 66 12-21.
- Choe, S., et al. (1992). THE CRYSTAL-STRUCTURE OF DIPHTHERIA-TOXIN. Nature 357(6375) 216-222.
- Combet, C., et al. (2000). NPS@: Network Protein Sequence Analysis. Trends in Biochemical Sciences 25(3) 147-150.
- Crevel, G., et al. (2001). The Drosophila Dpit47 protein is a nuclear Hsp90 co-chaperone that interacts with DNA polymerase alpha. Journal of Cell Science 114(11) 2015-2025.
- Cruz-Migoni, A., et al. (2011). A Burkholderia pseudomallei Toxin Inhibits Helicase Activity of Translation Factor eIF4A. Science 334(6057) 821-824.
- Cui, J. X., et al. (2010). Glutamine Deamidation and Dysfunction of Ubiquitin/NEDD8 Induced by a Bacterial Effector Family. Science 329(5996) 1215-1218.
- Currie, B. J. (2008). Advances and remaining uncertainties in the epidemiology of Burkholderia pseudomallei and melioidosis. Transactions of the Royal Society of Tropical Medicine and Hygiene 102(3) 225-227.

- D'Amato, G. and R. Patarca (1998). General biological aspects of oncogenesis. Critical Reviews in Oncogenesis 9(3-4) 275-373.
- Daggett, V., S. Schroder and P. Kollman (1991). CATALYTIC PATHWAY OF SERINE PROTEASES - CLASSICAL AND QUANTUM-MECHANICAL CALCULATIONS. Journal of the American Chemical Society 113(23) 8926-8935.
- Dance, D. A. B. (2000). Ecology of Burkholderia pseudomallei and the interactions between environmental Burkholderia spp. and human-animal hosts. Acta Tropica 74(2-3) 159-168.
- Doye, A., et al. (2002). CNF1 exploits the ubiquitin-proteasome machinery to restrict Rho GTPase activation for bacterial host cell invasion. Cell 111(4) 553-564.
- Drenth, J., et al. (1968). STRUCTURE OF PAPAIN. Nature 218(5145) 929-&.
- Eckert, M. (2012). Max von Laue and the discovery of X-ray diffraction in 1912. Annalen Der Physik 524(5) A83-A85.
- Emsley, P. and K. Cowtan (2004). Coot: model-building tools for molecular graphics. Acta Crystallographica Section D-Biological Crystallography 60 2126-2132.
- Eyal, E., et al. (2007). Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. Bioinformatics 23(13) I175-I184.
- Falbo, V., et al. (1993). ISOLATION AND NUCLEOTIDE-SEQUENCE OF THE GENE ENCODING CYTOTOXIC NECROTIZING FACTOR-I OF ESCHERICHIA-COLI. Infection and Immunity 61(11) 4909-4914.
- Falke, J. J. and G. L. Hazelbauer (2001). Transmembrane signaling in bacterial chemoreceptors. Trends in Biochemical Sciences 26(4) 257-265.
- Fenn, J. B., et al. (1989). ELECTROSPRAY IONIZATION FOR MASS-SPECTROMETRY OF LARGE BIOMOLECULES. Science 246(4926) 64-71.
- Fiorentini, C., et al. (1988). CYTOSKELETAL CHANGES INDUCED IN HEP-2 CELLS BY THE CYTO-TOXIC NECROTIZING FACTOR OF ESCHERICHIA-COLI. Toxicon 26(11) 1047-1056.
- Fiorentini, C., et al. (1997). Escherichia coli cytotoxic necrotizing factor 1 (CNF1), a toxin that activates the Rho GTPase. Journal of Biological Chemistry 272(31) 19532-19537.
- Flatau, G., et al. (2000). Deamidation of RhoA glutamine 63 by the Escherichia coli CNF1 toxin requires a short sequence of the GTPase switch 2 domain. Biochemical and Biophysical Research Communications 267(2) 588-592.
- Flatau, G., et al. (1997). Toxin-induced activation of the G protein p21 Rho by deamidation of glutamine. Nature 387(6634) 729-733.

- Foynes, S., et al. (2000). Helicobacter pylori possesses two CheY response regulators and a histidine kinase sensor, CheA, which are essential for chemotaxis and colonization of the gastric mucosa. Infection and Immunity 68(4) 2016-2023.
- Galan, J. E. and D. Zhou (2000). Striking a balance: Modulation of the actin cytoskeleton by Salmonella. Proceedings of the National Academy of Sciences of the United States of America 97(16) 8754-8761.
- Garavito, R. M., et al. (1977). CONVERGENCE OF ACTIVE-CENTER GEOMETRIES. Biochemistry 16(23) 5065-5071.
- Garrity, L. F. and G. W. Ordal (1997). Activation of the CheA kinase by asparagine in Bacillus subtilis chemotaxis. Microbiology-Uk 143 2945-2951.
- Gasteiger, E., et al. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research 31(13) 3784-3788.
- Gopalan, P., M. J. Dufresne and A. H. Warner (1986). EVIDENCE FOR A DEFECTIVE THIOL PROTEASE INHIBITOR IN SKELETAL-MUSCLE OF MICE WITH HEREDITARY MUSCULAR-DYSTROPHY. Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire 64(10) 1010-1019.
- Gorret, J., et al. (2009). Childhood delayed septic arthritis of the knee caused by Serratia fonticola. The Knee 16(6) 512-514.
- Green, D. W., V. M. Ingram and M. F. Perutz (1954). THE STRUCTURE OF HAEMOGLOBIN .4. SIGN DETERMINATION BY THE ISOMORPHOUS REPLACEMENT METHOD. Proceedings of the Royal Society of London Series a-Mathematical and Physical Sciences 225(1162) 287-307.
- Grimont, P. A. D. and F. Grimont (1978). GENUS SERRATIA. Annual Review of Microbiology 32 221-248.
- Harrison, M. J., N. A. Burton and I. H. Hillier (1997). Catalytic mechanism of the enzyme papain: Predictions with a hybrid quantum mechanical molecular mechanical potential. Journal of the American Chemical Society 119(50) 12285-12291.
- Hautbergue, G. M. and S. A. Wilson (2012). BLF1, the first Burkholderia pseudomallei toxin, connects inhibition of host protein synthesis with melioidosis. Biochemical Society Transactions 40 842-845.
- Holden, M. T. G., et al. (2004). Genornic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proceedings of the National Academy of Sciences of the United States of America 101(39) 14240-14245.
- Holm, L., et al. (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics 24(23) 2780-2781.

- Holm, L. and P. Rosenstrom (2010). Dali server: conservation mapping in 3D. Nucleic Acids Research 38 W545-W549.
- Howe, C., A. Sampath and M. Spotnitz (1971). PSEUDOMALLEI GROUP - REVIEW. Journal of Infectious Diseases 124(6) 598-&.
- Jensen, S., et al. (2003). Characterization of strains of Vibrio splendidus and V-tapetis isolated from corkwing wrasse Symphodus melops suffering vibriosis. Diseases of Aquatic Organisms 53(1) 25-31.
- Jeong, H., et al. (2005). Genomic blueprint of Hahella chejuensis, a marine microbe producing an algicidal agent. Nucleic Acids Research 33(22) 7066-7073.
- Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292(2) 195-202.
- Jones, D. T. and M. B. Swindells (2002). Getting the most from PSI-BLAST. Trends in Biochemical Sciences 27(3) 161-164.
- Kabsch, W. (2010). XDS. Acta Crystallographica Section D-Biological Crystallography 66 125-132.
- Kamphuis, I. G., et al. (1984). STRUCTURE OF PAPAIN REFINED AT 1.65 A RESOLUTION. Journal of Molecular Biology 179(2) 233-256.
- Kanaphun, P., et al. (1993). SEROLOGY AND CARRIAGE OF PSEUDOMONAS-PSEUDOMALLEI - A PROSPECTIVE-STUDY IN 1000 HOSPITALIZED CHILDREN IN NORTHEAST THAILAND. Journal of Infectious Diseases 167(1) 230-233.
- Kehry, M. R., T. G. Doak and F. W. Dahlquist (1985). ABERRANT REGULATION OF METHYLESTERASE ACTIVITY IN CHED CHEMOTAXIS MUTANTS OF ESCHERICHIA-COLI. Journal of Bacteriology 161(1) 105-112.
- Kelley, L. A. and M. J. E. Sternberg (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nature Protocols 4(3) 363-371.
- Kendrew, J. C., et al. (1958). 3-DIMENSIONAL MODEL OF THE MYOGLOBIN MOLECULE OBTAINED BY X-RAY ANALYSIS. Nature 181(4610) 662-666.
- Kirby, J. R., et al. (2001). CheC is related to the family of flagellar switch proteins and acts independently from CheD to control chemotaxis in Bacillus subtilis. Molecular Microbiology 42(3) 573-585.
- Kleywegt, G. J. and T. A. Jones (2002). Homo crystallographicus - Quo Vadis? Structure 10(4) 465-472.
- Kwon, S. K., Y. K. Park and J. F. Kim (2010). Genome-Wide Screening and Identification of Factors Affecting the Biosynthesis of Prodigiosin by Hahella chejuensis, Using Escherichia coli as a Surrogate Host. Applied and Environmental Microbiology 76(5) 1661-1668.

- Lah, T. T., et al. (1989). INHIBITORY PROPERTIES OF LOW-MOLECULAR MASS CYSTEINE PROTEINASE-INHIBITORS FROM HUMAN SARCOMA. Biochimica Et Biophysica Acta 993(1) 63-73.
- Lalucat, J., et al. (2006). Biology of Pseudomonas stutzeri. Microbiology and Molecular Biology Reviews 70(2) 510-+.
- Landraud, L., et al. (2004). E-coli CNF1 toxin: a two-in-one system for host-cell invasion. International Journal of Medical Microbiology 293(7-8) 513-518.
- Lee, H. K., et al. (2001). Hahella chejuensis gen. nov., sp nov., an extracellular-polysaccharide-producing marine bacterium. International Journal of Systematic and Evolutionary Microbiology 51 661-666.
- Lee, Y. H., et al. (2010). Identification of tomato plant as a novel host model for Burkholderia pseudomallei. Bmc Microbiology 10 11.
- Lefebre, M. D. and M. A. Valvano (2001). In vitro resistance of Burkholderia cepacia complex isolates to reactive oxygen species in relation to catalase and superoxide dismutase production. Microbiology-Uk 147 97-109.
- Lemichez, E., et al. (1997). Molecular localization of the Escherichia coli cytotoxic necrotizing factor CNF1 cell-binding and catalytic domains. Molecular Microbiology 24(5) 1061-1070.
- Lemonnier, M., L. Landraud and E. Lemichez (2007). Rho GTPase-activating bacterial toxins: from bacterial virulence regulation to eukaryotic cell biology. Fems Microbiology Reviews 31(5) 515-534.
- Lockman, H. A., et al. (2002). Yersinia pseudotuberculosis produces a cytotoxic necrotizing factor. Infection and Immunity 70(5) 2708-2714.
- Malina, A., J. R. Mills and J. Pelletier (2012). Emerging Therapeutics Targeting mRNA Translation. Cold Spring Harbor Perspectives in Biology 4(4) 17.
- Malorni, W. and C. Fiorentini (2006). Is the Rac GTPase-activating toxin CNF1 a smart hijacker of host cell fate? Faseb Journal 20(6) 606-609.
- Masuda, M., et al. (2000). Activation of Rho through a cross-link with polyamines catalyzed by Bordetella dermonecrotizing toxin. Embo Journal 19(4) 521-530.
- Matthaei, J. H. and M. W. Nirenberg (1962). CHARACTERISTICS OF RNA CODING UNITS. Federation Proceedings 21(2) 415-&.
- McCoy, A. J., et al. (2007). Phaser crystallographic software. Journal of Applied Crystallography 40 658-674.
- Menard, R., et al. (1991). CONTRIBUTION OF THE GLUTAMINE-19 SIDE-CHAIN TO TRANSITION-STATE STABILIZATION IN THE OXYANION HOLE OF PAPAIN. Biochemistry 30(37) 8924-8928.

- Mima, T. and H. P. Schweizer (2010). The BpeAB-OprB Efflux Pump of Burkholderia pseudomallei 1026b Does Not Play a Role in Quorum Sensing, Virulence Factor Production, or Extrusion of Aminoglycosides but Is a Broad-Spectrum Drug Efflux System. Antimicrobial Agents and Chemotherapy 54(8) 3113-3120.
- Muff, T. J. and G. W. Ordal (2007). The CheC phosphatase regulates chemotactic adaptation through CheD. Journal of Biological Chemistry 282(47) 34120-34128.
- Murshudov, G. N., A. A. Vagin and E. J. Dodson (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallographica Section D-Biological Crystallography 53 240-255.
- Nandi, T., et al. (2010). A Genomic Survey of Positive Selection in Burkholderia pseudomallei Provides Insights into the Evolution of Accidental Virulence. Plos Pathogens 6(4) 15.
- Nave, C. and E. F. Garman (2005). Towards an understanding of radiation damage in cryocooled macromolecular crystals. Journal of Synchrotron Radiation 12 257-260.
- Ngauy, V., et al. (2005). Cutaneous melioidosis in a man who was taken as a prisoner of war by the Japanese during World War II. Journal of Clinical Microbiology 43(2) 970-972.
- Ordal, G. W., L. Marquez-Magana and M. J. Chamberlin (1993). Motility and chemotaxis. Bacillus subtilis and other gram-positive bacteria: Biochemistry, physiology and molecular genetics 765-784.
- Orden, J. A., et al. (2007). Necrotoxigenic Escherichia coli from sheep and goats produce a new type of cytotoxic necrotizing factor (CNF3) associated with the eae and ehxA genes. International Microbiology 10(1) 47-55.
- Oswald, E., et al. (1994). CYTOTOXIC NECROTIZING FACTOR TYPE-2 PRODUCED BY VIRULENT ESCHERICHIA-COLI MODIFIES THE SMALL GTP-BINDING PROTEINS-RHO INVOLVED IN ASSEMBLY OF ACTIN STRESS FIBERS. Proceedings of the National Academy of Sciences of the United States of America 91(9) 3814-3818.
- Park, S. Y., et al. (2004). Structure and function of an unusual family of protein phosphatases: The bacterial chemotaxis proteins CheC and CheX. Molecular Cell 16(4) 563-574.
- Pause, A., et al. (1994). DOMINANT-NEGATIVE MUTANTS OF MAMMALIAN TRANSLATION INITIATION-FACTOR EIF-4A DEFINE A CRITICAL ROLE FOR EIF-4F IN CAP-DEPENDENT AND CAP-INDEPENDENT INITIATION OF TRANSLATION. Embo Journal 13(5) 1205-1215.

- Pei, S., A. Doye and P. Boquet (2001). Mutation of specific acidic residues of the CNF1 T domain into lysine alters cell membrane translocation of the toxin. Molecular Microbiology 41(6) 1237-1247.
- Ray, K., et al. (2009). Life on the inside: the intracellular lifestyle of cytosolic bacteria. Nature Reviews Microbiology 7(5) 333-340.
- Reechaipichitkul, W. (2004). Pulmonary melioidosis presenting with right paratracheal mass. Southeast Asian Journal of Tropical Medicine and Public Health 35(2) 384-387.Rippere-Lampe, K. E., et al. (2001). Cytotoxic necrotizing factor type 1-positive Escherichia coli causes increased inflammation and tissue damage to the prostate in a rat prostatitis model. Infection and Immunity 69(10) 6515-6519.
- Rippere-Lampe, K. E., et al. (2001). Mutation of the gene encoding cytotoxic necrotizing factor type 1 (cnf(1)) attenuates the virulence of uropathogenic Escherichia coli. Infection and Immunity 69(6) 3954-3964.
- Rosario, M. M. L., et al. (1995). CHEMOTACTIC METHYLATION AND BEHAVIOR IN BACILLUS-SUBTILIS - ROLE OF 2 UNIQUE PROTEINS, CHEC AND CHED. Biochemistry 34(11) 3823-3831.
- Rosario, M. M. L. and G. W. Ordal (1996). CheC and CheD interact to regulate methylation of Bacillus subtilis methyl-accepting chemotaxis proteins. Molecular Microbiology 21(3) 511-518.
- Rossmann, M. G. and D. M. Blow (1962). DETECTION OF SUB-UNITS WITHIN CRYSTALLOGRAPHIC ASYMMETRIC UNIT. Acta Crystallographica 15(JAN10) 24-&.
- Rotz, L. D., et al. (2002). Public health assessment of potential biological terrorism agents. Emerging Infectious Diseases 8(2) 225-230.
- Rullmann, J. A. C., M. N. Bellido and P. T. Vanduijnen (1989). THE ACTIVE-SITE OF PAPAIN - ALL-ATOM STUDY OF INTERACTIONS WITH PROTEIN MATRIX AND SOLVENT. Journal of Molecular Biology 206(1) 101-118.
- Sambrook, J., E. F. Fritsch and T. Maniatis (1989). MOLECULAR CLONING A LABORATORY MANUAL SECOND EDITION VOLS. 1 2 AND 3. Sambrook, J., E. F. Fritsch and T. Maniatis. Molecular Cloning: a Laboratory Manual, Second Edition, Vols. 1, 2 and 3. Xxxix+Pagination Varies(Vol. 1); Xxxiii+Pagination Varies(Vol. 2): Xxxii+Pagination Varies(Vol. 3) Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York, USA. Illus. Paper XXXIX+PAGINATION VARIES(VOL 1), XXXIII+PAGINATION VARIES(VOL 2), XXXII+PAGINATION VARIES(VOL 3).
- Sanger, F. (1988). SEQUENCES, SEQUENCES, AND SEQUENCES. Annual Review of Biochemistry 57 1-28.

- Sanger, F., S. Nicklen and A. R. Coulson (1977). DNA SEQUENCING WITH CHAIN-TERMINATING INHIBITORS. Proceedings of the National Academy of Sciences of the United States of America 74(12) 5463-5467.
- Saulmon, M. M., E. Karatan and G. W. Ordal (2004). Effect of loss of CheC and other adaptational proteins on chemotactic behaviour in Bacillus subtilis. Microbiology-Sgm 150 581-589.
- Schmidt, G. and K. Aktories (1998). Bacterial cytotoxins target Rho GTPases. Naturwissenschaften 85(6) 253-261.
- Schmidt, G., et al. (1997). Gln 63 of Rho is deamidated by Escherichia coli cytotoxic necrotizing factor-1. Nature 387(6634) 725-729.
- Schumann, R. R. (1992). FUNCTION OF LIPOPOLYSACCHARIDE (LPS)-BINDING PROTEIN (LBP) AND CD14, THE RECEPTOR FOR LPS/LBP COMPLEXES - A SHORT REVIEW. Research in Immunology 143(1) 11-15.
- Sheldrick, G. M. (2008). A short history of SHELX. Acta Crystallographica Section A 64 112-122.
- Steen, H. and M. Mann (2004). The ABC's (and XYZ's) of peptide sequencing. Nature Reviews Molecular Cell Biology 5(9) 699-711.
- Stone, R. (2007). Infectious disease - Racing to defuse a bacterial time bomb. Science 317(5841) 1022-1024.
- Swerdlow, H. and R. Gesteland (1990). CAPILLARY GEL-ELECTROPHORESIS FOR RAPID, HIGH-RESOLUTION DNA SEQUENCING. Nucleic Acids Research 18(6) 1415-1419.
- Szurmant, H., et al. (2003). Bacillus subtilis hydrolyzes CheY-P at the location of its action, the flagellar switch. Journal of Biological Chemistry 278(49) 48611-48616.
- Szurmant, H., T. J. Muff and G. W. Ordal (2004). Bacillus subtilis CheC and FliY are members of a novel class of CheY-P-hydrolyzing proteins in the chemotactic signal transduction cascade. Journal of Biological Chemistry 279(21) 21787-21792.
- Taylor, G. L. (2010). Experimental phasing and radiation damage Introduction to phasing. Acta Crystallographica Section D-Biological Crystallography 66 325-338.
- Toker, A. S. and R. M. Macnab (1997). Distinct regions of bacterial flagellar switch protein FliM interact with FliG, FliN and CheY. Journal of Molecular Biology 273(3) 623-634.
- Travaglione, S., A. Fabbri and C. Fiorentini (2008). The Rho-activating CNF1 toxin from pathogenic E. coli: a risk factor for human cancer development? Infectious agents and cancer 3 4.
- Vidyalakshmi, K., et al. (2008). Tuberculosis mimicked by melioidosis. International Journal of Tuberculosis and Lung Disease 12(10) 1209-1215.

- Wackett, L. P. and D. T. Gibson (1988). DEGRADATION OF TRICHLOROETHYLENE BY TOLUENE DIOXYGENASE IN WHOLE-CELL STUDIES WITH PSEUDOMONAS-PUTIDA F1. Applied and Environmental Microbiology 54(7) 1703-1708.
- Washington, E. J., M. J. Banfield and J. L. Dangi (2013). What a Difference a Dalton Makes: Bacterial Virulence Factors Modulate Eukaryotic Host Cell Signaling Systems via Deamidation. Microbiology and Molecular Biology Reviews 77(3) 527-539.
- Wiersinga, W. J., et al. (2006). Melioidosis: insights into the pathogenicity of Burkholderia pseudomallei. Nature Reviews Microbiology 4(4) 272-282.
- Winn, M. D., et al. (2011). Overview of the CCP4 suite and current developments. Acta Crystallographica Section D-Biological Crystallography 67 235-242.
- Winter, G., C. M. C. Lobley and S. M. Prince (2013). Decision making in xia2. Acta Crystallographica Section D-Biological Crystallography 69 1260-1273.
- Wongtrakoongate, P., et al. (2007). Comparative proteomic profiles and the potential markers between Burkholderia pseudomallei and Burkholderia thailandensis. Molecular and Cellular Probes 21(2) 81-91.
- Wuthiekanun, V. and S. J. Peacock (2006). Management of melioidosis. Expert review of anti-infective therapy 4(3) 445-455.
- Yee, V. C., et al. (1994). 3-DIMENSIONAL STRUCTURE OF A TRANSGLUTAMINASE - HUMAN BLOOD-COAGULATION FACTOR-XIII. Proceedings of the National Academy of Sciences of the United States of America 91(15) 7296-7300.
- Yu, H. and K. S. Kim (2010). Ferredoxin Is Involved in Secretion of Cytotoxic Necrotizing Factor 1 across the Cytoplasmic Membrane in Escherichia coli K1. Infection and Immunity 78(2) 838-844.