University College London

Wolfson Institute for Biomedical Research

Division of Medicine

Fragment Based Drug Discovery and Structural Biology of Norovirus 3CL Protease and Other Targets

A submission for the degree of Doctor of Philosophy

Jingxu Guo

December 2017

Declaration

I, Jingxu Guo, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed

_____

Jingxu Guo

Abstract

X-ray crystallography has been widely used for determination of protein structures for many years. Protein structures are essential in advancing our understanding of their function, thus providing useful information in many areas, such as drug discovery. Seven different proteins have been studied by X-ray crystallography and other methods, and the information has been gained with the intention of applying it to real-life challenges in the biomedical sphere.

The outbreaks of human epidemic nonbacterial gastroenteritis are mainly caused by noroviruses. Viral replication requires a 3C-like cysteine protease ($3CL^{pro}$) which processes a 200 kDa viral polyprotein into six functional proteins. The $3CL^{pro}$ has attracted much interest due to its potential to act as a target for antiviral drugs. The ligand-free crystal structure of the Southampton norovirus $3CL^{pro}$ (SV3CP) had been determined to 1.3 Å and a system for growing high-quality crystals has been established. This allowed for crystal-based fragment screening to be performed with non-covalent fragments, which identified several hits that will guide drug discovery for SV3CP. Screening with covalent fragments is on-going.

Three mutants of *Bacillus megaterium* porphobilinogen deaminase affecting a key catalytic residue have been analysed. Comparison with the wild-type enzyme shows significant domain movements and suggests that the enzyme adopts ‘open’ and ‘closed’ conformations in response to substrate binding.

Potato cathepsin D inhibitor (PDI) is a glycoprotein composed of 188 amino acids which inhibits both the aspartic protease cathepsin D and the serine protease

trypsin. The first crystal structure of PDI has been determined to a resolution of 2.1 Å, revealing that PDI adopts a typical *β*-trefoil fold with several protruding inhibitory loops, as is typical of the Kunitz-family protease inhibitors.

The family B DNA polymerase from *Pyrobaculum calidifontis* (Pc-polymerase) is distinct from other homologues (e.g. those from *T. gorgonarius*, *T. Kodakarensis* and *P. furiosus*) with a low amino acid sequence identity of 37%. The crystal structure of Pc-polymerase has been refined to a resolution of 2.8 Å and several unique features have been identified which may account for its high processivity and thermostability. A complex model with the primer-template duplex of DNA suggests that large movements of the thumb domain occur upon DNA binding.

The type III pullulan hydrolase from *Thermococcus kodakarensis* (TK-PUL) possesses both pullulanase and *α*-amylase activities and has many potential applications in the industrial food processing sector. The crystal structure of TK-PUL represents the first type III pullulan hydrolase to be analysed, revealing N-terminal and C-terminal domains with differences from homologous structures.

The crystal structure of the L-asparaginase from *Thermococcus kodakarensis* (TkA) has been determined at 2.2 Å, revealing a number of distinctive structural features. The enzyme has many applications in food processing and chemotherapy.

Finally, on-going studies of juvenile hormone diol kinase (JHDK) are described and a summary of the structures presented in earlier chapters of the thesis is given.

Impact statement

Structural biology has a major impact on studies of living systems at every level of organisation by offering a profound understanding of biological function in terms of molecular and supra-molecular structure. Of the various methods currently available for studying complex structures and assemblies, X-ray diffraction and cryo-electron microscopy (cryo-EM) stand out as the predominant techniques for providing information at molecular- and, in some cases, atomic-resolution on individual proteins and other biological macromolecules. In particular, the combination of X-ray diffraction, cryo-EM and solution scattering has probably had the most significant impact on studies of key biological assemblies that are central to the processes of life such as DNA replication and protein synthesis as well as cell division and signalling.

The astounding successes of structural biology in recent decades have impacted greatly on the physical, engineering and computational sciences, and methodological developments in structural biology are highly synergistic with technical advances in these fields. However, arguably the greatest societal impact of structural biology is in the pharmaceutical sector, where it provides an unsurpassed level of detail on macromolecules that are pivotal to the processes of health and the treatment of disease. Structural biology is the only experimental technique which permits the rational design of therapeutic agents and there are many examples of its highly successful use; the structure-based discovery of drugs for hypertension, AIDS and influenza are just a few examples of how it impacts on the health and quality of life for hundreds of millions of people with otherwise untreatable, terminal diseases.

Our work on *in-crystallo* compound screening against the noroviral 3CL protease will have an important bearing on engineering novel compounds as therapeutic agents through the elaboration of the inhibitors we have discovered. In addition, our work on porphobilinogen deaminase bears heavily on the study of its catalytic mechanism and how it relies on rigid body movements of the protein domains. DNA polymerase is recognised as playing a highly important role in the extraordinarily complex and co-ordinated enzymatic processes of DNA replication and error-correction. Studies of pullulan hydrolase and asparaginase from various hyperthermophilic species has yielded key information on the thermostability and catalytic mechanisms of these enzymes. Our characterisation of these proteins by a number of complementary methods, which are at the forefront of methodological developments in the field, is likely to impact heavily on many other research areas and has provided an interesting, informative and influential training.

The current global market for antibacterial drugs is estimated to be around £40 billion and is projected to rise substantially over the next few years. The high and ever-rising incidence of drug resistance in bacteria and viruses means that there is an overriding concern that the pharmaceutical sector has not been able to discover new antibiotic and antiviral therapies fast enough. The success of our preliminary structure-based drug screening studies of the 3CL protease attests that our work forms a very sound platform for discovery of novel antiviral compounds which are specific for a uniquely exploitable target.

Table of Contents

List of Figures

Chapter 1 Protein X-ray Crystallography

List of Tables

List of abbreviations

$3CL^{pro}$	3C-like protease
Abs	Absorbance
AIDS	Acquired immune deficiency syndrome
AIP	Acute intermittent porphyria
5-ALA	5-aminolaevulinic acid
ALAD	5-aminolaevulinic acid dehydratase
ALAS	5-aminolaevulinic acid synthase
ALL	Acute lymphoblastic leukaemia
AMP	Adenosine monophosphate
API-A	Arrowhead proteinase inhibitor A
APIs	Aspartic protease inhibitors
ASU	Asymmetric unit
ATP	Adenosine triphosphate
bp	Base pair
BPBGD	The porphobilinogen deaminase from *Bacillus Megaterium*
$CC_{1/2}$	Half-dataset correlation coefficient
CCD	Charge-coupled device
CCP4	Collaborative Computational Project Number 4
CDase	Cyclomaltodextrinase

CoA	Coenzyme A
Coot	Crystallographic Object-Oriented Toolkit
DCG	Dynamic Contact Graphs
DIALS	Diffraction Integration for Advanced Light Sources
DLS	Diamond Light Source
DMAB	*Para*-dimethylaminobenzaldehyde
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
DPM	Dipyrromethane
dSCP2	*Drosophila melanogaster* sarcoplasmic calcium-binding protein-2
EC	Enzyme commission number
EcA	The L-asparaginase I from *Escherichia coli*
EcAII	The L-asparaginase II from *Escherichia coli*
EDO	Ethylene glycol
EF-hand	Elongation factor hand
ER	Endoplasmic reticulum
ErAII	The L-asparaginase II from *Erwinia chrysanthemi*
ESBRI	Evaluating the Salt Bridges in Proteins
ETI	*Erythrina* trypsin inhibitor

FMDV	Foot-and-mouth disease virus
FOM	Figure of Merit
FUT2	Alpha-(1,2)-fucosyltransferase 2
GABA	Gamma-aminobutyric acid
GTP	Guanosine-5'-triphosphate
HBGA	*Histo*-blood group antigen
HCV	Hepatitis C virus
HIV	Human immunodeficiency virus
HMB	Hydroxymethylbilane
hrm	High remote
HRV	Human rhinovirus
IC_{50}	Half maximal inhibitory concentration
IEM	Immune electron microscopy
if	Inflection
IPTG	Isopropyl *β*-D-1-thiogalactopyranoside
JH	Juvenile hormone
JHa	Juvenile hormone acid
JHd	Juvenile hormone diol
JHDK	Juvenile hormone diol kinase
JHdp	Juvenile hormone diol phosphate

JHE	Juvenile hormone esterase
JHEH	Juvenile hormone epoxide hydrolase
kDa	Kilodalton
K_i	Inhibitory constant
K_M	Michaelis constant
KOD1	DNA polymerase 1 from *T. kodakarensis*
LB	Luria Broth
LC/MS	Liquid chromatography/mass spectrometry
LLG	Log-likelihood gain
MAD	Multi-wavelength anomalous dispersion
MAPI	Michael acceptor peptidyl inhibitor
MIR	Multiple isomorphous replacement
MR	Molecular replacement
mTOR	Mechanistic target of rapamycin
MX	Macromolecular Crystallography
NAG	N-acetylglucosamine
NCS	Non-crystallographic symmetry
NS	Non-structural protein
NV	Norovirus
OD	Optical Density

ORF	Open reading frame
P8K	Polyethylene glycol 8000
PanDDA	Pan-Dataset Density Analysis
PBG	Porphobilinogen
PBGD	Porphobilinogen deaminase
PBGS	Porphobilinogen synthase
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PDI	Potato cathepsin D inhibitor
PEG	Polyethylene glycol
PEO	Progressive external ophthalmoplegia
Pf	*Pyrococcus furiosus*
PfA	The L-asparaginase from *Pyrococcus furiosus*
pfu	DNA polymerases from *P. furiosus*
Ph.P	Phasing power
PhA	The type I L-asparaginase from *Pyrococcus horikoshii*
PHENIX	Python-based Hierarchical Environment for Integrated Xtallography
PIs	Protease inhibitors
pk	Peak
PSPI	Potato serine protease inhibitor

PV	Poliovirus
RASI	Rice bifunctional alpha-amylase subtilisin inhibitor
RdRp	RNA-dependent RNA polymerase
RF	Rotation function
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
SAD	Single-wavelength anomalous dispersion
SAPTF	Spherically averaged phased translation function
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SlJHDK	The juvenile hormone diol kinase from *Spodoptera litura*
SIMBAD	Sequence-independent molecular replacement based on available database
SIR	Single isomorphous replacement
SIRAS	Single isomorphous replacement with anomalous scattering
SMMA	The maltogenic amylase from *Staphylothermus marinus*
SPI	Serine protease inhibitor
STI	Soybean trypsin inhibitor
SV3CP	Southampton norovirus 3C-like protease
TA-PUL	The type III pullulan hydrolase from *thermosphaera aggregans*
TF	Translation function

TFZ	Translation function z-score
Tg	*Thermococcus gorgonarius*
ThMA	The maltogenic amylase from *Thermus* sp. IM6501
TIM	Triosephosphate isomerase
Tk	*Thermococcus kodakarensis*
TK-PUL	The type III pullulan hydrolase from *Thermococcus kodakarensis*
TkA	The L-asparaginase from *Thermococcus kodakarensis*
TLS	Translation Libration Screw
TS	Transition state
Ttha1563	The *Thermus thermophilus* HB8 pullulanase
UROS	Uroporphyrinogen synthase
VADAR	Volume Area Dihedral Angle Reporter
WCI	Wheat chymotrypsin inhibitor
WT	Wild type
XDS	X-ray Detector Software
XPV	Xeroderma pigmentosum variant

List of publications

- Guo, J., Coker, A.R., Wood, S.P., Cooper, J.B., Chohan, S.M., Rashid, N., Akhtar, M. Structure and function of the thermostable L-asparaginase from Thermococcus kodakarensis. *Acta Crystallographica. Section D, Structural Biology* 2017, 73, 889-895.

- Guo, J., Zhang, W., Coker, A. R., Wood, S. P., Cooper, J. B., Ahmad, S., Ali, S., Rashid, N. and Akhtar, M. Structure of the family B DNA polymerase from the hyperthermophilic archaeon *Pyrobaculum calidifontis*. *Acta Crystallographica. Section D, Structural Biology* 2017, 73, 420-427.

- Guo, J., Erskine, P., Coker, A.R., Wood, S.P., Cooper, J.B. Structural studies of domain movement in active-site mutants of porphobilinogen deaminase from Bacillus megaterium. *Acta Crystallographica. Section F, Structural Biology Communications* 2017, 73, 612-620.

- Guo, J., Erskine, P. T., Coker, A. R., Wood, S. P., Cooper, J. B., Mares, M. and Baudys, M. Structure of a Kunitz-type potato cathepsin D inhibitor, *Journal of Structural Biology* 2015, 192, 554-560.

- Guo, J., Erskine, P., Coker, A. R., Gor, J., Perkins, S. J., Wood, S. P. and Cooper, J. B. Extension of resolution and oligomerization-state studies of 2,4'-dihydroxyacetophenone dioxygenase from *Alcaligenes sp*. 4HAP, *Acta Crystallographica. Section F, Structural Biology Communications* 2015, 71, 1258-1263.

- Guo, J., Cooper, J. and Wood, S. The structure of endothiapepsin

complexed with a Phe-Tyr reduced-bond inhibitor at 1.35 Å resolution, *Acta Crystallographica Section F, Structural Biology Communications* 2014, 70, 30-33.

- Sousa, F.D., da Silva, B.B., Furtado, G.P., Carneiro, I.S., Lobo, M.D.P., Guan, Y., Guo, J., Coker, A.R., Lourenzoni, M.R., Guedes, M.I.F., Owen, J.S., Abraham, D.J., Monteiro-Moreira, A.C.O., Moreira, R.A. Frutapin, a lectin from *Artocarpus incisa* (breadfruit): cloning, expression and molecular insights. *Bioscience Reports*. 2017, 37, 1-14.

- Mills-Davies, N., Butler, D., Norton, E., Thompson, D., Sarwar, M., Guo, J., Gill, R., Azim, N., Coker, A. and Wood, S., Structural studies of substrate and product complexes of 5-aminolaevulinic acid dehydratase from humans, *Escherichia coli* and the hyperthermophile *Pyrobaculum calidifontis*, *Acta Crystallographica Section D, Structural Biology* 2017, 73, 9-21.

- Chataigner, L., Guo, J., Erskine, P. T., Coker, A. R., Wood, S. P., Gombos, Z. and Cooper, J. B. Binding of Gd^{3+} to the neuronal signalling protein calexcitin identifies an exchangeable Ca^{2+}-binding site, *Acta Crystallographica Section F, Structural Biology Communications* 2016, 72, 276-281.

- Keegan, R., Waterman, D. G., Hopper, D. J., Coates, L., Taylor, G., Guo, J., Coker, A. R., Erskine, P. T., Erskine, S. and Cooper, J. The 1.1 Å resolution structure of a periplasmic phosphate-binding protein from *Stenotrophomonas maltophilia*: a crystallization contaminant identified by

molecular replacement using the entire Protein Data Bank, *Acta Crystallographica Section D, Structural Biology* 2016, 72, 933-943.

- Keegan, R., Lebedev, A., Erskine, P., Guo, J., Wood, S. P., Hopper, D. J., Rigby, S. E. and Cooper, J. B. Structure of the 2,4'-dihydroxyacetophenone dioxygenase from *Alcaligenes sp*. 4HAP, *Acta Crystallographica. Section D, Structural Biology* 2014, 70, 2444-2454.

Submitted

- Guo, J., Keegan, R., Coker, A. R., Wood, S. P., Cooper, J. B., Rashid, N., Muhammad, M. A., Akhtar, M. and Ahmad, N. Structure and function of the type III pullulan hydrolase from *Thermococcus kodakarensis*.

In preparation

- Guo, J., Douangamath, A., Song, W., von Delft, F., Chan, A. W. E., London, N., Coker, A. R., Wood, S. P. and Cooper, J. B. Structure-based fragment screening and structural studies of the Southampton Norovirus 3CL protease.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Jon Cooper, for his truly exceptional supervision, encouragement and supporting during my study. He has taught me that attitude, patience and humility are in many respects key to successful science. I have been extremely lucky to have a supervisor who cared so much about my work and who responded so promptly to my questions and queries. Professor Cooper is also the kind of scientist I one day aspire to become.

I would also like to advance collective thanks to my secondary supervisor Dr Jon Wilden for many fruitful discussions, Professor Steve Wood for advice and explanations in both crystallography and chemistry and for thesis revision, Dr Alun Coker for his support and assistance, Dr Edith Chan, Dr Graham Taylor and Dr Jim Pitts for their constant help.

It is to the people in my laboratory that I also owe a huge thanks, Eyram Adjoatse, Huikai Shi, Kirsty Wood, Peter Hughes, Rebecca Coker, Sen Li, Weixiao Song, Wenjie Chen, William McCabe and Yiwei Guan, for the helpful discussions and enjoyable time.

I am truly grateful to Diamond Light Source facility and staffs for beamtime and support, also to the XChem group, especially Professor Frank von Delft and Dr Alice Douangamath for their patience and long-term support with the fragment screening experiments.

I would also like to thank my parents and my parents-in-law for their encouragement and support in pursuing this study and through it, my aspirations.

A special thanks must also go to my wife Wenling Zhang, for her constant love and support, without which this would have been a much tougher and lonelier experience, also for her expert suggestions and comments as a medicinal chemist.

Chapter 1

Protein X-ray Crystallography

1.1 Overview of Protein X-ray Crystallography

The development of protein X-ray crystallography can be traced back to England in the 1930s when Professor Dorthy Crowfoot Hodgkin and Professor John Desmond Bernal first reported on the diffraction pattern of a protein crystal, pepsin (Bernal and Crowfoot, 1934). Later in the 1950s Sir John Cowdery Kendrew and Max Ferdinand Perutz determined the very first protein structure of sperm whale myoglobin which allowed them to share the 1962 Nobel Prize in Chemistry (Kendrew *et al.*, 1958).

Protein X-ray crystallography is an experimental technique used to generate a 3-dimensional model of a protein molecule. It is essentially a very high resolution form of microscopy which enables visualisation of protein structures at or near atomic level in order to understand their function. Since the distance between individual atoms within a protein molecule is only around 1.5 Å (0.15 nm), it is not possible to see them under visible light (with wavelengths λ = 400-700 nm), because in order to 'see' something in detail, the wavelength of 'light' used must match the object being viewed. X-rays have a wavelength ranging from 0.01 to 10 nm which can be used to 'view' protein molecules in detail. When X-rays hit a target protein crystal, the diffracted beams cannot be focused by lenses so the diffraction patterns are recorded by detectors such as a charge-coupled device (CCD) or a pixel detector. These diffraction patterns are then analysed and used to build models of the protein molecules using computer programs.

There are some reasons why we use protein crystals rather than single molecules during a diffraction experiment. Firstly, the diffraction from a single molecule would be incredibly weak and extremely difficult to measure. Secondly, X-rays

are a form of electromagnetic radiation which has harmful effects on samples known as radiation damage, so single molecules are usually destroyed fairly quickly. The extremely high X-ray doses required by diffraction experiments, damage protein crystals significantly so data collection is always performed at 100K with carefully-chosen strategies in order to reduce radiation damage. Since protein crystals contain trillions (or even more) ordered molecules in 3-dimensions, they are used for data collection so as to magnify the signal to a measurable level and to reduce the effects caused by radiation damage. Figure 1.1 illustrates an overview of protein structure determination by X-ray crystallography, from a purified protein sample to the final structure.

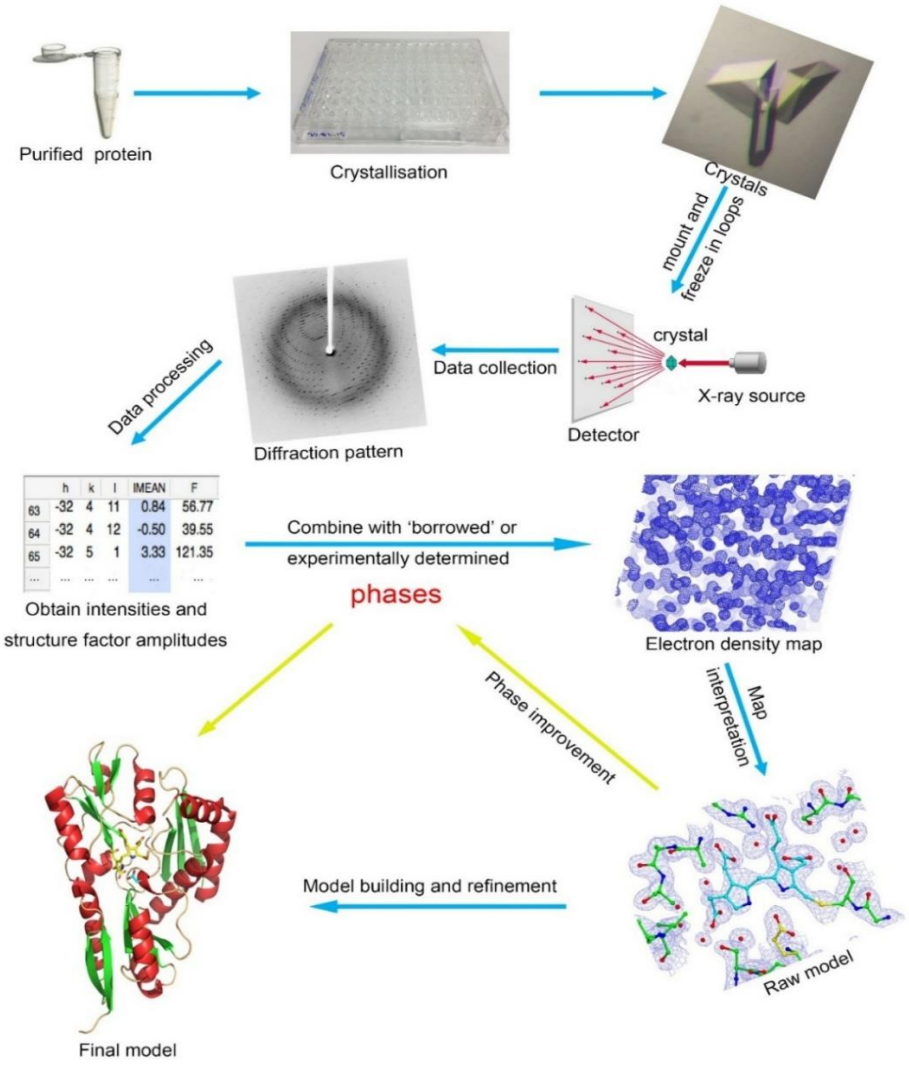


Figure 1.1 An overview of protein structure determination using X-ray crystallography. An initial screening for crystallisation conditions is taken with highly purified protein, which is often followed by optimisation steps in order to get protein crystals with good diffraction quality. The crystals are exposed to an X-ray beam and the diffraction patterns are recorded using a detector. The intensity (I) and position (hkl) of each diffraction spot are measured which provide information of the unit-cell dimensions and space group. An initial electron density map is calculated by combining the structure factor amplitudes ($|F| = I^{1/2}$) and the estimated phases, which are either borrowed from a homologous structure (molecular replacement) or determined by experimental phasing. A model is generated by interpreting the electron density map. Several rounds of model building and refinement are carried out to improve the accuracy of the phases and give a final molecular model which best fits the experiment data.

1.2 Protein crystallisation

1.2.1 The phase diagram

Obtaining well-diffracting crystals plays a key role in protein crystallography and is still the rate limiting step (Doerr, 2006; Geerlof *et al.*, 2006). Crystal growth is usually achieved by taking the protein solution through two rather distinct but inseparable steps: nucleation and growth. This is often performed by gradually reducing the solubility of the protein to reach a point where crystal nuclei can form and grow. However, nucleation is the most difficult problem to address because the molecules pass from a disordered state to an ordered one.

Figure 1.2 is the phase diagram for crystal growth which illustrates the phase transition behaviour of a protein solute affected by the protein and precipitant concentrations. Creation of a supersaturated state is required during crystallisation since the equilibrium is broken in this state and can be re-established by the formation and development of a solid state such as crystals (McPherson and Gavira, 2014). Different crystallisation systems take different routes through the phase diagram, for examples, vapour diffusion method takes route A whilst batch method takes route B, as indicated by the arrows in the figure. In vapour diffusion experiments, the solution is concentrated as water diffuses from the droplets into the well solution until the system enters the nucleation zone from the metastable zone. Then the concentration of protein in solution begins to decrease as a result of nucleation, which brings the system back to the metastable zone where nuclei grow to form crystals. In batch experiments, the system starts directly in the nucleation zone and shifts to the metastable zone, as the protein concentration decreases, where crystal growth from the nuclei happens (Sherwood and Cooper, 2011, Ch. 10).

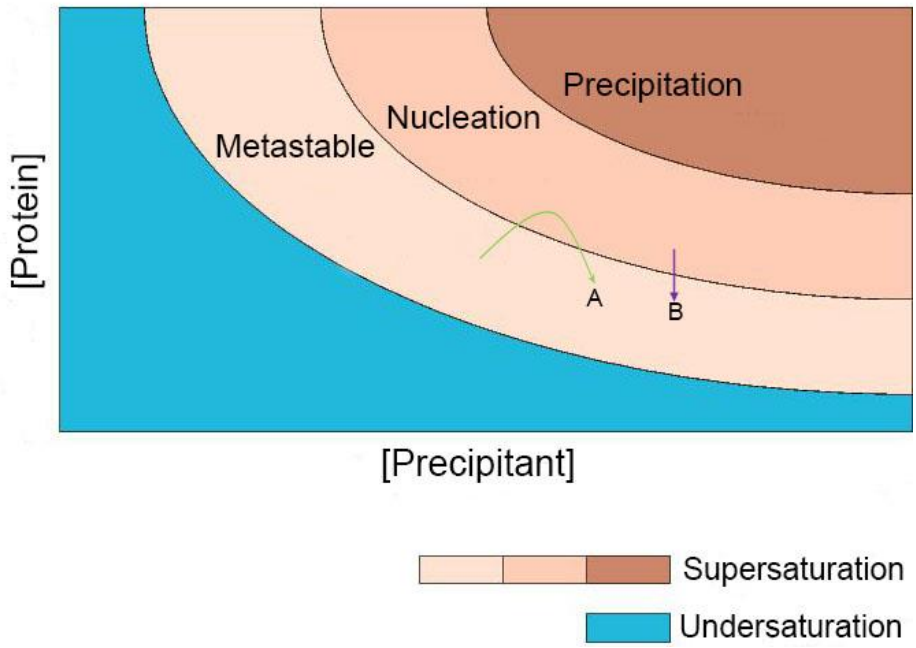


Figure 1.2 The phase diagram. A plot showing the effect of protein (*y* axis) and precipitant (*x* axis) concentrations on crystal growth. The diagram is divided mainly into two large regions: the undersaturation zone in which no precipitation can occur and the supersaturation zone which can be further divided into the metastable, nucleation and precipitation zones. Nuclei can be formed in the nucleation zone. Such nucleation does not happen in the metastable zone, but sizeable and ordered crystals can form in the presence of nuclei. In contrast, protein molecules usually aggregate and precipitate spontaneously in the precipitation zone. The optimal strategy in any crystallisation experiment is to bring the system just into the nucleation zone to form some but not too many nuclei. As crystals grow, the protein concentration reduces and the system is then moved back to the metastable zone where crystals can grow larger and nuclei formation is diminished. (Figure generated based on http://www-structmed.cimr.cam.ac.uk/Course/Crystals/Theory/phases.html).

1.2.2 Experiment techniques

One of the most frequently used methods to crystallise a protein is the *hanging-drop method* that is based on vapour diffusion (Figure 1.3a). In this method, a small volume of highly purified and concentrated protein (e.g. 10-20 mg/ml) is

mixed on a siliconised coverslip with the same volume of the relevant well solution, which often contains precipitant, salt and buffer solution. The coverslip is then sealed upside down tightly on top of the relevant greased well containing the reservoir solution. Since the well solution concentration in the droplet is half of that in the reservoir, the system will equilibrate slowly during the evaporation of water from the droplet and condensation into the reservoir via the vapour phase. Crystallisation will occur around the nuclei which are formed during the concentration of the precipitant in the droplet. Another vapour diffusion method is the *sitting-drop method* (Figure 1.3b) that allows larger volumes of protein to be used and may produce larger crystals. In this method, the droplet sits (always in a well) on top of a plastic bridge above the well solution (Rhodes, 2010, Ch. 3).

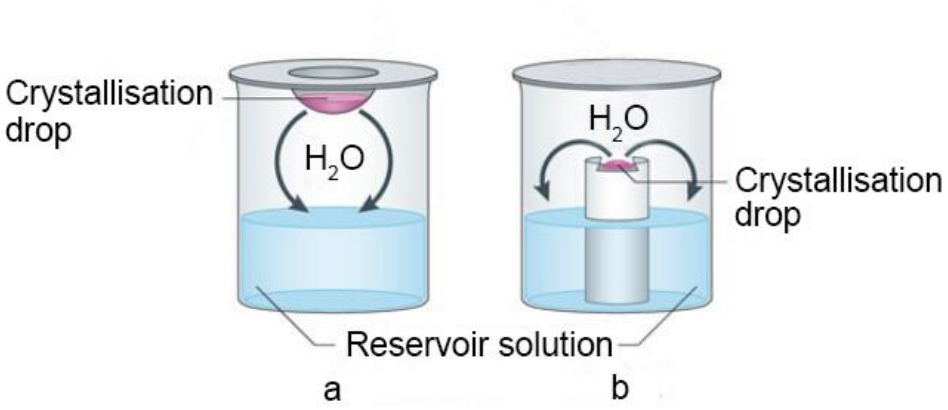


Figure 1.3 Vapour diffusion methods. a) The hanging-drop method. b) The sitting-drop method. [Figure generated based on (Ghosh *et al.*, 2015)].

Other crystallisation methods, such as the *batch method* and dialysis, are used less commonly. In the batch method, the protein is mixed with the precipitant so that their concentrations are just below those required to cause turbidity. The solution is then sealed in tubes or under a layer of oil on a supporter (the *microbatch method*) and left to crystallise. Crystallisation by dialysis involves

separating the protein and precipitant solutions with a semi-permeable membrane in between. The precipitant permeates slowly into the protein solution through the membrane which may lead to crystallisation.

1.2.3 Factors affecting protein crystallisation

Purity of a protein is the most important prerequisite for crystallisability but this should not be overemphasised. In general, a sample that contains at least 95% of the protein molecule of interest is considered pure enough for crystallisation (Sherwood and Cooper, 2011, Ch. 10). This also means that the protein should be homogenous because different forms of oligomer can affect the formation of the crystal lattice. A suitable protein concentration is also important because it is difficult for the system to enter the metastable zone with a low protein concentration. Conversely a protein may precipitate at very high concentration or form showers of tiny crystals. It is usually preferable to use fresh protein since it may degrade with time and the mixture becomes heterogeneous. The presence of a co-factor or a ligand often stabilises the protein and helps crystallisation.

Several other factors in the crystallisation solution, such as pH, salt and precipitant, have an influence on the protein solubility. The solubility of a protein is the lowest at its isoelectric point and most proteins tend to crystallise over a very narrow pH range. Salt concentration affects the solubility of a protein through the phenomena known as ‘salting in’ and ‘salting out’. A lower level of salt ions balances the electrostatic charges of a protein and increases its solubility whilst at a higher level, they start to compete for water molecules with the protein which reduces its solubility. Ammonium sulphate is commonly used as a precipitant in protein crystallisation. Other precipitants are usually polymers or organic solvents

which dehydrate protein solutions in similar ways. Polyethylene glycols (PEGs) in a variety of molecular weights are found to be the most effective both in terms of precipitation ability and cost effectiveness (McPherson, 1976). PEGs compete for water molecules and exert volume exclusion effects which induce separation of protein molecules from solution (Atha and Ingham, 1981). Unlike protein molecules, PEGs have no consistent conformation and occupy large space in solution. In addition, they can reduce the dielectric constant of solutions and increase the strength of electrostatic forces as a result. These effects force the protein molecules out of the solution (McPherson and Gavira, 2014). Temperature is another important factor that influences nucleation and crystal growth as well as the solubility of protein. Low temperature limits bacterial growth and decreases protein degradation. Crystals with better diffraction quality may be obtained at lower temperature which slows down crystal growth and can give more ordered crystals (Park *et al.*, 2010).

Since it is currently still impossible to predict the crystallisation conditions for a protein, initial screens for conditions can be performed using a wide range of sparse matrix crystallisation screening kits (also known as incomplete factorial screening) (Jancarik and Kim, 1991) that are available commercially. They were designed based on conditions that are known to work for other proteins and are generally still being developed. Application of these kits with automatic robots allow for screening of thousands of conditions with drop sizes on the nano-litre scales. Once crystals, or even tiny crystals, of the protein of interest are obtained, further optimisation is performed by gradually varying each of the factors or even omitting some of them carefully which, hopefully, gives well diffracted

crystals (Figure 1.4). This process is usually time-consuming and more difficult than one might expect (McPherson and Gavira, 2014).

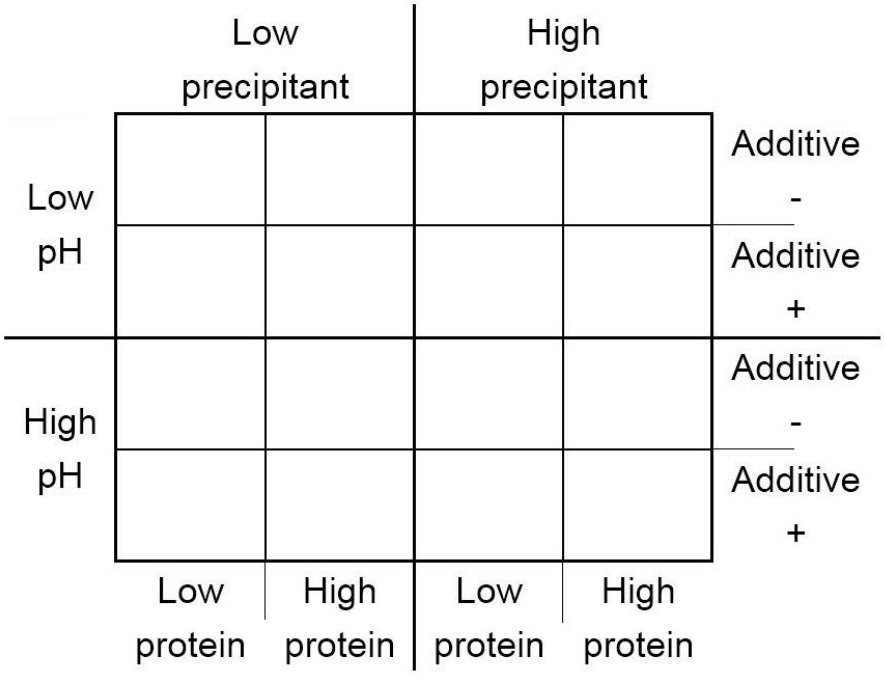


Figure 1.4 A factorial crystal screen. Crystallisation conditions obtained from commercial sparse matrix screening kits can always be optimised by using a factorial crystal screen in order to produce crystals with better diffraction quality. A factorial screen involves testing all or some of the factors systematically by gradually increasing and/or decreasing and/or omitting them entirely (Sherwood and Cooper, 2011, Ch. 10).

1.3 Data collection

Once well-ordered crystals of a suitable size have been obtained, they are usually mounted in loops and kept at 100K using a nitrogen gas stream during data collection. Other strategies may be used as well, for example, data collection at 15K gave more useful information, such as some of the hydrogen positions, compared with those performed at 100K or room temperature (Blakeley *et al.*, 2008; Vandavasi *et al.*, 2016). Also, there has been a growing interest in *in situ* data collection (collecting data directly from the crystallisation devices) at room

temperature which usually gives remarkably low mosaicity (Axford *et al.*, 2012; Michalska *et al.*, 2015). Generally, a crystal is exposed to an X-ray beam and while rotating it gradually during data collection, the resulting reflections are recorded by a detector as diffraction patterns (Figure 1.5). Depending on the symmetry, at least a minimum proportion of the axis rotation need to be collected in order to get a full data set. For example, at least 90-95% of the total number of unique reflections need to be obtained to any given resolution.

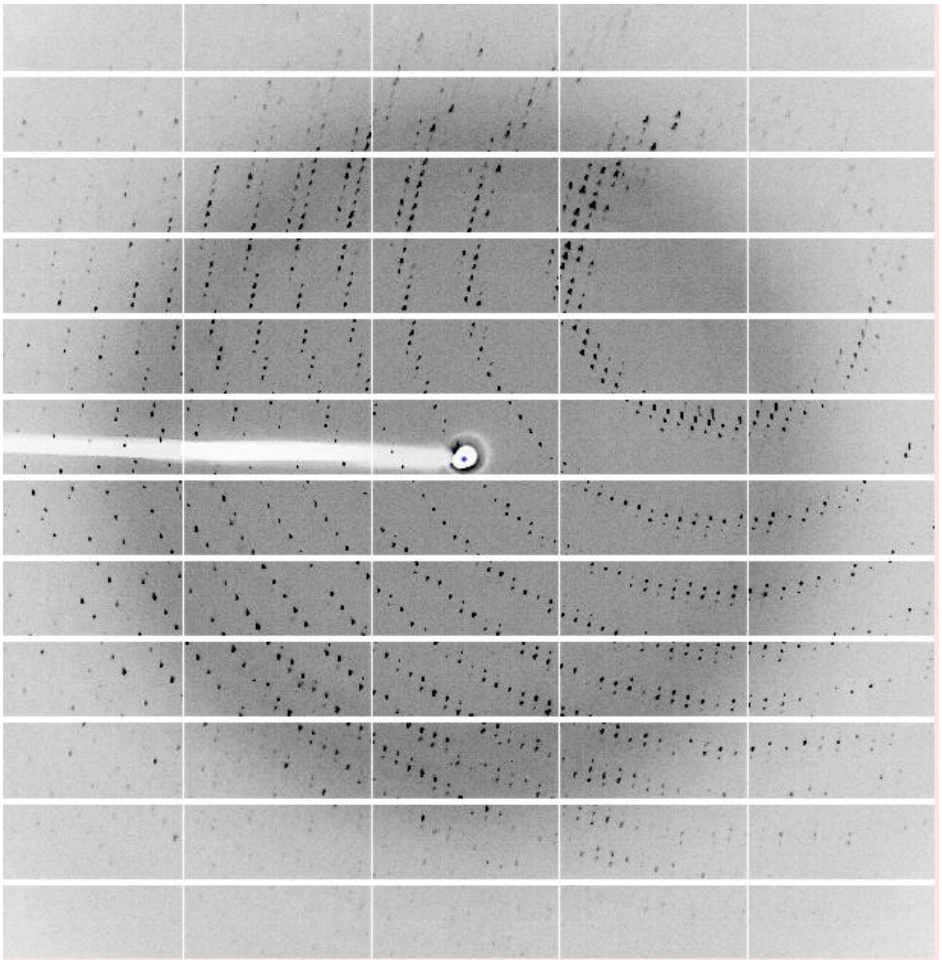


Figure 1.5 A typical diffraction pattern of protein X-ray crystallography. A diffraction pattern of juvenile hormone diol kinase (JHDK) collected at Diamond Light Source station I03 using 0.15° oscillation. Each of the spots (reflections) is the record of an X-ray beam diffracted by the crystal which satisfies Bragg's condition.

1.3.1 Bragg's law

Bragg diffraction occurs when X-ray beams pass through a crystal during data collection. X-ray waves are scattered from lattice planes with an inter-planar distance d_{hkl} by a scattering angle θ (Figure 1.6). When the difference between the path lengths ($2d_{hkl}\sin\theta$) of two scattered waves equals to n times of the wavelength (λ), where n is an integer, they interfere constructively and remain in phase. This is known as Bragg's law:

$$n\lambda = 2d_{hkl}\sin\theta \tag{1.1}$$

Only those very strong reflections, known as Bragg peaks, are recorded on diffraction patterns during data collection, at the positions where the scattering angles satisfy Bragg's law.

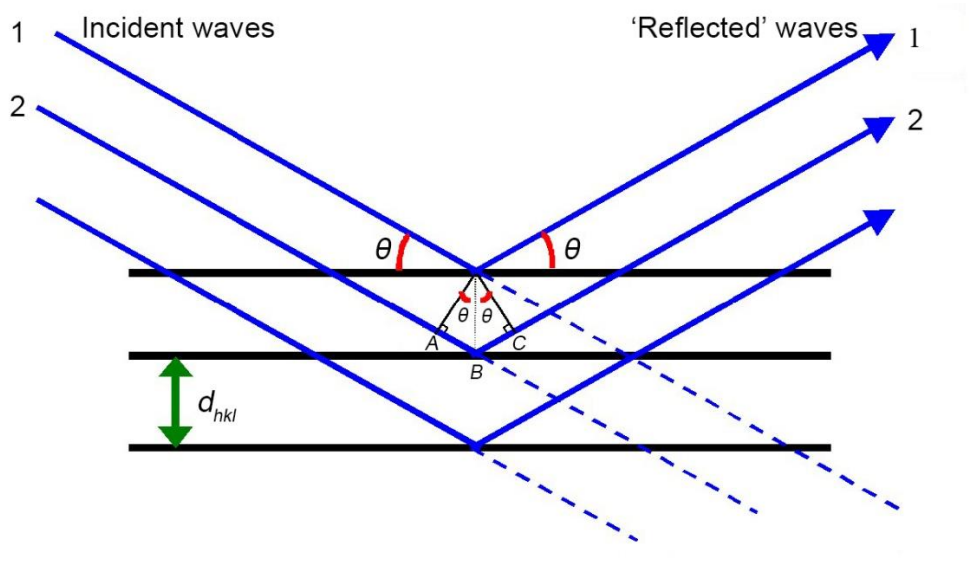


Figure 1.6 An explanation of Bragg's law. Two incident waves with identical wavelengths and phases are scattered by two lattice planes within a set. Wave 2 travels an extra distance which is $AB + BC = 2d_{hkl}\sin\theta$. Constructive interference will occur if the distance is equal to an integer multiple of the wavelength ($n\lambda$). (Figure generated based on http://www.slideshare.net/kumar_vic/solid-state-chemistry-17237117).

1.3.2 X-ray sources

X-rays are a form of electromagnetic radiation which is emitted either when fast-moving charged particles accelerate or decelerate violently, or when an electron in an orbital of higher energy in an atom falls into a vacant orbital which was occupied by another electron.

There are three X-ray devices commonly used to generate X-rays for protein crystallography: rotating anode generators and X-ray tubes which are used as laboratory sources, and synchrotrons. X-rays from a rotating anode generator are generated by colliding electrons into a metal rotating anode in an electric field under vacuum. The electrons are generated by heating a filament, usually tungsten, to reach a high temperature so that *thermionic emission* takes place. The rotation of the anode at high speed effectively dissipates heat and permits higher currents to be used. X-ray tubes have a similar way of producing X-rays but the target anode is fixed. Figure 1.7 illustrates a typical spectrum of X-rays produced by a rotating anode. It is composed of a continuous background spectrum of white radiation and a few peaks due to electronic transitions in the anode material. Copper is the most commonly used anode material which has characteristic peaks as follow (Sherwood and Cooper, 2011, Ch. 11):

$K\alpha_1$ = 1.54051 Å

$K\alpha_2$ = 1.54433 Å

$K\bar{\alpha}$ = 1.54184 Å

$K\beta$ = 1.39217 Å

The Kα_1 and Kα_2 components of the X-ray radiation are the most useful parts and are directed to the crystals as K$\bar{\alpha}$ since they are not always resolved. The others are eliminated or minimised by using devices such as filters or monochromators.

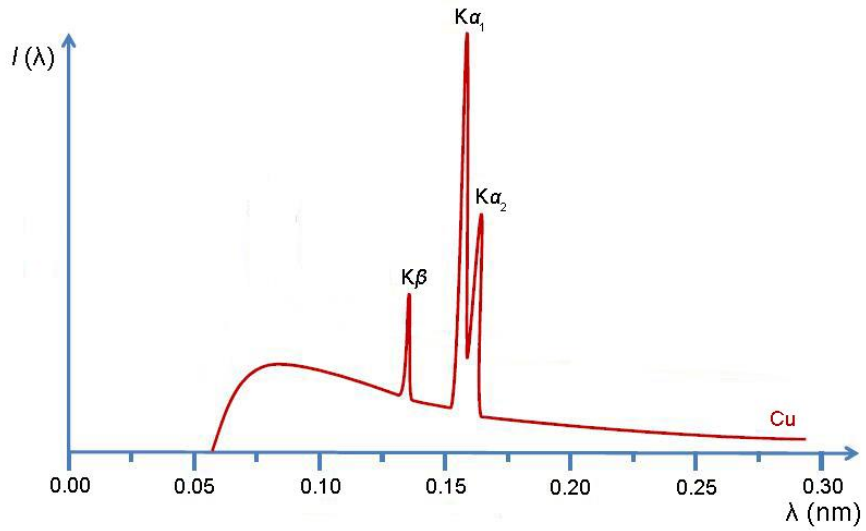


Figure 1.7 The X-ray spectrum derived from an X-ray tube. It shows a continuous spectrum with three peaks labelled as Kα_1, Kα_2 and Kβ. The Kβ peak is generated as a result of an M-K electron transition while the Kα peaks, which are not always resolved, are produced when electrons from the L shell displace those in the K shell (Drenth, 2007, Ch. 2) [Figure generated based on (Döbelin, 2013)].

Synchrotrons are the most powerful X-ray sources that are usually established as national or international facilities. At a synchrotron facility (Figure 1.8), an electron beam, generated by a linear accelerator, is fed into a booster ring to gain energy. The beam is then directed into a big storage ring with a lattice of bending magnets at each corner and 'insertion devices', called wigglers, as well as undulators in the straight sections. Radiation is produced when the charged particles change direction at the corners and oscillate in the insertion devices. The wavelength of the undulator radiation is 'tunable' by adjusting the gaps

between the magnets. X-rays generated by the modern 3rd generation synchrotron sources are very intense which allows the collection of high resolution data with short exposure times, however, radiation damage to crystals has to be considered carefully.

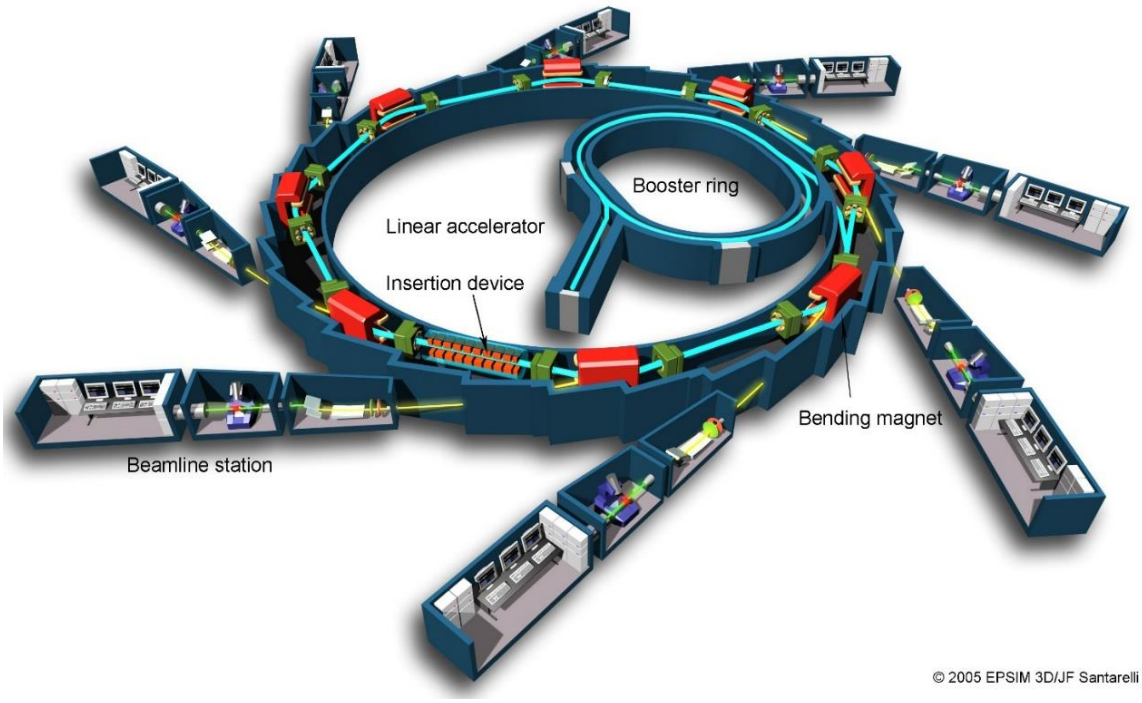


Figure 1.8 The general layout of a synchrotron radiation facility (Synchrotron Soleil). Electrons (light blue) are accelerated to very high speeds by the accelerator and the booster ring before being fed into the storage ring. Electrons are then accelerated in electric fields (green squares) and radiation, especially X-rays, is emitted when the electrons are bent by the bending magnets and deflected by the insertion devices. Radiation is directed into different beamlines for experiments. (Figure from https://en.wikipedia.org/wiki/SOLEIL).

1.3.3 Electronic detectors

Charge-coupled devices (CCDs) were used at synchrotron beamlines and are still used in many laboratory X-ray generators. These detectors consist of a large phosphor screen which absorbs the photons generated by the X-rays. The

photons are transmitted to the photon-sensitive CCD chip by a fibre-optic taper and the incident photons are then converted to electrical charge (in proportion to the number of photons in that specific location) that is stored in capacitors. The charge in each capacitor is finally shifted and dumped into an amplifier which gives a voltage that is digitised and stored.

Nowadays the third generation synchrotron facilities use pixel detectors that convert the energy of X-rays into electrical charges. However, unlike CCD detectors, they record incident X-rays directly without a phosphor-fibre-optic coupling and there is no need to shift each charge because each pixel behaves as a separate detector. The readout time is just a few milliseconds which allows 'real time' diffraction data recording. Pixel detectors have the advantages of low background and broad dynamic range (~ 10^6).

1.3.4 The rotation method

Protein X-ray data collection is usually carried out by slowly rotating the crystal through small angles about a fixed horizontal axis (the ϕ axis) perpendicular to the beam (shown in Figure 1.9). This is referred to as the rotation method. In order to minimise the influence of variations of the motor speed and the intensity of the incident beam during data collection with long exposures, each diffraction image can be recorded by repeating the rotation a few times, thus it is also known as the oscillation method (Arndt *et al.*, 1973; Wilson and Yeates, 1979). However, with modern pixel detectors which have very fast readout, the data are recorded in real time as the crystal is rotated. Data collection by rotating the crystal about only one axis causes the 'blind region' because the spots near the rotation axis in reciprocal lattice will never cross the Ewald sphere. This is usually not a

problem for crystals with moderate to high symmetry because there will be symmetry-related spots that are recorded elsewhere. However, these spots in the 'blind region' can be recorded properly by introducing other rotation axes such as the *K* axis or by collecting data from another randomly placed crystal.

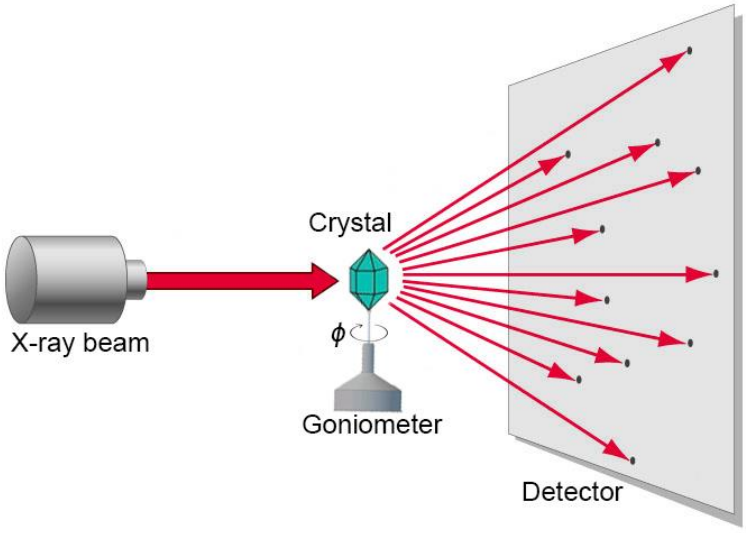


Figure 1.9 The rotation method scheme. Data collection is usually performed by slow rotation of the crystal about the fixed ϕ axis that is usually perpendicular to the X-ray beam. The distance between the crystal and the detector can be optimised in order to get well separated spots filling the active area of the detector.

1.3.5 Radiation damage

When a beam passes through a protein crystal, the absorption of photons causes radiation damage to the crystal because of the photoelectric effect or Compton scattering. The photoelectrons generated cause ionization events which subsequently cause the formation of radical species in the crystal. Also, the energy deposited by the beam raises the temperature in the crystal. The effects of radiation damage include loss of spot intensity, expansion of unit-cell volume, worsening of internal consistency of measurements (R_{merge}), increase of atomic

B-factor values in refined structures, etc. Structural changes, as a result of these effects, may cause non-isomorphism leading to problems in structure determinations (such as by *MAD* or *MIR*). Also when radiation damage is severe, in order to get complete data, multiple datasets must be collected and merged which introduces errors from crystal to crystal variation.

Radiation damage is reduced significantly by lowering the temperature to 100 K during data collection since many of the radical species diffuse much more slowly or not at all, and the effect of the local heating is reduced as well. 'Smart' data collection strategies may be adopted in which a complete data set is collected before significant radiation damage happens.

1.4 Data processing

The aim of data processing is to properly index each recorded reflection, integrate the intensity and calculate the structure factor amplitude for the following structure determination.

1.4.1 Unit-cell parameters, crystal orientation, Bravais lattice, space group, real and reciprocal space parameters

The first step of data processing involves the determination of unit-cell dimensions and crystal orientation. The unit-cell dimensions are described by the lengths of three unit-cell edges (a, b, c) and three crystallographic angles (α, β, γ). The unit-cell dimensions give useful information in the determination of which of the 7 crystal systems (triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal and cubic) the crystal falls into. Combining these 7 crystal systems with 5 lattice types (*P* - primitive, *C* – base centred, *I* – body centred, *F* – face

centred and *R* - rhombohedral) gives 14 Bravais lattices, as shown in Figure 1.10. There are also three possible symmetry operations that can be applied to the asymmetric unit within protein crystals known as rotation, translation and screw (a combination of both rotation and translation). Combining these symmetry operations with the 14 Bravais lattices results in 65 allowed space groups in protein crystallography. This is far less than the 230 space groups because proteins nearly always contain only L-amino acids thus some of the symmetry operations (such as mirrors) are not possible in protein crystals.

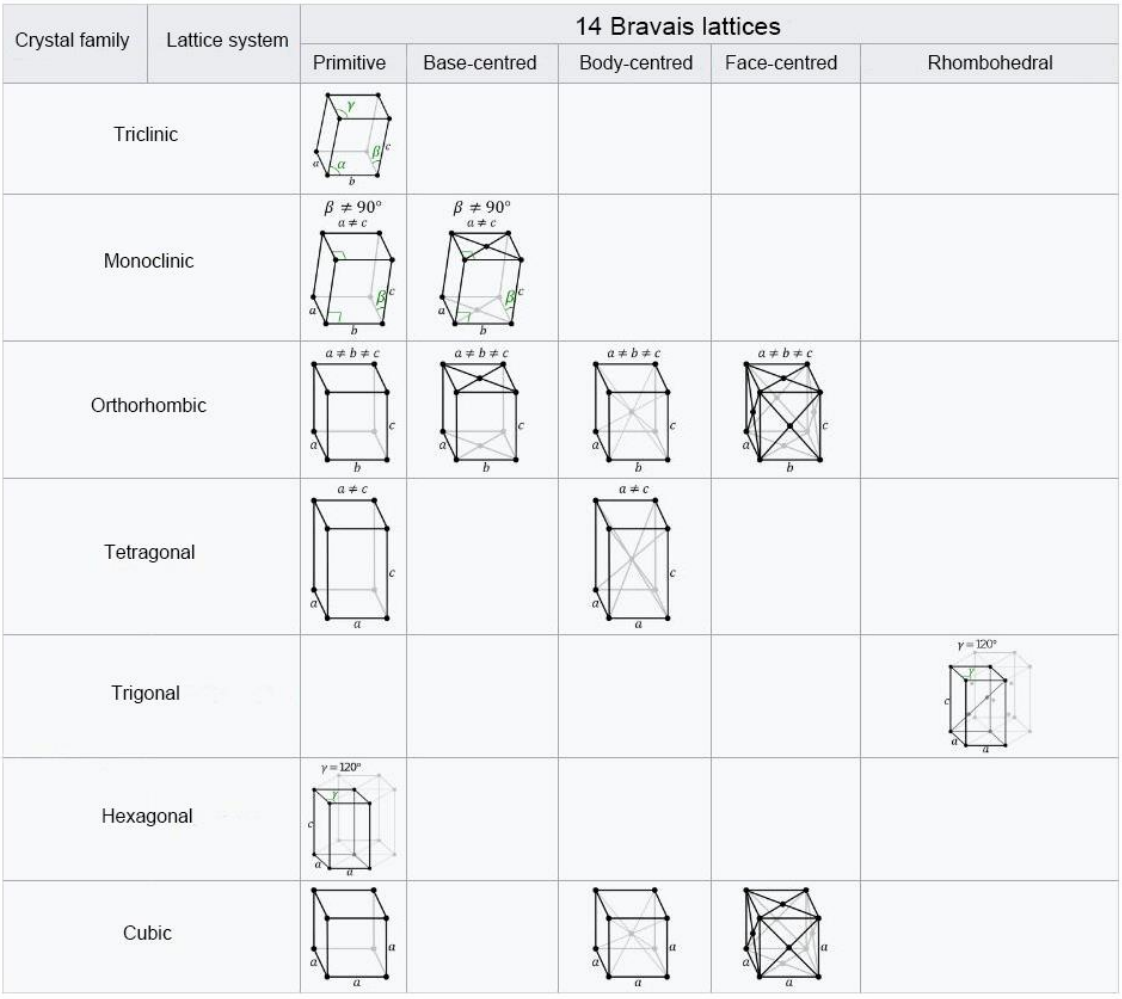


Figure 1.10 The 14 Bravais lattices in 3D space [Figure generated based on wikipedia and Hahn (2002). https://en.wikipedia.org/wiki/Bravais_lattice#CITEREFHahn2002].

Autoindexing allows the determation of unit-cell dimensions and crystal orientation from diffraction images of a randomly placed crystal (Kabsch, 1988) and has been used by many data integration programs such as *iMosflm* (Battye *et al.*, 2011; Powell *et al.*, 2013), *XDS* (Kabsch, 2010) and *DIALS* (Waterman *et al.*, 2013). For any 2-dimensional diffraction image, the calculation of the coordinates (*x*, *y*, *z*) of each spot in the reciprocal space can be carried out automatically by using this method from its coordinates (X_d, Y_d) on the detector and the crystal-to-detector distance D (Leslie, 2006). The coordinates of each spot, together with a ϕ angle (initially assigned as being the midpoint of the ϕ range), usually give enough informaiton to determine the unit-cell dimensions and its orientation. For a successful autoindexing step, the predicted spot positions must agree with the observed spots, by visual inspection and by checking the *RMSD* values between the observed and calculated spot coordinates.

After indexing, the unit-cell parameters and orientation matrix are refined together with other parameters such as the beam position, mosaicity and detector position. This is usually called post-refinement because it requires that one round of integration has already been performed.

During the integration step, the intensity of each spot is calculated by summing the pixels of the peak and subtracting the pixel counts of the background. The background counts cannot be measured directly from the pixels around each diffraction peak within a mask. Whilst for any weak reflections, more accurate methods, such as ‘profile fitting’ (Balzar, 1992; Enzo *et al.*, 1988), can be performed. A ‘profile’ is the average spot shape of reflections nearby in the same part of the detector and is included during intensity calculation. The standard

deviation of each intensity (σI) is calculated which is then used as weights in the subsequent data processing, structure determination and refinement steps.

At this stage, the most probable Laue group symmetry can be determined by using programs like *Pointless* (Evans, 2006). It scores all possible Laue groups that are consistent with the crystal class by matching potential symmetry equivalent reflections based on two scoring mechanisms: a 'likelihood estimate' and a 'combined Z-score'. Relevant systematic absences are tested for all accepted Laue groups which gives a combined score for possible point groups.

When X-rays or other short wavelength radiation is applied to any real crystal lattice, the resulting diffraction spots can be considered as forming a reciprocal lattice. The reciprocal space lattice contains imaginary points whose positions correspond to the normal of the real space planes, and the length of these vectors is equal to the reciprocal of the real interplanar distance. Reciprocal space is the Fourier transform real space, so it is also known as Fourier space. A diffraction pattern of a crystal is generated by the diffraction of the atomic components at the crystal lattice points.

1.4.2 Data reduction – scaling and merging

There are many physical factors affecting the diffraction intensities measured by integration of the recorded reflections thus the intensities are not all on the same scale. For example, the intensity of a spot measured at the start of an experiment could be very different from the measurement of the same spot made later due to crystal shape and radiation damage. Therefore it is very important to correct for these factors before the averaging of the intensities of the symmetry-related reflections is carried out. This process is referred to as scaling which, by fitting a

scaling model (e.g. a scale factor and a temperature factor for each diffraction image) puts all observations on a common scale (Evans, 2006). The quality of scaling, in other words the data quality, is assessed by the inspection of a few parameters based on the internal consistency and comparison of the corrected *I* and σ*I* (Evans, 2006).

When the data have been scaled, the multiple observations for each reflection are then merged into an average intensity. Those observations with large deviations from the mean are rejected during the merging step because they always contain significant errors due to spurious spots resulting from ice and salt diffraction, or from sudden fluctuations in the beam intensity. Programs such as *Scala* (Evans, 2006), *Aimless* (Evans and Murshudov, 2013a) and *XSCALE* (Kabsch, 2010) can be used to perform scaling and merging tasks.

The R_{merge} (also known as R_{sym}, equation 1.2) indicates how well the different observations agree. An overall R_{merge} values of < 5%, 5-10%, 10-20% and > 20% indicate very good, usable, marginal and questionable data quality (McRee, 1999). However, R_{merge} is not always reliable because it inherently depends on multiplicity, so better indicators such as the R_{meas} was introduced. The R_{meas} (also referred to as $R_{r.i.m.}$, equation 1.3) is the redundancy-independent version of the R_{merge} (Diederichs and Karplus, 1997). Another data quality indicator is the $CC_{1/2}$ (referred to as half-dataset correlation coefficient, equation 1.4), which measures the Pearson correlation coefficient of the averaged intensities between two randomly split half subsets of each unique reflection in an unmerged dataset (Karplus and Diederichs, 2012). This is a very good indicator and is commonly used to select the high-resolution cut-off.

$$R_{merge} = \frac{\sum_{hkl}\sum_{j}|I_{hkl,j} - \langle I_{hkl}\rangle|}{\sum_{hkl}\sum_{j} I_{hkl,j}} \quad (1.2)$$

$$R_{meas} = \frac{\sum_{hkl}\sqrt{\frac{N_{hkl}}{N_{hkl}-1}}\sum_{j}|I_{hkl,j} - \langle I_{hkl}\rangle|}{\sum_{hkl}\sum_{j} I_{hkl,j}} \quad (1.3)$$

$$CC_{1/2} = \frac{\sum(I_{x,j} - \langle I_x\rangle)(I_{y,i} - \langle I_y\rangle)}{[\sum(I_{x,j} - \langle I_x\rangle)^2 \sum(I_{y,i} - \langle I_y\rangle)^2]^{1/2}} \quad (1.4)$$

Where $I_{hkl,j}$ is the intensity for the j^{th} reflection, $\langle I_{hkl}\rangle$ is the averaged intensity, N_{hkl} is the number of reflections, $I_{x,j}$ is the intensity for the j^{th} reflection in the random dataset x, $\langle I_x\rangle$ is the averaged intensity for the random half subset x, and so forth.

1.4.3 Calculation of structure factor amplitudes from intensities

Since calculation of the electron density function requires structure factors, F_{hkl}, it is necessary to obtain the F_{hkl} for each reflection. The structure factor F_{hkl} is composed of the structure factor amplitude $|F_{hkl}|$ and its phase. However, the phase cannot be determined directly yet, which forms the well-known 'phase problem' in macromolecular X-ray crystallography. The 'phase problem' can be solved by other methods which will be mentioned in the next section. The structure factor amplitude is calculated from the corresponding intensity according to equation 1.5 shown below (in real practice, one needs to consider the geometry and polarization factor as well as the absorption correction factor):

$$|F_{hkl}| = \sqrt{I_{hkl}} \quad (1.5)$$

1.4.4 Space group determination

The possible Bravais lattice types are examined in the indexing step and the Laue group is usually indicated by using programs like *Pointless*. By detecting the translational symmetry operators, which are indicated in the diffraction pattern as systematic absences, we can determine the possible space group which defines the symmetry in space. However, since these indicators are not always reliable, we should bear in mind that the space group is only a hypothesis until the structure has been solved (Evans, 2011).

1.4.5 Data quality assessment

Besides those data quality indicators, such as the R_{merge} and $CC_{1/2}$ mentioned in the previous sections, there are still a few things that should be checked before any structure determination is carried out. The completeness tells us the percentage of unique reflections that have been measured to a given resolution, and should ideally be above 90-95%. The $I/\sigma I$ value indicates the signal-to-noise ratio and an acceptable value was considered to be 2 or higher in the outer resolution shell, but much lower values are acceptable nowadays since refinement programs are able to handle weak data better than before. Sometimes the crystal order is anisotropic which means there is a directional dependence in diffraction quality. This may lead to high *R*-factors during refinement and give blurred electron density maps which make model building difficult or impossible. The presence of translational non-crystallographic symmetry (NCS) and/or twinning may give partial or wrong solutions so that these problems need to be handled properly.

1.5 Solving the phase problem

1.5.1 The structure factor, electron density and Fourier transforms

In X-ray crystallography, the structure factor F_{hkl} is a mathematical description of how the crystal diffracts the incident radiation producing a reflection on the detector. The structure factor encapsulates the amplitude and phase of the X-rays scattered by the electrons distributed within the crystal, or specifically within the unit cell because the crystal is composed of a great number of unit cells all scattering in phase.

The result from a protein X-ray crystallography experiment is an electron density map into which a structural model of the protein can be fitted. Both the structure factor and the electron density are related by Fourier transforms. The Fourier transform of the structure factor is the electron density and the Fourier transform of the electron density gives the structure factor. This is one of the important properties of the Fourier transform, known as the Fourier inversion theorem. So by taking an inverse Fourier transform of the structure factor equation (1.6), the electron density equation is obtained as shown in equation 1.7.

$$F_{hkl} = V\int\int\int \rho(x,y,z)e^{2\pi i\,(hx+ky+lz)}dxdydz \tag{1.6}$$

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}F_{hkl}e^{-2\pi i(hx+ky+lz)} \tag{1.7}$$

where V is the unit cell volume.

Because the structure factor F_{hkl} is a complex number, equation 1.7 can be rewritten as equation 1.8.

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|\,e^{i\phi_{hkl}}e^{-2\pi i(hx+ky+lz)} \qquad (1.8)$$

1.5.2 The phase problem

From equation 1.8, we know that in order to calculate the electron density we need to know both the structure factor amplitude $|F_{hkl}|$ and the phase ϕ_{hkl} for all the structure factors. As mentioned before, the $|F_{hkl}|$ can be determined from the I_{hkl} according to equation 1.5, while the ϕ_{hkl} cannot be determined directly which constitutes the 'phase problem'. This is shown in the Argand diagram below in Figure 1.11.

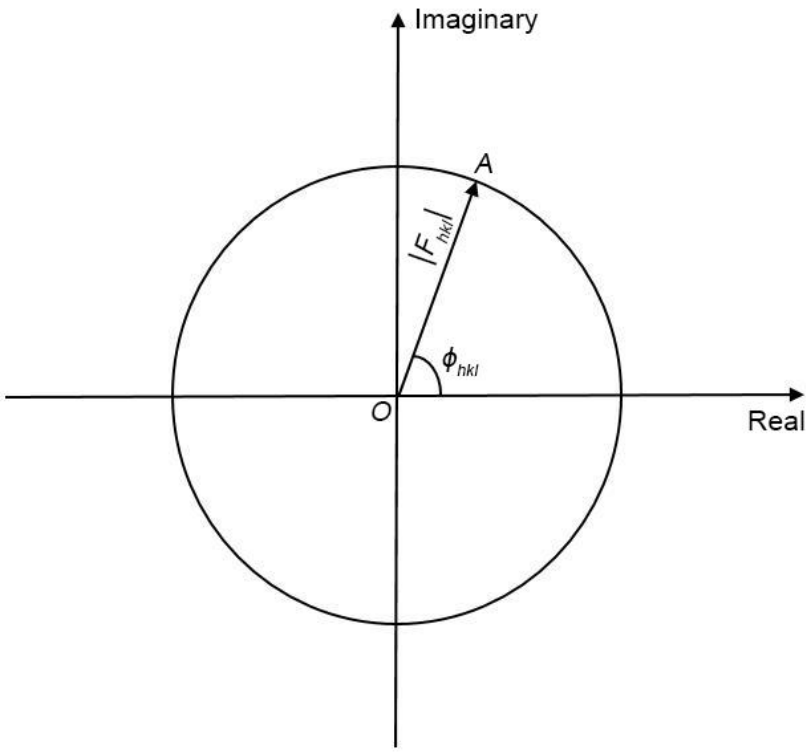


Figure 1.11 The phase problem. The structure factor can be illustrated by an Argand diagram in which the complex number F_{hkl} is represented by its length $|F_{hkl}|$ with the angle ϕ_{hkl}. The amplitude $|F_{hkl}|$ is directly observed but if the phase angle ϕ_{hkl} is not known, F_{hkl} can be any straight line from the origin O to the circle of radius $|F_{hkl}|$.

1.5.3 The Patterson function and Patterson map

The Patterson function is the Fourier transform of the relative intensities and does not require phase information, as shown below:

$$P(u,v,w) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l} |F_{hkl}|^2 \, e^{-2\pi i(hu+kv+lw)} \qquad (1.9)$$

A Patterson map (Figure 1.12) is a vector map derived from the relative Patterson function and represents all interatomic vectors between pairs of atoms. It contains not only intramolecular vectors within a molecule but also intermolecular vectors due to the presence of crystallographic and non-crystallographic symmetry. For a cell containing N atoms, there will be N(N-1) peaks in the Patterson map at points other than the origin and a huge peak at the origin corresponding to the vectors between each atom and itself. So as the number of atoms in the cell increases, it becomes difficult to interpret the Patterson map. However, in protein X-ray crystallography, it can be used in locating certain heavy atoms within the unit cell which is a useful technique in experimental phasing methods. In addition, the Patterson map of the known structure may be used to interpret the Patterson map of the unsolved homologous structure, which is used in some of the molecular replacement programs.

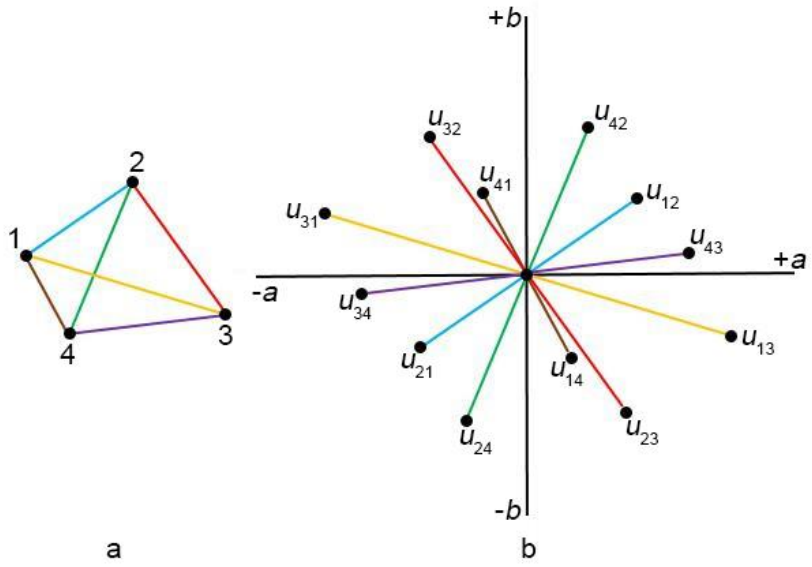


Figure 1.12 The 2D Patterson map for 4 atoms. a) The positions of the 4 atoms. b) The corresponding Patterson map showing the peaks representing all interatomic vectors relative to the origin. These four atoms in the unit cells give rise to 12 [N(N-1)] Patterson peaks at points other than the origin.

1.5.4 Molecular replacement

The molecular replacement (MR) method uses the phases calculated from a known structure as initial phase estimates for the unsolved structure (target structure) which is similar (homologous) to the known structure. The target and the search molecules must have a reasonable ‘similarity’ (e.g. a reasonable amino acid sequence identity or *RMSD* value) to have a good chance of success. The known structure is used as the ‘search model’ to determine the orientation and position of the target molecule in the unit cell. This is performed in two steps, known as the rotation function and the translation function, by molecular replacement. With a successful solution, a preliminary model of the target molecule is obtained with a correct orientation and position in the unit cell. The phases are then calculated based on this model and used with the structure factor

amplitudes from the experiment to produce an electron density map. The search model can be rebuilt to fit into the electron density map followed by converting the amino acid sequence to that of the target molecule.

The basic calculations of molecular replacement (simply illustrated in Figure 1.13 in two-dimensional space) involve the determination of a rotation matrix [*C*] and a translation vector *d*. If the coordinates of the search model and the target structure are represented by the matrices x and x', the transformation between them can be described by equation 1.10:

$$x' = [C]x + d \tag{1.10}$$

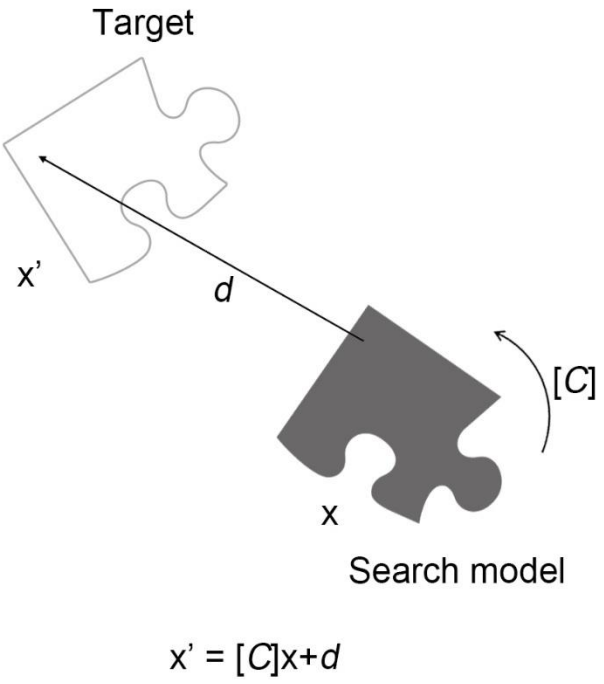


Figure 1.13 The basic concept of molecular replacement. It is carried out by superimposing the search model x on the target molecule x' in the unit cell. The transformation process can be describe by equation x' = [*C*]x + *d* where [*C*] is the rotation matrix that defines the 'new' orientation and *d* is the translation vector which defines the 'new' position of the search model.

Due to more and more structures (search models) being available in the Protein Data Bank (PDB), the number of PDB X-ray depositions solved by molecular replacement has increased dramatically in the recent years. In total, around 60% of X-ray structures were solved by this method between 1970 and Feb 2013 whereas over 70% were solved by this method from 2010 to Feb 2013, as shown in Figure 1.14 (Scapin, 2013).

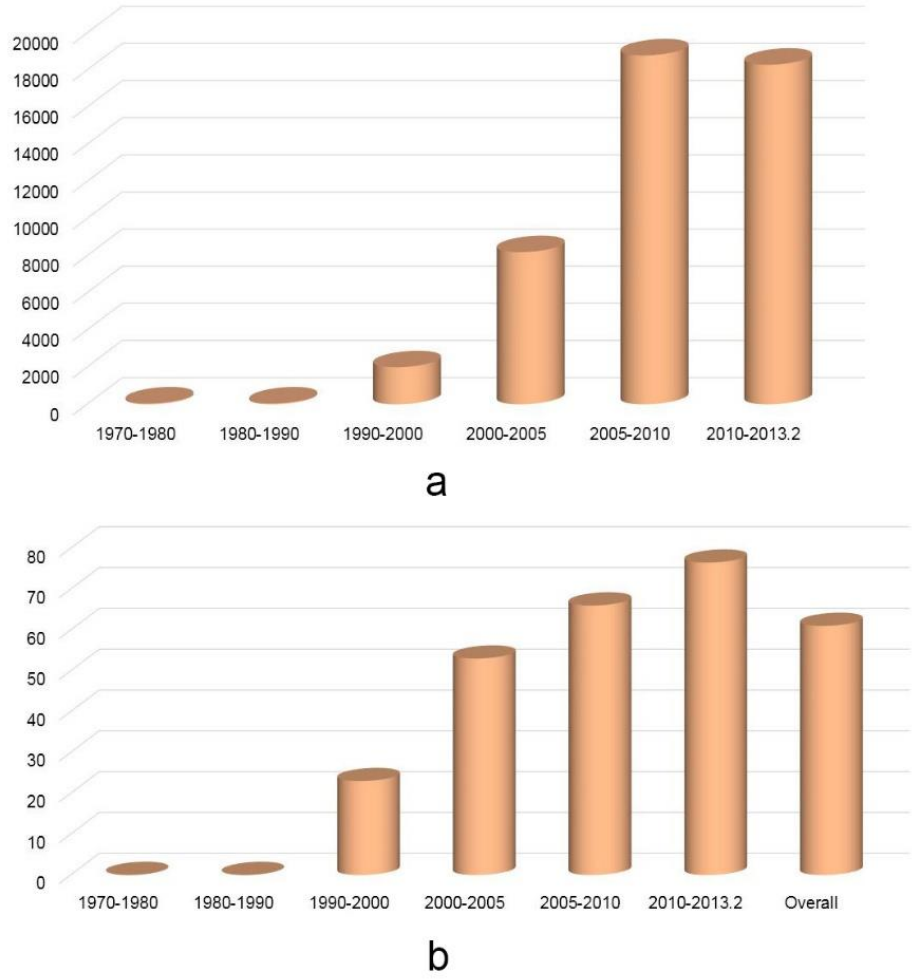


Figure 1.14 PDB deposition statistics. a) Number of X-ray structure depositions solved by molecular replacement in the PDB between 1970 and Feb 2013. b) The number of deposited structures solved by molecular replacement as a percentage of the total number of PDB X-ray depositions to Feb 2013. [Figure generated based on (Scapin, 2013), reproduced with permission of the International Union of Crystallography (http://journals.iucr.org)].

1.5.4.1 The Patterson method

1.5.4.1.1 The rotation function

The rotation function (equation 1.11) exploits the fact that the intramolecular vectors depend only on the orientation of the molecule, not on its position in the unit cell. This function, which is required to determine the orientation of the search model, involves calculating its Patterson function and comparing it with that of the target molecule calculated from the experimental data. The Patterson function of the search model is then repeatedly rotated by small angles and compared with that of the target molecule within a limited radius in three dimensions. The unknown translation can be ignored by restricting the radius at this stage. The agreement of the two Pattersons is assessed by the rotation function.

$$R = \int P_T(u)\mathcal{R}P_S(u)du \qquad (1.11)$$

In the equation, $P_T(u)$ is the target molecule Patterson function given by the matrix u, $\mathcal{R}P_S(u)$ is the search model Patterson function which has been rotated by the matrix [C].

The rotation function can be calculated either in real space or in reciprocal space. The computational process can be speeded up by selecting the strongest grid points of the search model Patterson function and calculating the agreement with the target Patterson function using the equivalent grid points (Huber, 1965). The fast rotation function (Crowther, 1972), which is computationally efficient, calculates the rotation function from spherical harmonics and Bessel functions by the fast Fourier transform (FFT).

1.5.4.1.2 The translation function

Unlike the rotation operation, the intermolecular vectors are sensitive to translation operations while the intramolecular vectors are not affected. The translation operation is based on the crystallographic symmetry-related operators. The Patterson based translation function is defined as:

$$T(t) = \int P_c(u,t) P_o(u) du \qquad (1.12)$$

where P_c (u,t) is the Patterson function of the search model and P_o (u) is the target structure Patterson function.

The translation search involves moving a correctly oriented search model on a grid and inspecting the agreement between the calculated target Patterson function and the Patterson function of the search model at each step (Fujinaga and Read, 1987). The translation operation can be performed in both real and reciprocal space and it can be calculated by the FFT (Harada *et al.*, 1981; Tickle and Driessen, 1996).

1.5.4.2 The maximum likelihood method

The basic theory supporting the maximum likelihood method is quite simple: the best model has the best match with the observed data. The consistency with the observations is assessed statistically by the probability that the observations would have been measured given the model that is being tested. In other words, having the model in the correct orientation and position, there should be a high probability that the data as measured would have been observed. So the likelihood goes higher if changing the model makes the observations more probable, which means the model is getting better. The log of the likelihood is

commonly used in this method because it is easier to deal with. The likelihood function is expressed by:

$$\mathcal{L} = -\sum_{hkl} \log P(|F_o|;|F_c|) \quad (1.13)$$

In real cases, what is needed is actually the probability distribution of the measurements, given as a function of model parameters and sources of errors. Sources of errors are measurement errors and errors in the model. However, measurement errors are less significant than those of the model. The effect of the errors in the model is estimated by introducing a parameter σ_A (Srinivasan and Ramachandran, 1965), which estimates the agreement between the observed and calculated normalised structure factors.

1.5.4.3 Molecular replacement programs

The Patterson method is widely used by molecular replacement programs such as *Molrep* (Vagin and Teplyakov, 2010) and *AMoRe* (Navaza, 1994; Navaza, 2001). In addition to the standard searches, real space searching is also possible. Partial solutions can be used as fixed models to search for more monomers in one more runs. For example, if one molecule (A) is found in the ASU which actually contains two molecules (A and B) in *Phaser MR* (McCoy *et al.*, 2007), the program can accept the input coordinates of molecule A without modification and skip to searh for molecule B.

Phaser, based on the maximum likelihood method, is an efficient molecular replacement program. Both the CCP4 (Winn *et al.*, 2011) and the *PHENIX* (Adams *et al.*, 2010) program suites have *Phaser* included. It has been proved that *Phaser* is significantly better than traditional methods especially in difficult

cases (McCoy *et al.*, 2007; Scapin, 2013). Things such as anisotropy and translational NCS corrections can be performed automatically. A very good feature of *Phaser* and *Molrep* is that it allows possible (or selected) space groups to be searched when there is ambiguity about the crystal symmetry. Also, multiple ensembles can be used as search models. An overview of *Phaser* is shown in Figure 1.15.

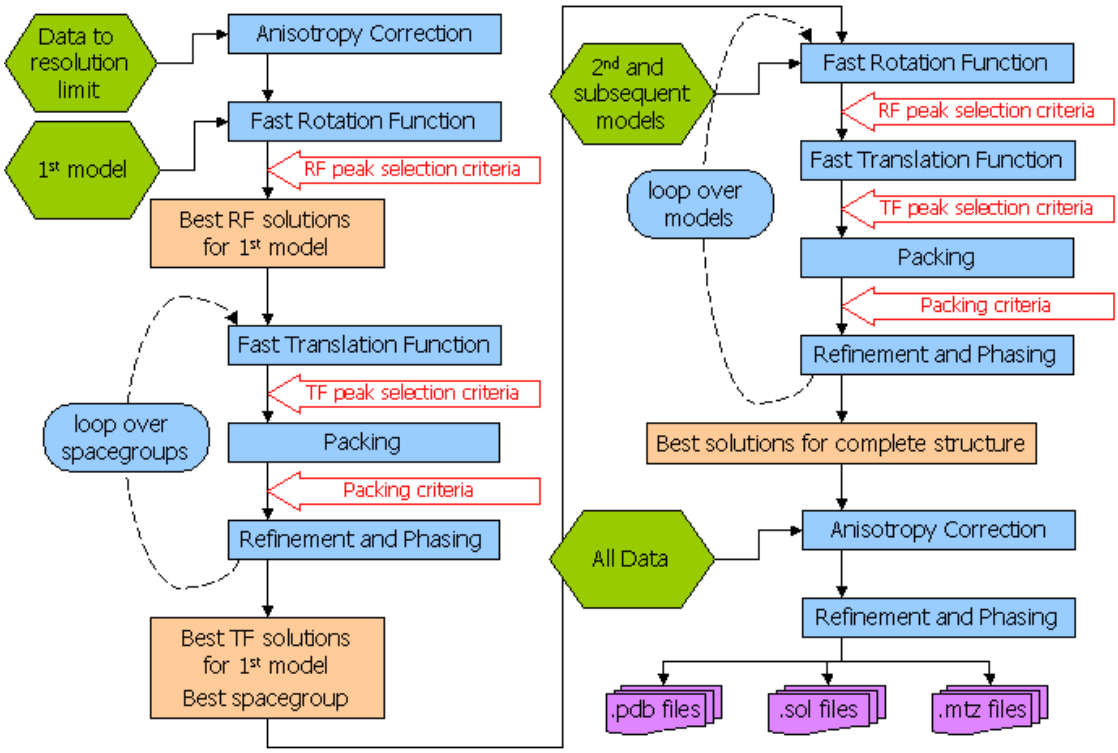


Figure 1.15 An overview for automated molecular replacement in *Phaser*. (Figure is originally from http://www.phaser.cimr.cam.ac.uk/index.php/File:Phaser_MR_auto.gif).

Molecular replacement pipelines such as *MrBUMP* (Keegan and Winn, 2008) and *BALBES* (Long *et al.*, 2008) are available as both computer programs and online services. They start with a structure factor file and a target sequence and produce refined models and phased structure factors sorted by scores. Model selection and modification, molecular replacement and refinement are all carried out automatically.

1.5.5 Experimental phasing

The phase problem may be solved by purely experimental means without borrowing the phase from a homologous structure and this is referred to as experimental phasing. Chemical modifications of the molecules within the crystal are usually required so that the diffraction intensities are altered. In X-ray crystallography, anomalous scattering is a change in the phase of a diffracted X-ray, caused by certain atoms due to strong X-ray absorbance, which is different from that caused by the other atoms in the crystal. This can be achieved by the introduction of electron-rich 'heavy' atoms because atoms such as carbon, nitrogen and oxygen do not contribute to anomalous scattering. Atoms such as Hg or Pt are good candidates to make heavy metal derivatives for experimental phasing. In addition, replacing methionine residues with selenomethionine in a recombinantly expressed protein is also a good way. The development of technology has made it possible to perform experimental phasing with 'not so heavy' atoms such as S and Ca^{2+} (Mueller-Dieckmann *et al.*, 2007), which may be naturally present in the protein already. The anomalous signal from these 'heavy' atoms allows for estimation of the phases of diffraction spots by several methods known as single isomorphous replacement (*SIR*), multiple isomorphous replacement (*MIR*), multi-wavelength anomalous dispersion (*MAD*) or single-wavelength anomalous dispersion (*SAD*). The wavelength of the X-ray beam being used should be close to the absorption edge of the 'heavy' atom during experiment phasing.

1.5.5.1 Characteristics of anomalous scattering

Anomalous scattering happens when the energy of the incident X-ray equals that which is required to promote an electron of an atom to a higher energy orbital. This energy is referred to as the absorption edge. The gained energy is then re-emitted as the excited electron goes back to its original energy level. When anomalous scattering happens, the scattered wave (X-ray) has a component that is phase-shifted by 90°, and each reflection will therefore possess an anomalous component.

1.5.5.2 Friedel's law

A Friedel pair is a couple of Bragg reflections: h, k, l and $\bar{h}$, $\bar{k}$, $\bar{l}$. Friedel's law states that the two members of a Friedel pair have the same amplitude but opposite phase: $|F_{hkl}| = |F_{\bar{h}\bar{k}\bar{l}}|$, $\phi_{hkl} = -\phi_{\bar{h}\bar{k}\bar{l}}$ (Figure 1.16a).

Friedel's law is broken whenever there is anomalous scattering. Two parameters f' and f'', known as the anomalous scattering coefficients, are introduced to modify the structure factors. These two values are dependent on scattering angle and wavelength. As is illustrated in Figure 1.16b, the structure factors are no longer the same since f'' is always 90° ahead in the Argand diagram. The difference between the amplitudes of such structure factors is named a Bijvoet difference (Blow, 2003) and these measurements are fundamental to experimental phasing methods such as *SAD* or *MAD*.

Data are usually collected at three different wavelengths in a *MAD* experiment: a peak wavelength at the absorption edge where f'' is maximum; an inflection point wavelength where f' is at a minimum; and a remote wavelength where f'' is at a

low value. These wavelengths as well as the f’ and f’’ values should be determined by an X-ray fluorescence experiment prior to the *MAD* data collection. In a *SAD* experiment, data are collected at the peak wavelength only but with a high redundancy (Dauter and Adamiak, 2001).

Experimental phasing can be seen as a divide-and-conquer process which is divided into two steps: solving the substructure (e.g. locating heavy atom sites) first and then extend the phase from the substructure to the whole structure.

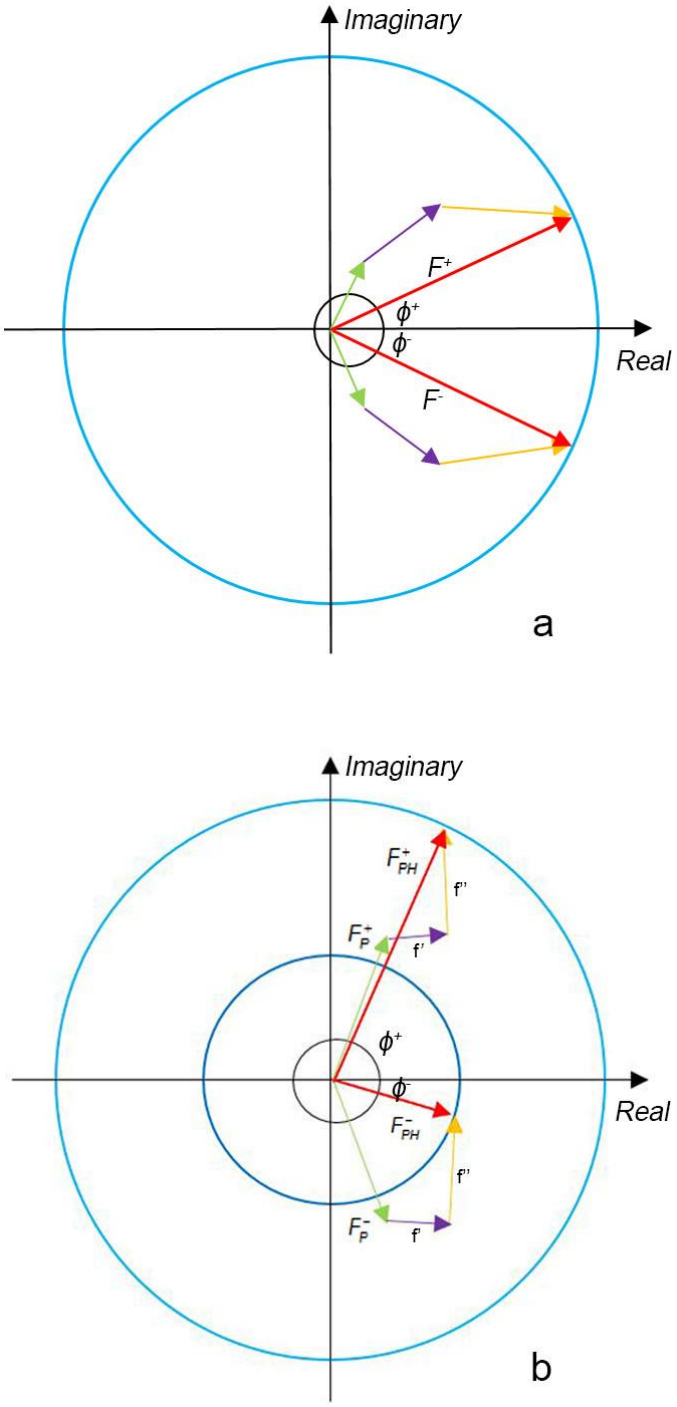


Figure 1.16 Normal and anomalous scattering. a) A Friedel pair composed of F^+ and F^- from a normal scattering. The vectors F^+ and F^- (red) are of identical amplitudes but opposite phase angles, this is due to the contributions of the atomic vectors which have identical amplitudes but opposite phases in each pair (green, purple and yellow). b) A description of the structure factors in anomalous scattering. F_P^+ and F_P^- indicate a Friedel pair without any anomalous contribution while they are modified (by f' and f'') to give F_{PH}^+ and F_{PH}^- when there are anomalous contributions.

1.5.5.3 Identification of heavy atom locations based on the difference Patterson function

If we calculate a Patterson function for a structure containing heavy atoms (*H*), the Patterson peaks will correspond to interatomic vectors between: non-*H* & non-*H* pairs, non-*H* & *H* pairs and *H* & *H* pairs. Since vectors between any non-*H* & non-*H* pairs are the same in the native (*P*) and derivative (*PH*) crystals, they can be cancelled out by calculating a difference Patterson function (equations 1.14 and 1.15) which only has peaks arising from vectors between non-*H* & *H* as well as *H* & *H* pairs.

$$\Delta P = P_{PH} - P_P \tag{1.14}$$

$$\Delta P(u,v,w) = \frac{1}{V}\sum_h\sum_k\sum_l [|F_{PH(hkl)}|^2 - |F_{P(hkl)}|^2]e^{-2\pi i(hu+kv+lw)} \tag{1.15}$$

It is more common to calculate the difference Patterson function from the isomorphous differences (Δ_{iso}) according to equations 1.16 and 1.17, which gives the heavy atom peaks with little contaminating noise.

$$\Delta_{iso} = |F_{PH}| - |F_P| \tag{1.16}$$

$$\Delta P(u,v,w) = \frac{1}{V}\sum_h\sum_k\sum_l [|F_{PH(hkl)}| - |F_{P(hkl)}|]^2 e^{-2\pi i(hu+kv+lw)} \tag{1.17}$$

The difference Patterson function significantly reduces the peak numbers and makes the Patterson map much less complex to interpret. A Patterson function calculated with anomalous differences also gives information on the heavy atom vectors. The anomalous difference is expressed as equation 1.18 and the corresponding anomalous difference Patterson is referred to as equation 1.19.

$$\Delta_{ano} \approx 2 \mid F_H^{"} \mid \sin(\phi_{PH} - \phi_H) \tag{1.18}$$

$$\Delta P(u,v,w) = \frac{1}{V} \sum_h \sum_k \sum_l [2 \mid F_H^{"} \mid \sin(\phi_{PH} - \phi_H)]^2 e^{-2\pi i(hu+kv+lw)} \tag{1.19}$$

Where $F_H^{"}$ corresponds to the imaginary component of the heavy atom scattering.

1.5.5.4 Direct methods for locating heavy atom sites

Direct methods are commonly used for determining phases of chemical molecules or small macromolecules (e.g. < 2000 atoms) diffracting to atomic resolution. However, they can also be used for locating heavy atom sites in experimental phasing. Based on how the structure factor magnitudes are compared, direct methods may be classified as: (i) inequalities, (ii) equalities and (iii) probabilities. To be concise, only the third one is going to be described here because its use is more general. The basic theory of the third class is that given a small set of starting phases, such as might be derived from the origin-fixing set or just random phases, a more complete phase set can be constructed by applying phase probability relationships. These relations are based on the tangent formula shown in equation 1.20, as described by Karle and Hauptman (Karle and Hauptman, 1956). An electron density map can be generated with the expanded phase set.

$$\tan\phi_h \approx \frac{<\mid E_{h'} \mid\mid E_{h-h'} \mid \sin(\phi_{h'} + \phi_{h-h'}) >_{h'}}{<\mid E_{h'} \mid\mid E_{h-h'} \mid \cos(\phi_{h'} + \phi_{h-h'}) >_{h'}} \tag{1.20}$$

Where $|E|$ is the normalised structure factor amplitude and *hkl* is represented for brevity as *h*. $<\ldots>_{h'}$ means that an average is taken over all values of *h*'. The formula assumes that *h* is the new set of planes whose phase need to be assigned from the 'known' set *h*'.

It is difficult to perform this method on a large macromolecule since there are too many probabilities to handle, but it can be applied to finding heavy atoms by use of Δ_{iso} or Δ_{ano} which give the positions of heavy atoms.

1.5.5.5 Phase calculation from heavy atoms

Once the heavy atom sites have been determined, the phases for the protein molecule can be determined followed by calculation of the electron density map. In the isomorphous replacement method, the phases of the protein molecule are obtained by using phasing circles as shown in Figure 1.17. After the heavy atoms sites have been determined, the phases and the amplitudes of F_H can be calculated whilst the amplitudes of F_{PH} and F_P are already known. To determine the phases of F_P and F_{PH}, the F_H is drawn firstly from the origin on an Argand diagram. Then a circle of radius $|F_P|$ is drawn centred at the origin. Another circle is drawn of radius $|F_{PH}|$ with its centre at the head of F_H. The two circles intersect at two points representing possible solutions of the unknown protein phase in the *SIR* method (Figure 1.17a). This ambiguity is solved by the introduction of more heavy atom derivatives and this is referred to as the *MIR* method. In the *MIR* method (Figure 1.17b), all the derivatives are treated in the same way as just described for the *SIR* method. Different derivatives will have different intersection points but one of them should be the same for all derivatives within a reasonable margin of error. The phase ambiguity is solved by applying this and the phase problem is finally solved.

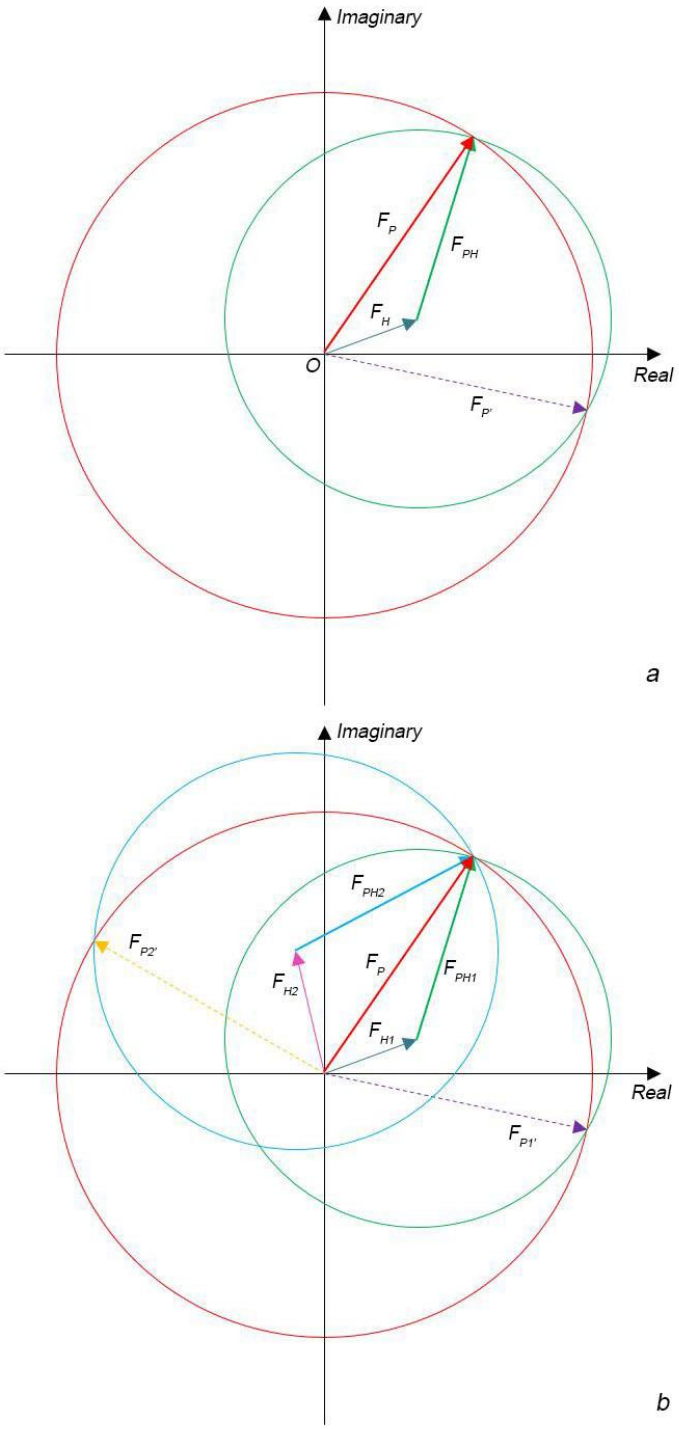


Figure 1.17 Harker construction for *SIR* and *MIR*. a) In the *SIR* method, the red circle is drawn centred at the origin with a radius of $|F_P|$. A second circle (green) of radius $|F_{PH}|$ is drawn centred at the head of F_H. The two circles intersect at two points which gives two possible phases for vectors F_P and $F_{P'}$. b) In the *MIR* method with two derivatives, the same circles (red and blue) for the second derivative can be drawn and both circles intersect at one point which solves the phase ambiguity.

In single isomorphous replacement with anomalous scattering (*SIRAS*), the phase ambiguity can be solved by the introduction of anomalous information. Similar to the *SIR* method, a circle of radius $|F_P|$ can be drawn with its centre at the origin. Circles of radius $|F_{PH}^{+}|$ and $|F_{PH}^{-}|$ are then drawn centred at head of $+F_H''$ and $-F_H''$. Only one of the F_P estimates is consistent in these two diagrams as well as in the *SIR* one, thus breaking the ambiguity.

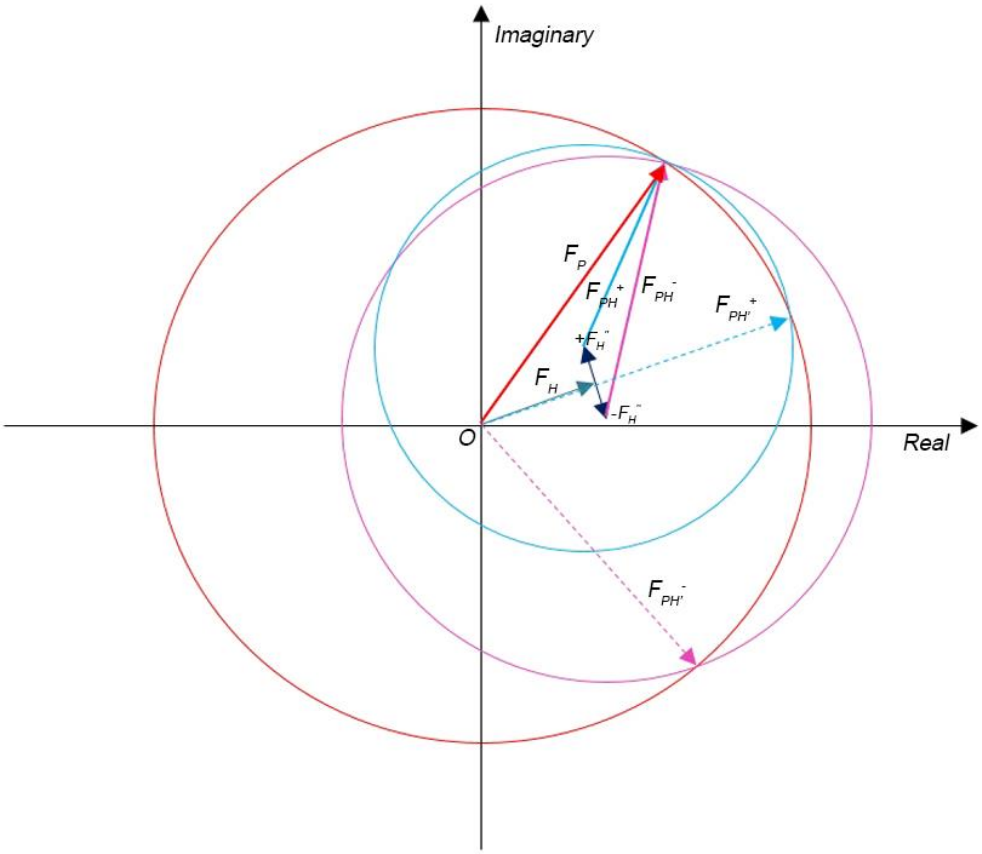


Figure 1.18 Harker construction for *SIRAS*. The phase possibilities without anomalous effects is represented by the circle centred at the origin with a radius of $|F_P|$. Anomalous effects are represented by $+F_H''$ and $-F_H''$ which give rise to two vectors with circles of radii $|F_{PH}^{+}|$ and $|F_{PH}^{-}|$. Each of them produces two phase possibilities but only the phase represented by vector F_P is consistent when considering all the effects.

In a *SAD* experiment, two circles (Figure 1.19) can be drawn according to the method described in anomalous scattering. They interact at two points, giving two

possibilities of the phases of F_P. Because F_P is unknown in terms of both phase and amplitude, this ambiguity cannot be broken at this stage, but instead it may be solved by density-modification (mentioned briefly below).

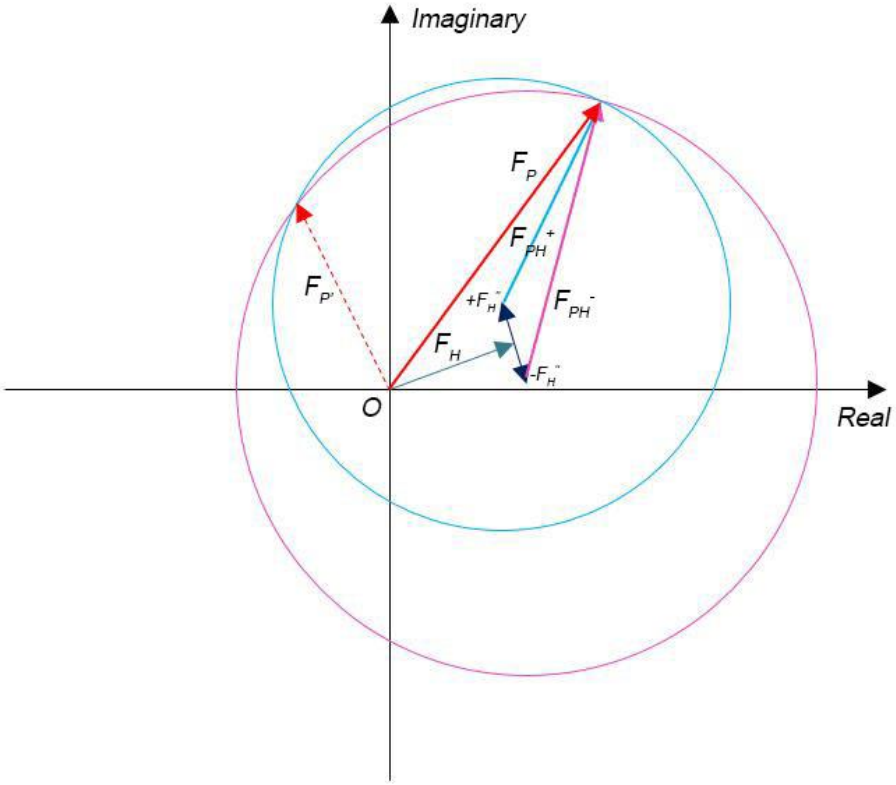


Figure 1.19 Harker construction for *SAD*. Two circles are drawn according to the method described in anomalous scattering, leaving two possible phases for F_{HP} indicated by vectors F_P and F_P'. This ambiguity needs to be solved later by density-modification.

In a *MAD* experiment, changes in the heavy atom structure factor are induced by different X-ray energy levels (wavelengths), so it is treated as the combination of different *SAD* experiments. However, the data from different wavelengths can be treated as 'derivative' structure factors or, in other words, as a 'pesudo-*MIR*' with one of the data sets being chosen as the 'native'.

The initial map from *SIR*, *MIR*, *MAD* and *SAD* will contain noise due to errors in the phases, such as negative electron density, strong electron density in solvent regions and differences between the electron density for NCS related copies. Density-modification recognises and reduces the noise and this should improve promising phases and potentially turn good phases into very good ones. However, this approach cannot turn bad phases into good ones. By applying this method, the phase ambiguity which arises in *SIR* or *SAD* may be resolved and/or an improved electron density map may be obtained.

The experimental phasing quality is commonly indicated by the mean Figure of Merit (*FOM*, equation 1.21), which will be higher given a better solution. However, it can be artificially high if the rms lack-of-closure error is underestimated. Other parameters include the phasing power (Ph.P, equation 1.22) and the Cullis *R*-factor (equation 1.23) (Cullis *et al.*, 1961). Ph.P values of 0.5-1.0, >1.5, and >2.0 indicate usable, very good and excellent quality. The Cullis *R*-factor should always be significantly less than 90%, values below 70% indicate very good phasing quality (Sherwood and Cooper, 2011, Ch. 14).

$$m = \frac{\sum_i P(\phi_i)\cos(\phi_{best} - \phi_i)}{\sum_i P(\phi_i)} \qquad (1.21)$$

Where ϕ_{best} is the centroid phase (Blow and Crick, 1959) which minimises the rms error.

$$Ph.P = \frac{<|F_H(calc)|>}{<\varepsilon(\phi)>} \qquad (1.22)$$

Where <…> means the rms value. ε_ϕ is the lack-of-closure error for ϕ.

$$R_{Cullis} = \frac{\sum |(|F_{PH}| - |F_P|) - F_H(calc)|}{\sum (|F_{PH}| - |F_P|)} \quad (1.23)$$

1.5.5.6 Phasing with maximum likelihood

As mentioned in section 1.5.4.2, the likelihood method measures the consistency between the model and the observations. This can also be used in *SAD*. Given the calculated heavy atom structure factors F_H^+ and F_H^-, this method describes a probability distribution P_{SAD} (which is also a Gaussian distribution) of the model structure factor amplitudes $|F^+|$ and $|F^-|$ and finds the best match (McCoy *et al.*, 2004).

$$P_{SAD} = P(|F^+|, |F^-|; F_H^+, F_H^-) \quad (1.24)$$

There are many highly automated programs and online services that can be used to carry out experimental phasing including programs *SHELX* (Sheldrick, 2008) and *Autosol* (Terwilliger *et al.*, 2009) as well as online services *SHELX* (Sheldrick, 2010; Skubák and Pannu, 2013a) and *CRANK*2 (Skubák and Pannu, 2013a).

1.6 Model building and refinement

1.6.1 Obtaining the trial structure model

With the phased structure factors obtained, an electron density map can be generated from equation 1.7:

$$\rho(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l F_{hkl} e^{-2\pi i(hx+ky+lz)} \quad (1.7)$$

It reveals the features of the real structure but is also far away from accurately representing these features. This is because the structure factors contain errors

raising from different sources, e.g. experimental errors and errors from the difference between the search model and the target molecule. A molecular model can be fitted into this raw electron density map either manually or automatically by the use of various computer programs. In the case of a molecular replacement solution, since an initial model with the correct orientation and position has already been obtained, rebuilding is usually performed with graphics programs such as *Coot* (Emsley and Cowtan, 2004; Emsley *et al.*, 2010). When experimental phasing has been performed, it is better to run auto-building programs such as *Buccaneer* (Cowtan, 2006) or *AutoBuild* (Terwilliger *et al.*, 2008b). The model obtained at this stage often does not represent the true structure accurately so several rounds of rebuilding and refinement are usually needed.

1.6.2 Refinement

The purpose of refinement is to improve the phases as well as the interpretation of the electron density map. This is achieved statistically by fitting the atomic model better with the diffraction data. Fourier transforms allow the refinement to move forward and backward between real and reciprocal space because the model is in real space while the data is in reciprocal space.

1.6.2.1 Least-squares refinement

In general, the least-squares method minimises the sum of the squares of the errors between observed and calculated data. In protein crystallographic refinement, this is specifically to minimise the sum of the weighted squares of the differences between the observed and the calculated structure factor amplitudes. This can be expressed as:

$$\phi = \sum w(|F_o| - |F_c|)^2 \qquad (1.25)$$

Where ϕ is the sum of squared differences between the observed and the calculated structure factor amplitudes and w_{hkl} is the weighting factor which is an estimate of the precision of the measured intensities.

A new set of F_c values is calculated after each cycle of refinement and the shifts in the atomic parameters are computed until there is no significant shift compared with the previous cycle if the function has reached its minimum where the refinement converges. The refinable parameters can include the atomic coordinates (x, y, z), the occupancy and the temperature factor (*B*-factor) for each atom. Least-squares calculations require the data to parameter ratio to be at least 2:1, and this ratio can be increased by introducing known rules of stereochemistry for amino acids. Constraints are rigid mathematical rules that can fix the values derived from the parameters (e.g. bond length) of the model during refinements, others (e.g. dihedral angles) are allowed to vary. Restrained refinement allows the derived values (e.g. bond lengths and bond angles) to vary within a certain range based on additional information. NCS, if present, can also be used as constraints or restraints. TLS (translation, libration and screw) refinement approximates anisotropy with much fewer parameters. Simulated annealing, which is performed by computationally heating and slowly cooling the structure, can be used to cross barriers between minima.

1.6.2.2 Refinement based on maximum likelihood

The maximum likelihood method optimises the model such that the probability of having obtained the experimental dataset is maximised. Errors in the search model arise from Gaussian errors in the atomic positions, which give rise to Gaussian errors in the amplitudes and phases of the atomic structure factor contributions. The total structure factor is obtained by summing up the atomic structure factor contributions and their errors and the resulting distribution is a two-dimensional Gaussian centred on DF_c with a variance of σ_Δ^2. The probability distribution can be expressed as function 1.26. Since the phase of the observed structure factor is not known, the probability functions need to be converted to 'integrate out' the phase information (known as a nuisance variable). Removal of this nuisance variable gives a Rice distribution (Read, 1990; Sim, 1959) and the probability can be described by the Rice function 1.27.

$$P(F_o;F_c) = \frac{1}{\pi\sigma_\Delta^2}\exp\left(-\frac{|F_o - DF_c|^2}{\sigma_\Delta^2}\right) \quad (1.26)$$

$$P = \frac{2|F_o|}{\sigma_\Delta^2}\exp\left(-\frac{|F_o|^2 + D^2|F_c|^2}{\sigma_\Delta^2}\right) I_0\left(\frac{2|F_o|D|F_c|}{\sigma_\Delta^2}\right) \quad (1.27)$$

Where D is the Luzzati weighting factor which is the fraction of the calculated structure factor that is correlated to the true structure factor. I_0 is the modified Bessel function of order [for details of these parameters see (McCoy *et al.*, 2004)].Constraints and restraints described in section 1.5.6.2.1 can also be used in refinement based on this method.

The quality of refinement is assessed by an 'R_{factor}' which measures the difference between each calculated structure factor amplitude and the observed one. It is defined by the following equation:

$$R = \frac{\sum ||F_o| - |F_c||}{\sum |F_o|} \quad (1.28)$$

The R_{factor} may not be the best indicator of convergence since the model can be over-fitted. Therefore another parameter 'R_{free}' is introduced to avoid the over-fitting problem. This value is generally calculated in the same way as the R_{factor} but from 5% of the reflections that have been excluded from the refinement. The R_{free} value is often slightly higher than the R_{factor} value but they should drop in a similar way during a refinement.

Refinement based on maximum likelihood is used most often these days, such as with the computer programs *Refmac*5 (Murshudov *et al.*, 2011) in CCP4 (Winn *et al.*, 2011) and *Phenix.refine* (Afonine *et al.*, 2012).

1.7 Structure validation and deposition

It is always important to validate the structure after refinement to avoid any improper stereochemistry. Parameters such as bond lengths, bond angles, van der Waals contacts and torsion angles need to be monitored carefully. The Ramachandran plot (Fig 1.20) allows the visualisation of the main chain torsion angles and any unfavourable regions or outliers can be identified immediately. Many programs or online services can be used to validate refined structures including *MolProbity* (Chen *et al.*, 2010), *PDB_REDO* (Joosten *et al.*, 2014) and *ProCheck* (Laskowski *et al.*, 1993). Note that glycine is more flexible while proline is more rigid compared with other amino acid residues. Side chains may adopt different conformations so they should be checked visually. Loop regions may be flexible giving them poor electron density which makes it difficult to interpret the map, and in addition they may adopt different conformations. The *B*-factor is a

good indicator since the disordered regions tend to have high *B*-factors while the more ordered buried regions have low values. High *B*-factors in a buried region almost certainly indicates errors.

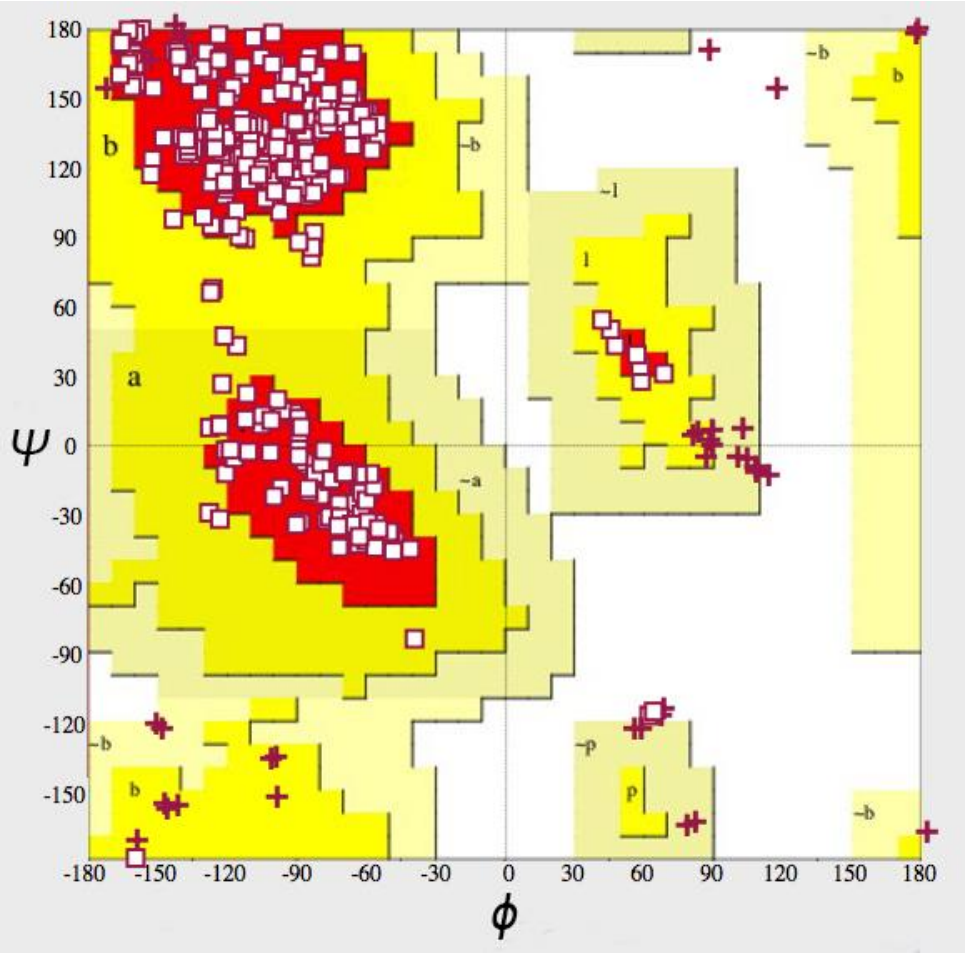


Figure 1.20 A Ramachandran plot. The red, yellow and light yellow regions represent the favoured, allowed, and 'generously allowed' regions. Glycine residues are indicated by crosses while others are indicated by boxes. (Figure generated by using 'The Structure Analysis and Verification Server' from University of California, Los Angeles).

The structures can finally be deposited in the PDB and will be validated again before release; the validation report is required by most journals.

Chapter 2

Structure based drug discovery and structural studies of the Southampton norovirus 3CL protease

2.1 Introduction

2.1.1 Norovirus and gastroenteritis

Gastroenteritis is inflammation of the gastrointestinal tract involving the stomach and small intestine and causes diarrhoea and vomiting. It is a common cause of morbidity and mortality among people of all ages but particularly in young children. For example, it accounts for the death of 2,195 children every day worldwide which makes it the second leading cause of death among children under the age of 5, more than the combination of AIDS, malaria and measles (Liu *et al.*, 2012). Gastroenteritis can be traced back to the 1930s, when Zahorsky first described a 'winter vomiting disease' or 'hyperemesis hemis' that is characterised by the sudden outbreak of vomiting and diarrhoea which peaks in winter (Zahorsky, 1929). Almost about 40 years later in 1972, through immune electron microscopic (IEM) examination of stools of volunteers treated with a faecal filtrates of students affected by gastroenteritis during an outbreak in Norwalk, Ohio in 1968 (Kapikian *et al.*, 1972), it was identified that this syndrome was caused by 'Norwalk virus'. It is now known that many things can cause gastroenteritis including parasites, bacteria and viruses, however, human caliciviruses have been recognised as the leading cause of gastroenteritis worldwide among people of all ages. The *Caliciviridae* family contains five genera know as norovirus, vesivirus, nebovirus, sapovirus and lagovirus (Clarke *et al.*, 2012). While among all the members in this family, norovirus is the most common cause of disease in human (Lambden *et al.*, 1993).

The clinical symptoms of norovirus infection include vomiting, fever, headache, watery diarrhoea, abdominal cramps, chills, myalgia and the presence of mucus in stools which often appear after a 12-48 hours incubation period. A sudden

onset of vomiting may occur on the first day of the infection which, as was suggested by Meeroff *et al*. (1980), is because of a delay in gastric emptying. The most frequent symptom of norovirus infection is diarrhoea and virus shedding in faeces begins 18 hours after infection (Atmar *et al.*, 2008). The mechanism of this is not fully understood but it may be attributed to dysfunction of the epithelial barrier during infection which was observed in *ex vivo* culture of biopsies from infected humans (Troeger *et al.*, 2009). Different genotypes (see section 2.1.3 for details) of norovirus may be different in pathogenic strains, e.g., GII.4 which is more often associated with vomiting, a longer duration of diarrhoea and a greater number of diarrhoea stools compared with others (Desai *et al.*, 2012; Friesema *et al.*, 2009; Huhti *et al.*, 2011).

Several features make noroviruses highly transmissible. 1) The infectious dose is low; as few as 10 viral particles are enough to cause infection in some cases (Teunis *et al.*, 2008). 2) The viral shedding can last for years which increases the risk of secondary spread. 3) The viruses are genetically diverse which limits cross-protection and long-term immunity. 4) Also they have high stability against nearly all active elements in cleaning products such as chlorine, detergent and alcohol, and are resistant to a wide range of temperatures from freezing to 60 °C. Generally, faecal-oral transmission is the most important route. Transmission happens through stools, infectious vomit, contaminated environmental surfaces and aerosolisation in closed settings such as hospitals, cruise ships, hotels and day-care centres (Widdowson, 2005). Another important mode is foodborne transmission, e.g. seafood near coastal areas may be contaminated by faecal discharge (Le Guyader *et al.*, 2012) or food products may be contaminated by infected personnel during handling. Foodborne transmission has the potential to

transmit viruses worldwide, given the globalisation of the food chain and thus increases the chance of viral infection.

While norovirus infections are self-limiting in healthy adults, they can be much more severe in young children, elderly people and those who have impaired immunity. Patients with impaired immune function may be chronically infected which, in some cases, lasts for longer than a year (Siebenga *et al.*, 2008). Chronic infections have been reported in people of all ages who are undergoing chemotherapy, infected with HIV and those who have inherited immune deficiencies (Green, 2014; Siebenga *et al.*, 2008). In younger children with congenital immunodeficiencies, noroviruses are frequently detected and the infections usually last for more than 9.5 months (Frange *et al.*, 2012). The most frequently reported norovirus infection-associated deaths is pneumonia and other causes of death include gastroenteritis, acute gastrointestinal bleeding, malnutrition, cardiac complications, necrotizing enterocolitis, sepsis and colon perforation (Trivedi *et al.*, 2013). In countries with limited resources, gastroenteritis is an important cause of morbidity and mortality.

Noroviruses account for more than 50% of gastroenteritis and at least 90% of nonbacterial acute gastroenteritis worldwide, as reported by the Centers for Disease Control and Prevention in the US (2011). Studies have shown that in the United States alone, noroviruses cause on average 19-21 million acute gastroenteritis cases, 1.7-1.9 million outpatient visits, 400,000 emergency department visits, 56,000-71,000 hospitalisations and 570-800 deaths annually. Scallan *et al*. (2011) estimated that 99% of all viral foodborne illness incidents are caused by noroviruses which corresponds to 5.5 million per year in the US. Also from 2009 to 2013, around 62.5% of norovirus cases needed long-term care

facilities in order to control the transmission (Vega *et al.*, 2014). Statistics are generally similar in Europe (Baert *et al.*, 2009; Phillips *et al.*, 2010). Globally, noroviruses lead to a total of $4.2 billion in direct health system costs and $60.3 billion in social cost per year (Bartsch *et al.*, 2016).

Clinical treatments and interventions have been hampered because no licensed vaccine or antiviral is available at the moment. Treatment with human immunoglobulin did show some benefit, but did not result in clearance of the virus (Florescu *et al.*, 2008). Despite the fact that discovery of a vaccine is hindered by the lack of small-animal models and cell culture systems, the first norovirus vaccine have now passed phase I and II clinical trials. Intramuscular vaccination did reduce the incidence and severity of vomiting and diarrhoea, however, the incidence of protocol-defined illness was not significantly reduced (Bernstein *et al.*, 2015). Treatment with a small antimicrobial compound nitazoxanide resolved the acute gastroenteritis in a patient, but asymptomatic shedding was still observed (Siddiq *et al.*, 2011). Oral treatment with another compound ribavirin gave both successful and unsuccessful cases (Woodward *et al.*, 2015). Intervention with natural human interferon-*α* in gnotobiotic pigs reduced virus shedding which was later return to normal level after the termination of the intervention (Jung *et al.*, 2012).

2.1.2 Norovirus classification

Noroviruses cannot be classified according to their serotypes since they cannot be cultured *in vitro*, except for the murine strains. For this reason, they are genetically classified into 7 established genogroups, known as GI to GVII (Figure 2.1), based on amino acid sequence in the complete VP1 capsid protein. They are further segregated into at least 40 genotypes which have 9, 21 or 22, 3, 2, 2,

2 and 1 members in genogroups GI, GII, GIII, GIV, GV, GVI and GVII, respectively (Vinjé, 2015). Noroviruses from groups GI and GII infect humans except for GII.11, GII.18 and GII.19, which infect pigs. Members of GIV.1 subgroup are also involved in human diseases, while GIV.2 have been detected in feline and canine. GII viruses are the most frequently detected (89%) with GII.4 are the major cause of norovirus outbreaks worldwide (Siebenga *et al.*, 2009).

Many noroviruses have been reported such as Norwalk virus (Jiang *et al.*, 1993), Hawaii virus (Lew *et al.*, 1994a), Snow Mountain virus (Lochridge and Hardy, 2003), Desert Shield virus (Lew *et al.*, 1994b), Southampton virus (Clarke and Lambden, 1997) and Lordsdale virus (Lambden *et al.*, 1993).

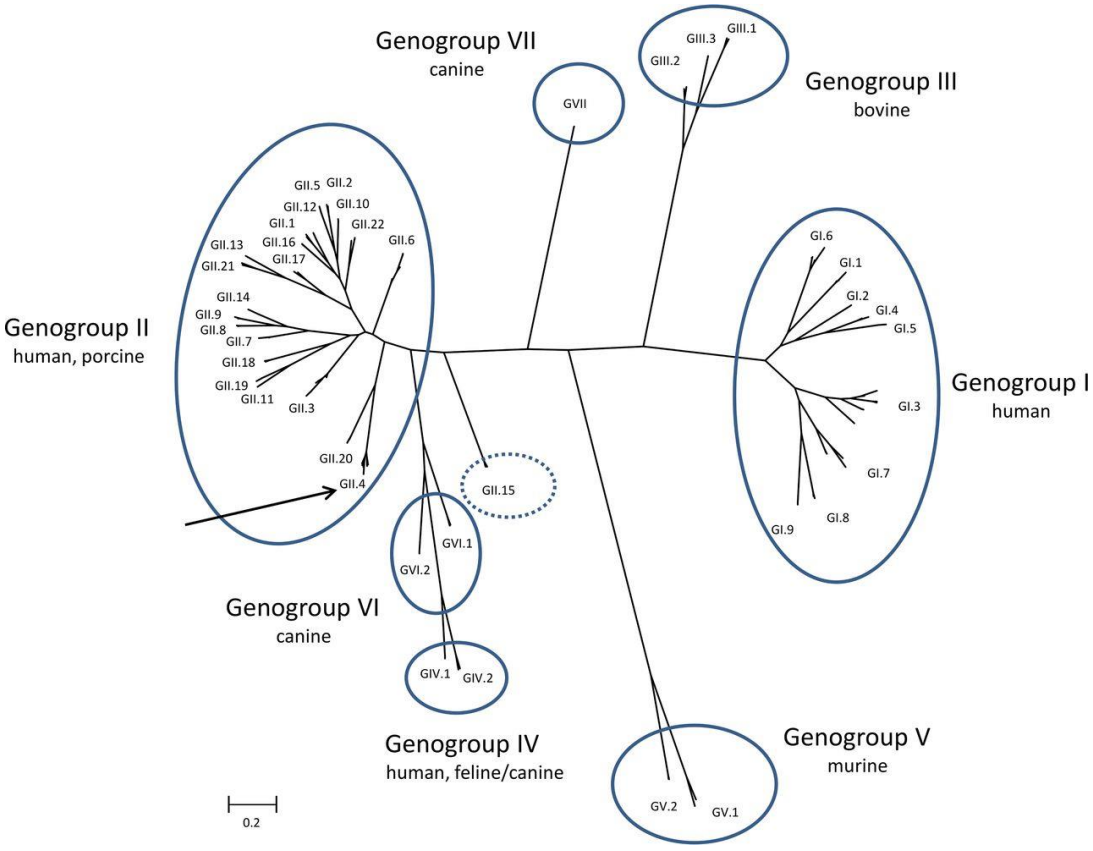


Figure 2.1 Genogroups and genotypes of noroviruses. The arrow reflects GII.4 viruses which constitute the majority of norovirus infections all over the world. GII.15 may be reclassified into a separate group but further approval is required. The scale bar indicates the number of amino acid substitutions per site. [Figure from (Vinjé, 2015)].

2.1.3 Susceptibility and resistance to norovirus infection

Since it is not possible to cultivate human noroviruses and there has been no animal model until now, most of the immunology and pathogenesis data are from human volunteer studies (Johnson *et al.*, 1990; Parrino *et al.*, 1977). Wyatt and co-workers (1974) indicated that individuals recovering from norovirus infection showed a 6-14 weeks resistance to further infections, suggesting the development of at least partial protection against noroviruses. However, other conflicting studies showed that, in some cases, people possessing higher levels of antibody had a higher susceptibility to norovirus infection compared with those with lower levels or no antibody (Parrino *et al.*, 1977), which indicated the infection is complicated. Interestingly, in their long term norovirus infection resistance experiments, 50% of the volunteers never became infected in the initial and further tests, while the others who displayed symptoms initially were re-infected after re-exposed to noroviruses 27-42 months later. This suggested that some individuals possess an innate resistance to norovirus infection but the reason was not clear at that time.

Such individuals appeared in familial clusters (Koopman *et al.*, 1982) and recent studies revealed that noroviruses bind to specific human *histo*-blood group antigen (HBGA) receptors in the gut of hosts and the infection is correlated with the host genotype (Lindesmith *et al.*, 2003). HBGAs are complex carbohydrates expressed on the surface of specific cells. They participate in cell to cell interactions, self and non-self recognition, as well as the binding of viral particles to host cell surfaces which cause viral internalisation and infection (Hutson *et al.*, 2004). They are determinants of the ABO blood group and Lewis blood group system (Marionneau *et al.*, 2001). H HBGAs, the precursors of HBGAs, are

generated by the reactions catalysed by α(1,2)-fucosyltransferase 2 (FUT2) and are then converted to A and B HBGAs. The VP1 of different noroviruses bind specifically to different HBGAs, resulting in different susceptibilities of humans to specific strains of norovirus (Lindesmith *et al.*, 2003; Thorven *et al.*, 2005). Some people who do not have A, B and H HBGAs in their body, because of the lack of FUT2, are known as non-secretors. In contrast, those who do possess a *FUT2* gene are secretors (Le Pendu *et al.*, 2006). About 20% of Northern Europeans are non-secretors (Thorven *et al.*, 2005). The ABO blood group system is independent of secretor status because it is FUT1 rather than FUT2 which does HBGAs synthesis on the surfaces of erythrocytes (Shirato *et al.*, 2008). However, individuals with blood type A or O have shown higher susceptibilities to norovirus infection compared with the other two blood types (Hutson *et al.*, 2002; Thorven *et al.*, 2005).

A human challenge study in healthy adults showed that 70% of the secretors were infected while only 6% of the non-secretors were infected and displayed minimal disease (Frenck *et al.*, 2012), confirming that HBGAs have a major effect on the susceptibility of individuals to norovirus infections. The host-specificity may explain why some volunteers with high levels of antibody had an increased chance of developing illness. Also, the variable host susceptibility observed in volunteer studies as well as in norovirus outbreaks may be explained by the varying expression of the HBGA receptors and the strain-specific binding.

2.1.4 Genome structure

The norovirus genome (Figure 2.2) consists of a single-stranded positive-sense RNA of 7.5-7.7 kb in length and contains three open reading frames (ORFs) (Lambden *et al.*, 1993), except for the murine norovirus which has a fourth

alternative ORF (McFadden *et al.*, 2011). ORF1 encodes a 200 kDa non-structural polyprotein which is co- and post-translationally cleaved into six or seven non-structural proteins by the viral 3C-like protease (NS6). The products of the proteolysis are shown in Figure 2.2, from N-terminus to C-terminus: a p48 protein (NS1-2), an NTPase (NS3), a 3A-like protein (p22, NS4), a viral genome-linked protein (VPg, NS5), a 3C-like protease ($3CL^{pro}$, NS6) and an RNA-dependent RNA polymerase (RdRp, NS7) (Blakeney *et al.*, 2003). ORF2 and ORF3 encode the capsid protein VP1 and the minor structural protein VP2, respectively.

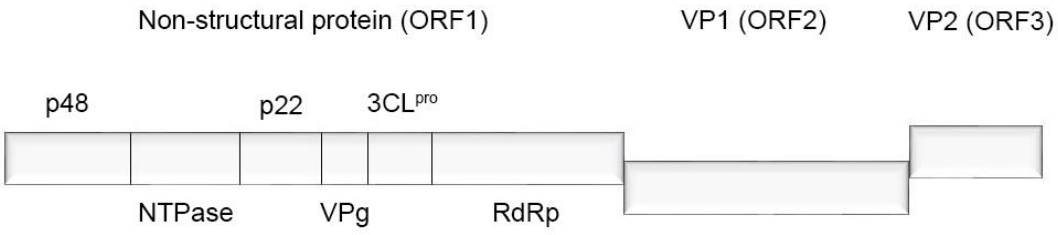


Figure 2.2 Diagrammatic representation of norovirus genome structure.

2.1.5 The 3C-like protease

Noroviruses encode a single protease named 3C-like protease ($3CL^{pro}$) because of its similarity to the picornavirus 3C protease. As mentioned before, norovirus $3CL^{pro}$ cleaves the polyprotein and generates several non-structural proteins including the $3CL^{pro}$ itself. The protein is a cysteine protease which shows a typical chymotrypsin-like fold containing two domains: a *β*-barrel domain and a *β*-sheet domain separated by a groove where the active site is located (Bazan and Fletterick, 1988; Boniotti *et al.*, 1994). The active site is characterised by a catalytic dyad (Cys139-His30) (Someya *et al.*, 2002) or triad (Cys139-His30-Glu54) (Tiew *et al.*, 2011) and shows a strong preference for a –D/E-F/Y-X-L-Q-G-P- (X can be H, Q or E) sequence corresponding to the subsites S_5-S_4-S_3-S_2-

S_1-S_1'-S_2' (Tiew *et al.*, 2011). Studies have indicated that norovirus 3CL proteases have a preferential order of processing the polyprotein, for example, the Southampton virus $3CL^{pro}$ has a preference for cleavage at LQ-GP and LQ-GK, but it can also cleave at ME-GK, FE-AP and LE-GG (Hussey *et al.*, 2011). Although several structures of norovirus 3CL proteases have been determined (Hussey *et al.*, 2011; Nakamura *et al.*, 2005; Zeitler *et al.*, 2006), the structural basis of how these enzymes recognise different sites with different affinities is still unknown. The key role of norovirus $3CL^{pro}$ in the processing of the polyprotein and its requirement for viral replication make it an excellent target for antiviral drug discovery.

2.1.6 Viral structure

The first 3-dimensional structure of norovirus (Norwalk virus) capsid (VP1) was determined by cryo-electron microscopy in 1994 (Prasad *et al.*, 1994) and X-ray crystallography in 1999 (Figure 2.3a) (PDB ID: 1IHM) (Prasad *et al*., 1999). The capsid is composed of a single structural protein of approximately 58 kDa (Figure 2.3b). The viruses are 405 Å in diameter and show *T*=3 icosahedral symmetry. The 3-dimensional structure has 32 large surface hollows (90 Å in width and 50 Å in depth) surrounded by 90 distinctive arch-like capsomeres, each of which is formed by a dimer of the capsid protein VP1. VP1 is composed of the shell (S) domain and the protrusion (P) domain which is further divided into the P1 and P2 sub-domains. The S domain is related to the formation of viral structure and is highly conserved. The P1 sub-domain is moderately flexible and links the S domain and P2 sub-domain together. The P2 sub-domain is highly variable and is located on the surface. It has been suggested that this sub-domain contains specific antigenic determinants, i.e. receptor- and neutralising antibody-binding

sites, and is involved in the up-regulation and stabilisation of VP1 in norovirus (Bertolotti-Ciarlet *et al.*, 2003). The high stability of noroviruses to acidic conditions, such as in the stomach, may be explained by the lack of a capsid envelope. Capsid envelopes are generally degraded in the acidic conditions in the stomach so that those rely on a capsid envelope cannot survive in stomach (Maillard, 2001).

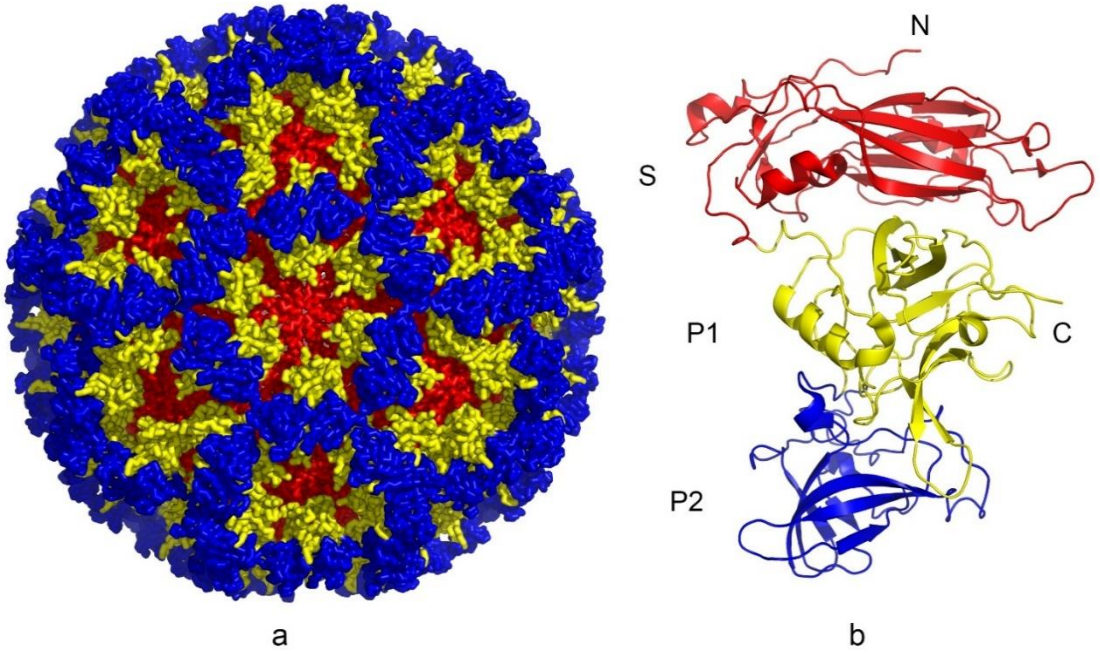


Figure 2.3 Crystal structure of the Norwalk virus capsid. a) Overall X-ray structure of the capsid containing 180 molecules of VP1. b) Structure of the VP1 subunit with the S, P1 and P2 domains showing in red, yellow and blue, respectively.

2.1.7 The VP2 protein

The function of the VP2 protein is not well-understood. Based on its basic nature and internal location in the virion, VP2 is predicted to be involved in RNA binding and genome packaging into progeny virions (Vongpunsawad *et al.*, 2013). In addition, it is suggested to contribute to virion and the VP1 stabilities (Bertolotti-

Ciarlet *et al.*, 2003; Sosnovtsev *et al.*, 2005). Also, it negatively regulates the viral RdRp (Subba-Reddy *et al.*, 2011).

2.1.8 Viral proteases as targets for antiviral drug discovery

Therapeutic agents targeting viral replication had not developed much for a few decades since the first one targeting on vaccinia virus was identified almost 70 years ago (Hamre *et al.*, 1950). With a better understanding of the viral life cycle in the early 1990's, many key enzymes have been identified to play vital roles in viral replication including viral proteases. Most viruses which infect humans encode at least one viral protease (Jaeger *et al.*, 2012) which is responsible for the processing of viral non-structural polyproteins in order to produce functional viral proteins. Viral proteases have attracted great interest as therapeutic targets and many compounds targeting these proteases have been approved for clinical use (Courter *et al.*, 2008; Kim *et al.*, 2012a).

2.1.8.1 HIV protease inhibitors

HIV protease belongs to the aspartic protease family and is involved in the processing of the Pol polyprotein to yield 3 non-structural proteins and in the processing of the Gag polyprotein to yield 4 structural proteins (Kohl *et al.*, 1988). Saquinavir was the first approved HIV protease inhibitor in 1995, which is built on a peptidomimetic scaffold that has a non-hydrolysable amide bond between residues that are at the S_1 and S_1' recognition sites of the protease (Martin, 1992). Several inhibitors based on the same mechanism have been approved since then including ritonavir, indinavir, nelfinavir, amprenavir, lopinavir, atazanavir, fosamprenavir and darunavir. Other inhibitors, such as tipranavir, have a coumarin scaffold instead (De Clercq, 2013).

2.1.8.2 Hepatitis C virus NS3-4A protease inhibitors

Unlike the noroviral and HIV proteases, processing of the polyprotein in hepatitis C viruses (HCVs) relies on several proteases: the signal peptide protease (Hüssy *et al.*, 1996), the NS2/3 (Grakoui *et al.*, 1993) protease and the NS3-4A protease (Bartenschlager *et al.*, 1994). Many efficient HCV NS3-4A protease inhibitors are now on the market including asunaprevir, paritaprevir, simeprevir, vaniprevir and grazoprevir. All the approved NS3-4A protease inhibitors are used for the treatment of HCV genotype 1 infections, which is the most prevalent genotype (Messina *et al.*, 2015). In addition to the approved ones, other NS3-4A protease inhibitors, such as danoprevir, faldaprevir, vedroprevir, sovaprevir and narlaprevir are under clinical development (De Clercq, 2014, 2015).

2.1.8.3 Norovirus $3CL^{pro}$ inhibitors

There is currently no approved norovirus $3CL^{pro}$ inhibitors available but several products have been reported to have inhibitory activity against Norovirus 3CL proteases *in vitro* such as transition state (TS) mimics (Mandadapu *et al.*, 2013a), peptidyl TS inhibitors (Mandadapu *et al.*, 2012) and macrocyclic inhibitors (Damalanka *et al.*, 2017).

These transition state related peptidyl inhibitors were designed based on a structure incorporated with a recognition element (a peptidyl fragment) and a warhead or transition state mimic (Kankanamalage *et al.*, 2015). Peptidyl aldehydes (Figure 2.4 GC373) and *α*-ketoamides (Figure 2.4 GC375) are TS inhibitors which showed great inhibition against not only norovirus $3CL^{pro}$, but also the 3C or 3C-like proteases in picornaviruses and coronaviruses in cell-based assays (Kim *et al.*, 2012b). The aldehydes and *α*-ketoamides act as warheads

which form a reversible adduct with the catalytic residue Cys139 in the active site. The crystal structure of norovirus $3CL^{pro}$ complexed with an aldehyde bisulphite adduct (Figure 2.4 GC376) showed that the bisulphite group seemed to be removed and the compound was converted to the aldehyde form that is covalently linked to Cys139 (Kim *et al.*, 2012b). These compounds are named as latent TS inhibitors. TS mimics, such as *α*-hydroxyphosphonate, are converted to the aldehyde form either with or without catalytic action of the enzyme and form a tetrahedral adduct with the Cys139 residue (Kankanamalage *et al.*, 2015).

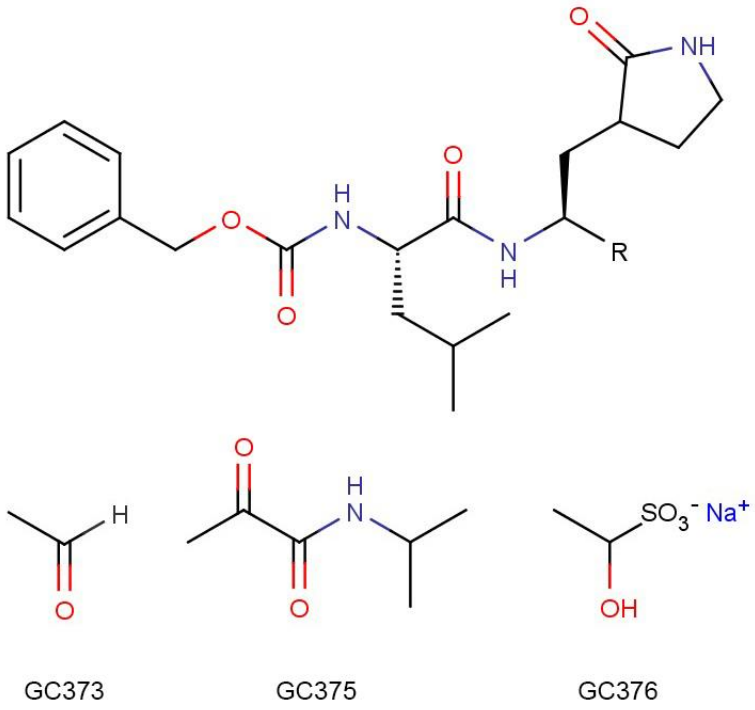


Figure 2.4 Structures of norovirus $3CL^{pro}$ dipeptidyl inhibitors. The dipeptidyl fragment is shown on top. The warhead (R) is replaced by an aldehyde, an *α*-ketoamide or an aldehyde bisulphite adduct sodium salt for inhibitors GC373, GC375 and GC376, respectively.

Macrocyclic inhibitors against 3C or 3CL proteases from different viruses have also been reported (Damalanka *et al.*, 2017; Mandadapu *et al.*, 2013b). The basic theory was to design macrocyclic molecules (Figure 2.5) that are able to 1) maintain the favourable interactions which have been identified from the known

protein-inhibitor complex structures of noroviral 3C proteases and 2) have a flexible diversity site that can form hydrogen bonds and hydrophobic interactions with the S_3 and S_4 sub-sites. As for the inhibitors mentioned above, each of these also possessed an aldehyde or an aldehyde bisulphite adduct (*α*-ketoamides diminished activity) which forms covalent bond with the Cys139 residue. However, macrocyclic molecules have their own advantages including higher pharmacological activity and selectivity, improved permeability and stability (Bhat *et al.*, 2015; Marsault and Peterson, 2011).

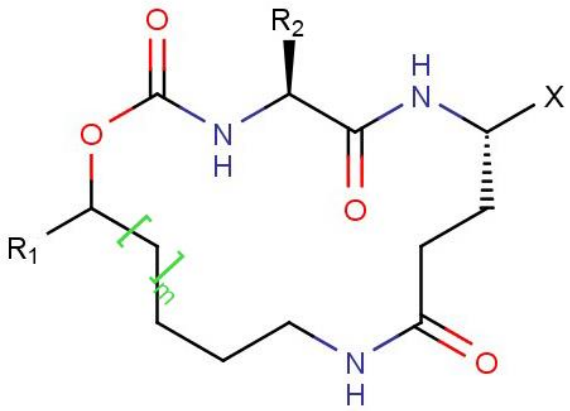


Figure 2.5 General structure of norovirus $3CL^{pro}$ macrocyclic inhibitors. R_1 and R_2 groups are replaced by different chemical and amino acid structures, respectively. The letter m indicates different numbers of $-CH_2-$ group in the square brackets. X is the warhead that stands for an aldehyde or an aldehyde bisulphite adduct. [Figure generated based on (Damalanka *et al.*, 2017)].

Hussey *et al*. (2011) reported an X-ray complex structure of the Southampton norovirus 3CL protease (SV3CP) with an inhibitor (Figure 2.6) composed of part of the preferable substrate (EFQLQ) and a Michael acceptor moiety which is linked to the S_1 residue Gln. This bond is attacked by the enzyme and a covalently bound complex is formed. Interestingly, His30 is pushed away by the inhibitor to some distance which disrupts the catalytic triad.

Finally, it is possible to develop broad spectrum inhibitors targeting 3C or 3CL proteases since the substrate specificity of many of these enzymes is quite similar (Kim *et al.*, 2012b).

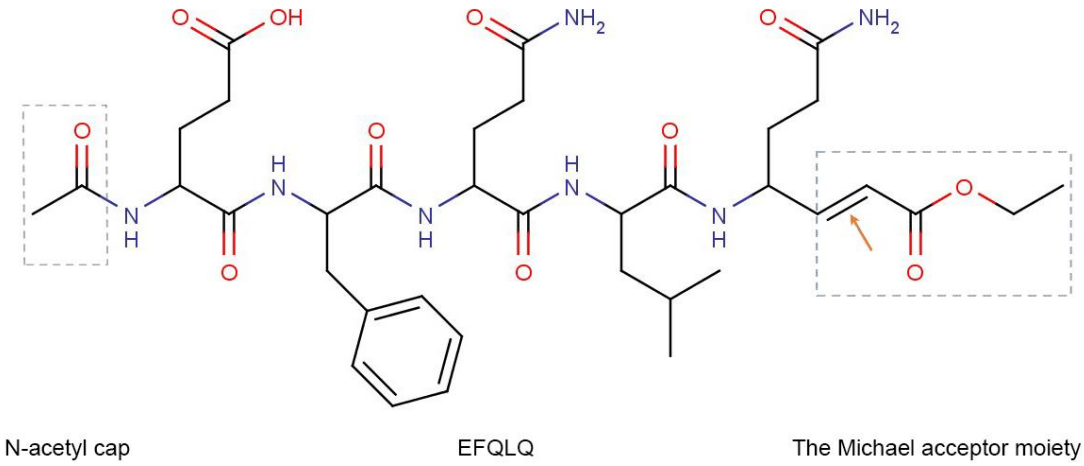


Figure 2.6 Structure of the Michael acceptor peptidyl inhibitor (MAPI) for SV3CP, designed by Hussey *et al*. The arrow indicates the nucleophilic attack site by the Cys139 residue.

2.2 Project aim

1. To express, purify and crystallise SV3CP and determine its ligand-free structure for crystal-based fragment screening. To find out the differences between the ligand-free and the ligand-bound (PDB ID: 2IPH) structures and whether there is any domain movement.

2. To perform crystal-based fragment screening with both covalent and non-covalent fragment libraries which will allow fragments that bind in the active site to be identified and to find ways to develop these hits into good inhibitors.

3. Since the 2IPH complex structure has only part of the substrate polypeptide bound in the active site of SV3CP, to try to find out how the other part binds in the active site.

2.3 Methods

2.3.1 Mutagenesis, expression, purification and crystallisation

Mutagenesis of SV3CP, in which the catalytic residue Cys139 was mutated to Ala, was carried out by a project student Nat Tan. The presence of the mutation was confirmed by gene sequencing and expression of the mutated enzyme was verified by Western blot.

The same methods were used for DNA transformation, protein expression and purification for both the wild-type (WT) and the mutated proteins (C139A). For each protein, DNA transformation was undertaken according to the method described in Method i of the appendices, protein expression was performed following Method ii of the appendices using the standard method without heat shock. Protein purification was achieved firstly by using an SP Sepharose (GE Healthcare, Buckinghamshire, UK) cation-exchange column (binding buffer: 10 mM Na_2HPO_4, 5 mM *β*-mercaptoethanol, pH 7.5, eluted with a gradient up to 1 M NaCl in binding buffer), followed by use of a Superdex 75 (GE Healthcare, Buckinghamshire, UK) gel-filtration column (buffer: 100 mM NaCl in cation-exchange binding buffer).

Screening for crystallisation conditions for both the WT and C139A proteins was accomplished using the sitting-drop method at 21 °C with the screening kits Structure Screen 1 & 2, JCSG-*plus*, PACT *premier*, MIDAS and Morpheus from Molecular Dimensions (Suffolk, UK) . A TTP Labtech Mosquito crystal screening robot (TTP Labtech, Hertfordshire, UK) was used to dispense 400 nl of each protein, at 5 mg/ml along with 10 mg/ml, plus 400 nl of the corresponding well solution into each drop. Two different crystal forms for each protein (Figure 2.7)

appeared in many conditions in different kits and further optimisation revealed the optimal conditions (table 2.1).

In order to obtain a complex structure of SV3CP and the substrate (DEFQLQGK), co-crystallisations were carried out with the C139A protein and the substrate (1×, 5×, 10× and 20× in molar excess) with both crystal forms.

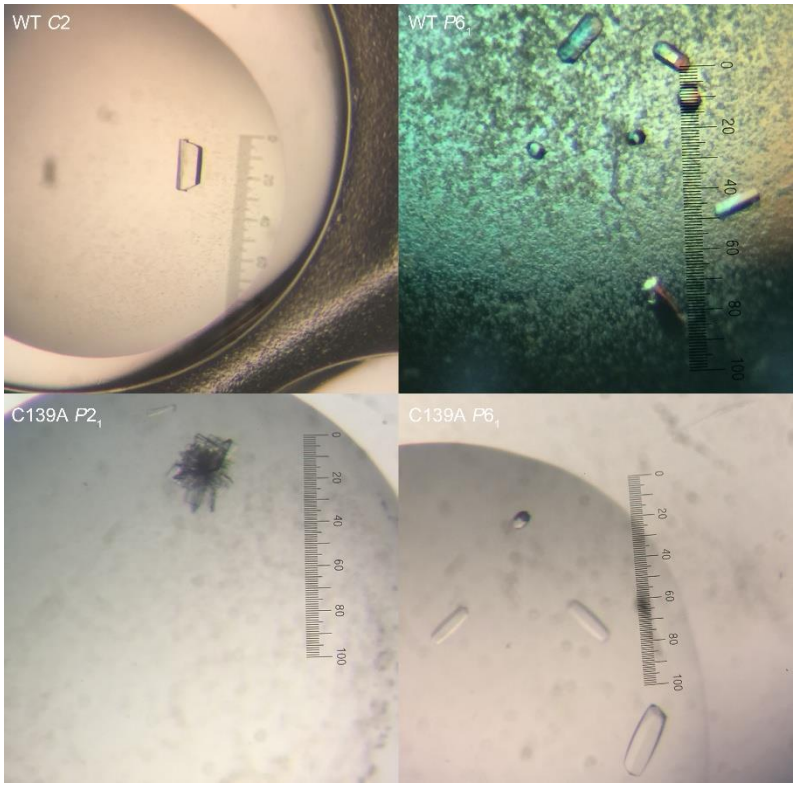


Figure 2.7 Different crystal forms of SV3CP WT and C139A proteins. One small unit on the ruler is 10 microns.

Table 2.1 Optimal crystallisation conditions for the WT and C139A SV3CPs.

Protein	Crystal form	Conc. (mg/ml)	Crystallisation condition
WT	*C*2	4	0.2 M Ammonium citrate dibasic, 12% (w/v) PEG 3,350
	$P6_1$	5	Morpheus C6*
C139A	$P2_1$	5	30% (w/v) PEG 1,500
	$P6_1$	5	1.6 M Ammonium sulphate, 0.1 M MES pH 6.5, 10% (v/v) 1,4-Dioxane

* Morpheus C6: for details see Gorrec (2009).

2.3.2 Data collection, data processing and structure determination

Selected crystals were cryo-protected in 30% glycerol and mounted in loops before flash-cooling. X-ray data were collected at beamlines I03 and I04-1 at Diamond Light Source (DLS, Didcot, England). Fine-sliced data were collected as guided by the strategies suggested by the program *EDNA* (Incardona *et al.*, 2009). Data were processed automatically by the program x*ia*2 (Winter, 2010) at DLS, which revealed the space groups of the crystals, as shown in table 2.1. No twinning, tNCS or significant anisotropy was spotted by the analysis using *Phenix.xtriage* (Zwart *et al.*, 2005). Solvent content (Table 2.2) for each crystal was estimated using *Matthews_coef* (Kantardjieff and Rupp, 2003).

The *C*2 crystal structure of the WT protein was determined by use of the program *Phaser MR* (McCoy *et al.*, 2007) using the protein structure from the published SV3CP-MAPI complex (PDB ID: 2IPH) as a search model. This refined structure was then used as the search model for the determination of the remaining other

three structures. Several rounds of manual rebuilding and correction were performed using *Coot* (Emsley and Cowtan, 2004) followed by restrained refinement using *Refmac*5 (Murshudov *et al.*, 2011) in the CCP4 (Winn *et al.*, 2011) program suite. Since the *C*2 crystal diffracted to atomic resolution, a final round of anisotropic refinement was conducted using *Refmac*5, while for others, TLS refinement was performed instead. All the structures were validated using *MolProbity* (Chen *et al.*, 2010). All the statistics for data collection, data processing and refinement are shown in table 2.2.

Table 2.2 X-ray statistics for the SV3CP structures. Values in parentheses are for the high resolution shell.

	WT *C*2	WT $P6_1$	C139A $P2_1$	C139A $P6_1$
Beamline	I04-1	I03	I04-1	I04-1
Wavelength (Å)	0.9282	0.9762	0.9282	0.9282
Space group	*C*2	$P6_1$	$P2_1$	$P6_1$
Unit-cell parameters				
a (Å)	63.1	130.2	37.8	130.5
b (Å)	89.4	130.2	36.8	130.5
c (Å)	61.6	120.0	112.7	121.1
α (°)	90.0	90.0	90.0	90.0
β (°)	96.5	90.0	97.9	90.0
γ (°)	90.0	120.0	90.0	120.0
Resolution (Å)	41.25-1.30 (1.32-1.30)	57.21-2.77 (2.82-2.77)	55.80-2.74 (2.75-2.74)	82.6-2.75 (2.80-2.75)
R_{merge} (%)	4.4 (130.5)	13.2 (421.2)	8.3 (54.9)	6.5 (20.0)

R_{meas} (%)	4.8 (143.0)	13.5 (431.4)	9.6 (72.6)	7.0 (224.6)
$CC_{½}$ (%)	100.0 (55.0)	99.9 (54.7)	99.8 (96.5)	99.9 (54.0)
Completeness (%)	99.9 (98.7)	100.0 (100.0)	98.3 (97.9)	100.0 (100.0)
Average $I/\sigma(I)$	13.6 (1.0)	14.8 (0.8)	13.8 (2.9)	14.5 (1.1)
Multiplicity	6.6 (6.1)	20.6 (21.4)	6.4 (6.8)	7.9 (7.9)
No. of observed reflections	544,844	606,221	52,785	239,338
No. of unique reflections	83,130	29,412	8,199	30,491
Wilson plot *B*-factor (Å^2)	16.1	104.3	58.7	96.5
Solvent content (%)	44.9	67.6	38.6	68.0
No. of molecules per ASU	2	4	2	4
R_{factor} (%)	16.4	19.3	19.8	18.7
R_{free} (%)	21.3	24.3	24.4	24.5
RMSD bond lengths (Å)	0.033	0.020	0.143	0.020
RMSD bond angles (°)	2.592	2.292	1.798	2.121
No. of reflections in working set	83,130	29,378	8,338	30,457
No. of reflections in test set	4,243	1,521	429	1,605

Mean protein *B*-factor (Å^2)	20.7	98.1	46.1	71.7

2.3.3 Polymeric status determination

The polymeric status of WT SV3CP protein was determined by gel filtration using a Superdex 75 column and the same buffer was used in purification. SV3CP (4 mg) was loaded on the column together with two marker proteins known as ovalbumin (4 mg, 42.7 kDa) and ribonuclease (4 mg, 13.7 kDa) and the result was checked by SDS-PAGE.

2.3.4 Fragment screening with crystals

2.3.4.1 Crystal preparation

The native *C*2 crystals were selected to perform fragment screening since, on average, they are of good quality and diffract well. Crystals were prepared in Swissci 3-drop crystallisation plates (Hampton Research, CA, USA) in 200 nl drops containing 100 nl of the protein (4mg/ml) and 100 nl of well solution. Since all of the fragments were dissolved in 100% DMSO and some of them were also dissolved in 100% ethylene glycol (EDO, the alternative option in case crystals could not survive in DMSO, but fewer fragments were available), crystal stability in these two solvents was tested in the range (v/v) of 0, 10%, 20%, 30% and 40%, and in time scales of 1h, 3h and overnight. In order to make the experiment more efficient, the crystals were also tested with and without additional cryo-protectant for data collection. It was found that these crystals could survive in up to 40% of both solvents for many hours and additional cryo-protection other than DMSO or EDO was not required.

2.3.4.2 Fragment soaking, crystal harvesting and data collection

The plates containing crystals were scanned using a Desktop Minstrel UV instrument (Rigaku Automation, CA, USA) for crystal imaging. All the crystals were then ranked manually by use of the program *TeXRank* (Ng *et al.*, 2014) which also gave the coordinates of a target position for the fragment to be ejected to. Each fragment from the DSLP library (non-covalent, 776 fragments) (Cox *et al.*, 2016), the Maybridge Ro3 core set (non-covalent, 68 fragments used) (Fisher Scientific UK Ltd, Loughborough, UK) and a covalent library (1000 electrophiles, offered by Dr Nir London, Weizmann Institute of Science, Israel) was ejected into a single drop in series of 2.5 nl droplets by acoustic dispensing using a Labcyte Echo 550 liquid handler (Labcyte Inc, CA, USA) which gave a final concentration of 200 mM and 8 mM for the non-covalent and covalent fragments, respectively, in each drop (Collins *et al.*, 2017). Fragment soaking was conducted in batches to give an average soaking time of approximately 2.5 hours. Crystal fishing (harvest) was manual but significantly efficient with the help of an OLT Crystal Shifter (Oxford Lab Technologies, Oxford, UK). All the crystals were mounted in loops which were about the same size or slightly smaller than the crystals, this allowed for highly automated, unattended data collection aiming for the centre of the loops. Since the crystals belong to the *C*2 space group, 180° of data were collected for each crystal which only took approximately 60 seconds.

2.3.4.3 Fragment screening by LC/MS and co-crystallisation with the covalent library

Since soaking with the covalent library damaged most of the crystals and resulted in poor diffraction, optimisations were undertaken with different soaking time

scales and fragment concentrations. It was finally identified that the crystals could only survive in 8 mM fragments for 15 min or 2 mM fragments for 15 h. Attempts for both conditions with a few fragments did not give any hit.

It was then decided to carry out co-crystallisation experiments with the covalent library. To make the experiment time and cost effective, an initial screening with LC/MS was performed followed by co-crystallisation using the top 100 hits from the LC/MS screening. LC/MS screening was accomplished in collaboration with Dr Nir London from Weizmann Institute of Science, Israel. In brief, each protein sample at a final concentration of 2 μM was incubated with a pool containing 5 electrophiles at a final concentration of 6 μM each at 4 °C for 24 h. The sample was then applied to LC/MS and 200 hits were detected with measurable labelling. The protein was incubated separately with the top 100 hits from LC/MS overnight at 4 °C with protein to fragment ratios of 1:1 and 1:3 (in molar excess). The incubated samples were then subjected to co-crystallisation with and without seeding. We are currently waiting for crystal growth.

2.3.4.4 Data processing, analysis and hit identification

All the 4 datasets were processed automatically by *xia*2 (Winter, 2010) at DLS and the processed data were subjected to *XChemExplorer* (Krojer *et al.*, 2017) and *PanDDA* (Pearce *et al.*, 2016) for further analysis and hit identification. In brief, these two programs perform averaging on approximately 2000 datasets collected during the experiment, revealing clear electron density for only the changed state, given that most datasets do not have a ligand bound and are very similar. Even weak ligands with low occupancy can be identified by applying the analysis. All the hits were checked visually which was facilitated by using the

program *Pandda.inspect* in the *PanDDA* suite (Pearce *et al.*, 2016). The interesting hits were further refined using *Refmac*5 (Murshudov *et al.*, 2011) followed by further inspections using *Coot* (Emsley and Cowtan, 2004). Omit maps for the ligands were generated using the program *Composite omit map* (Terwilliger *et al.*, 2008a) in the *PHENIX* program suite (Adams *et al.*, 2010). Interactions between ligands and SV3CP were analysed using *LigPlot*$^{+}$ (Wallace *et al.*, 1995). All the figures in this thesis were prepared using programs *MarvinSketch* (ChemAxon, 2013), *PyMOL* (The *PyMOL* Molecular Graphic System, Schrödinger, LLC) and *CueMol* (Molecular Visualization Framework http://www.cuemol.org).

2.4 Results

2.4.1 Tertiary structure of SV3CP

Figure 2.8 illustrates the overall structure of SV3CP which is composed of an N-terminal and a C-terminal domain. The N-terminal domain contains an *α*-helix and a twisted 7-stranded antiparallel *β*-sheet forming an incomplete *β*-barrel. A typical N-terminal antiparallel *β*-barrel has been reported in many chymotrypsin-like viral 3C or 3C-like proteases such as FMDV 3C^{pro} (Birtley *et al.*, 2005), PV 3C^{pro} (Mosimann *et al.*, 1997) and coronavirus 3CL^{pro} (Anand *et al.*, 2002). The C-terminal domain is made up of 6 *β*-strands forming an antiparallel *β*-barrel. The catalytic triad C139-H30-E54 in the active site located in a cleft between the N-terminal and the C-terminal domains. The *β*-hairpin formed by *β*9 and *β*10 adopts different conformations in the mutant $P2_1$ and 2IPH structures compared with the native *C*2 structure.

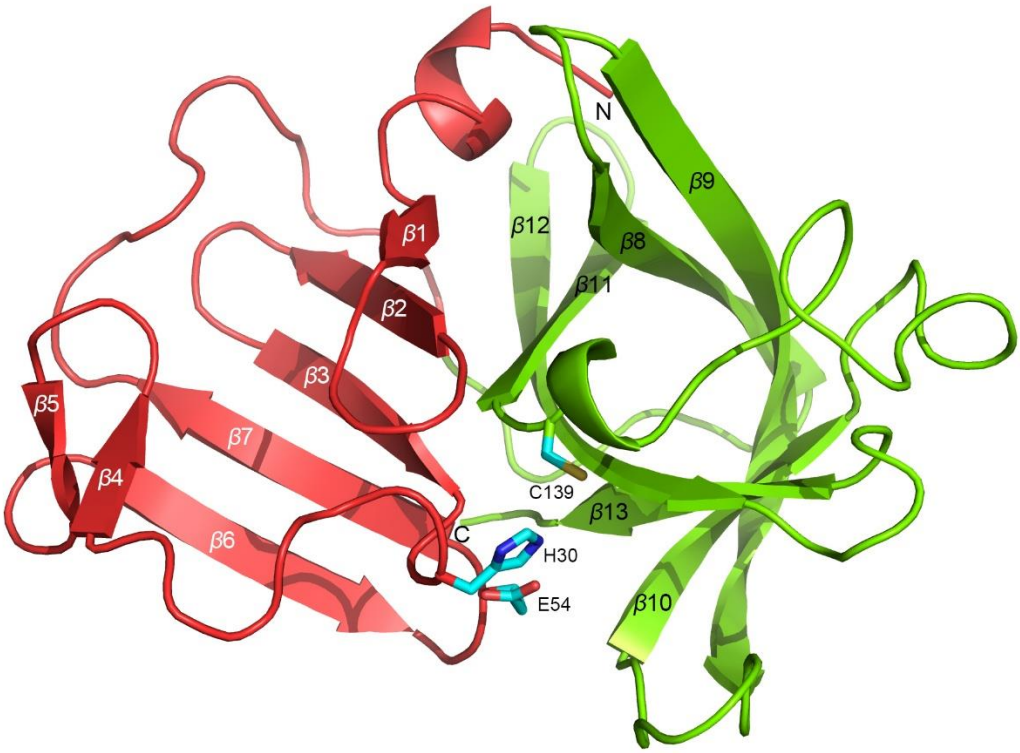


Figure 2.8 The overall structure of SV3CP with all *β*-strands labelled. The protein is composed of an N-terminal domain (red) containing a twisted 5-stranded antiparallel *β*-sheet and a C-terminal domain (green) consisting of a 6-stranded *β*-barrel. The catalytic triad is shown in ball-and-stick.

2.4.2 Polymeric status

Polymeric status of WT SV3CP (19.1 kDa) was determined by gel filtration with two marker proteins known as ovalbumin (42.7 kDa) and ribonuclease (13.7 kDa). Gel-filtration showed only two main peaks (Figure 2.9) with peak 2 being ribonuclease, as shown in Figure 2.10. Peak 1 contained both SV3CP and ovalbumin which were not separated, however, ovalbumin came out earlier than SV3CP, as indicated by the SDS-PAGE, suggesting SV3CP forms dimers in solution. This is consistent with some other noroviral 3C proteases that have been reported (Leen *et al.*, 2012; Zeitler *et al.*, 2006), however, analysis with the *Pisa* website (Krissinel and Henrick, 2007) suggested a tetramer is also stable in

solution. Indeed, the interface area between the A-B and C-D chains is 883.0 Å^2 while it is 692.3 Å^2 between the A-D and B-C chains, indicating more stable complexes formed by the A-B and C-D chains compared with those formed by the A-D and B-C chains (Figure 2.11).

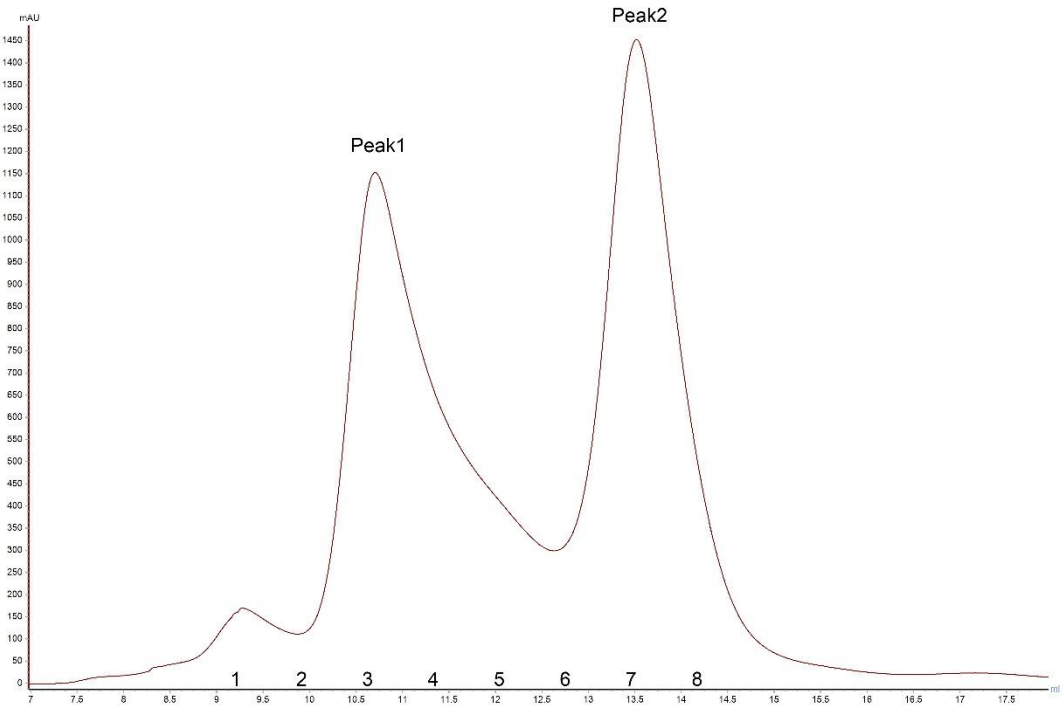


Figure 2.9 Gel filtration for SV3CP. Two proteins known as ovalbumin and ribonuclease were used as markers.

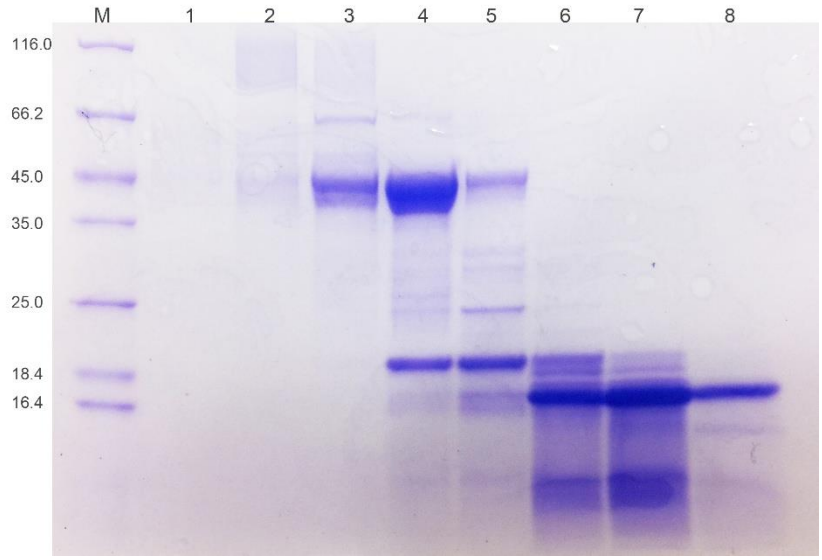


Figure 2.10 SDS-PAGE for the SV3CP gel filtration fractions.

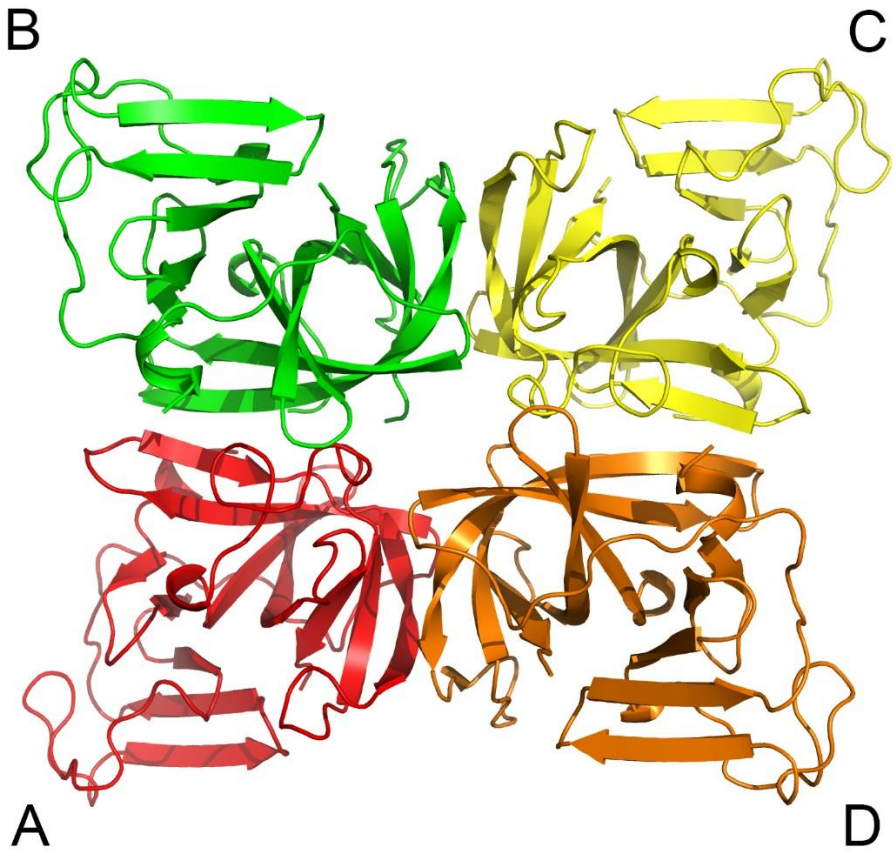


Figure 2.11 Polymeric status of SV3CP. The asymmetric unit (ASU) of the *C*2 WT SV3CP is composed of two dimeric structures shown as AB and CD. Less stable interfaces may form between chains A and D, B and C, as suggested by the *Pisa* website, resulting a possible tetrameric structure which makes up the ASU of the $P6_1$ WT and the C139A SV3CP structures.

2.4.3 Crystal-based fragment screening with non-covalent libraries

Most crystals used in the non-covalent fragment screening experiment diffracted to a resolution ranging from 1.5 to 1.8 Å with good crystallographic statistics, allowing accurate analysis of the results. Screening with the DSPL library and part of the Maybridge Ro3 library identified 19 hits in total, as illustrated in Figure 2.12, which bind in five different sites. The protease active site (site A) is a long groove with the catalytic Cys139 residue locating in the middle. Two fragments (Figure 2.13, J01 and J02) were identified binding at different sides of the Cys139

residue in site A. Five hits (Figure 2.13, J03-J07) have been found to bind in the putative RNA binding site (site B) including one (J07) which also binds in site C. Site C lies in a pocket between chains A and B and the symmetry related chains A' and B', with 11 hits being identified (J07-J17) to bind here. Two other fragments were also spotted that bind at site D (J18) and site E (J19). For details of ligands J08-J19, see Figure A of the appendices. Presence of the ligands were further confirmed by calculating the omit maps, see Figure 2.14 for examples of J01 and J02.

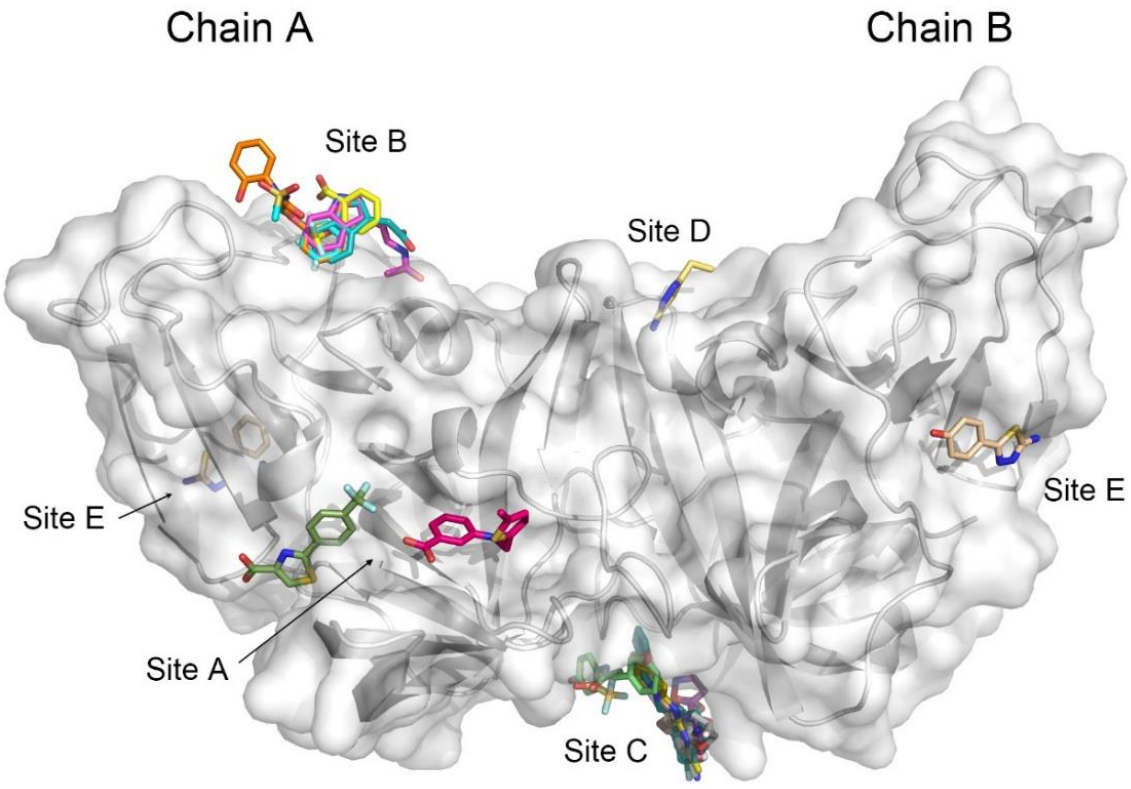


Figure 2.12 Fragment screening hits from the non-covalent libraries. All the ligands bind in five sites labelled as A, B, C, D and E. Site A is the active site of the protease and site B is the putative RNA binding site.

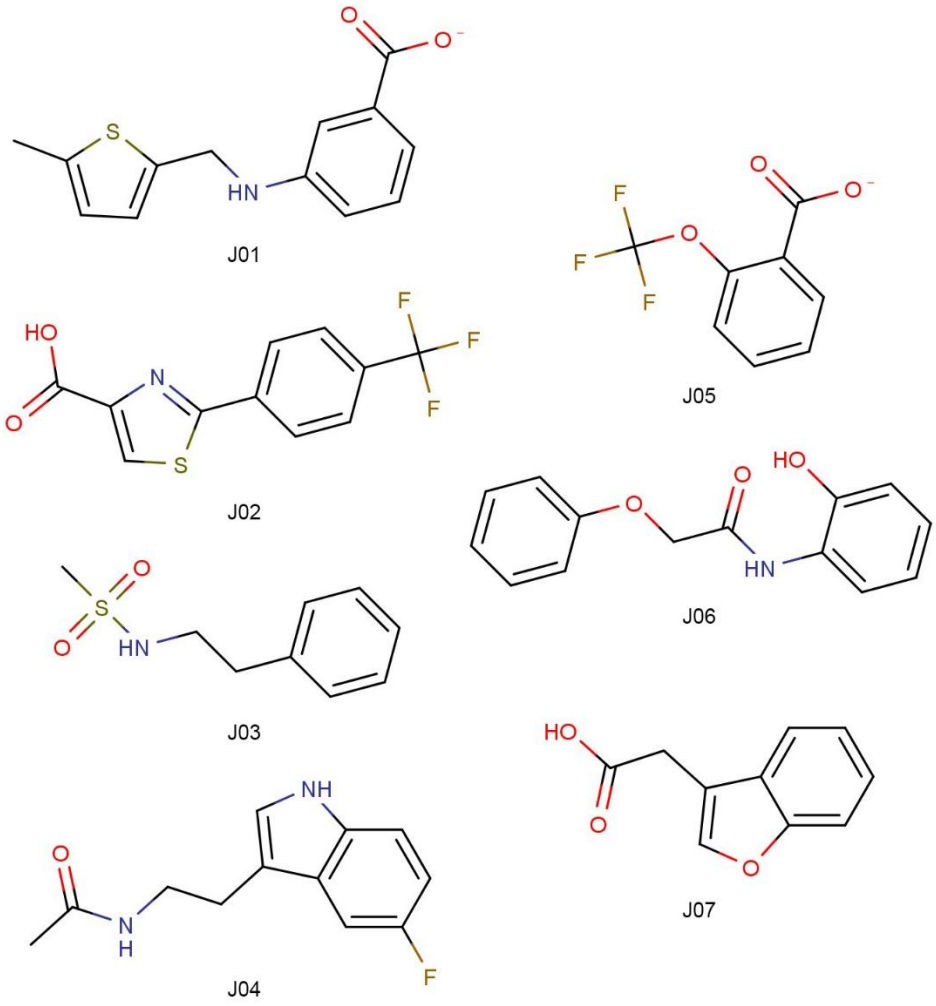


Figure 2.13 2D structure of fragments J01 to J07.

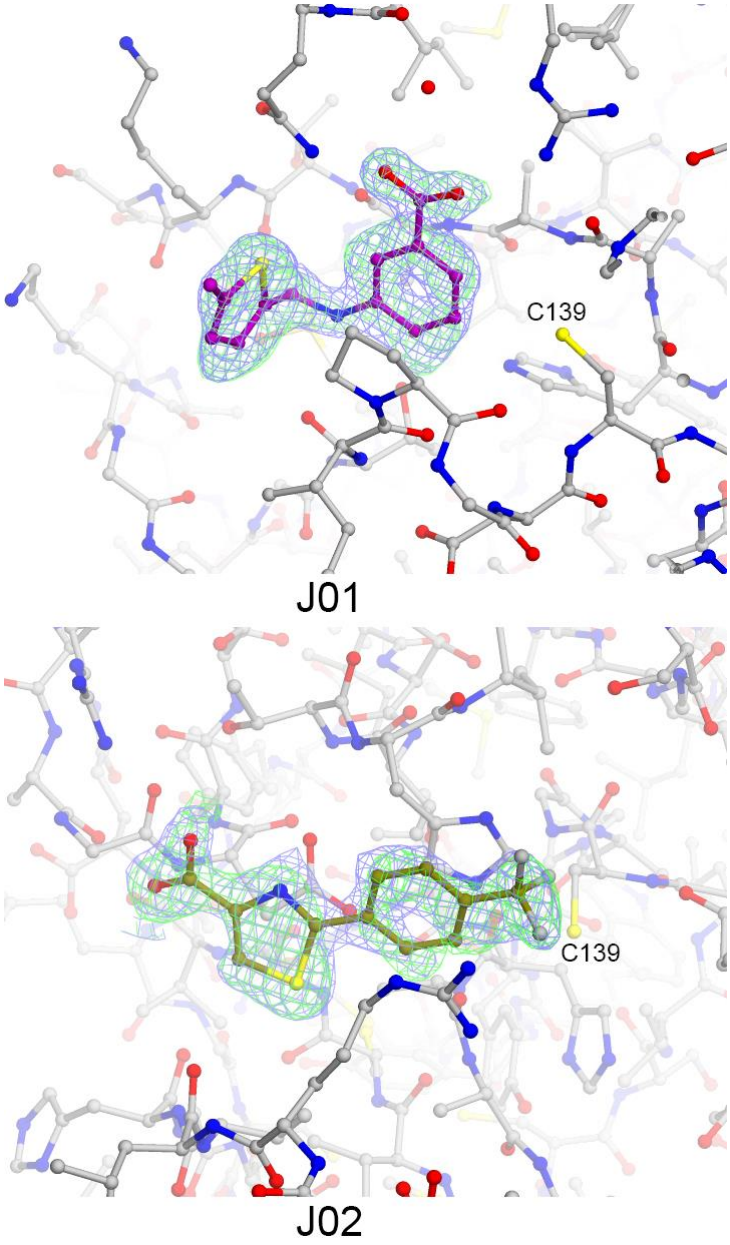


Figure 2.14 $2F_o$-F_c and omit maps for fragments J01 and J02. $2F_o$-F_c map and omit maps are coloured as blue and green, respectively.

2.5 Discussion

2.5.1 The important role of a *β*-hairpin in substrate recognition

The N-terminal domain of ligand-free SV3CP and the ligand-bound enzyme is almost the same. Superimposition of the N-terminal domains by C^{α} atoms shows that most part for the C-terminal domain in these two structures is quite similar

as well, except for a *β*-hairpin (*β*9, *β*10) in the 2IPH structure having moved by 7.5 Å, as indicated by the asterisk in Figure 2.15a. Binding of the MAPI inhibitor has pushed the *β*-hairpin back away from *β*11 and *β*12, which have also moved to some extent. These movements open up the channel A formed by the four *β*-strands, suggesting a conformational change in response to inhibitor binding (Figure 2.15b and 2.15c). Interestingly, it has been reported that the equivalent *β*-hairpin in FMDV 3C protease is crucial for its catalytic activity (Sweeney *et al.*, 2007). Indeed, this *β*-hairpin folds over the active site of the protease and is characterised by significantly higher *B*-factors than the average in the ligand-free structure. In contrast, in the 2IPH structure, the *B*-factors for the residues forming this *β*-hairpin are similar to the other residues and the *β*-hairpin is stabilised by the contacts between the P_2 to P_4 positions of the MAPI and the protease. In the complex structure of human rhinovirus 3C protease (HRV $3C^{pro}$) with a peptidyl inhibitor (PDB ID: 1CQQ), Leu127 in this *β*-hairpin makes hydrophobic contact with the P_2 and P_4 side chains, suggesting that it is very important for substrate binding (Matthews *et al.*, 1999). An Ile residue is positioned at the equivalent site in the SV3CP structure, making it very likely to be a key residue for substrate recognition.

Compared with the ligand-free structure, the binding pocket is 3 Å deeper in the 2IPH structure in superposition. Interestingly, binding of the inhibitor also pushes the His30 residue away from the Cys139 and Glu54 residues, resulting in decomposition of the catalytic triad which further enhances the inhibitory activity. There is no electron density for the last 8 residues (ASEGETTL) at the C-terminal end of the 1.3 Å ligand-free structure, which is possibly due to proteolysis caused by the enzyme itself since the electron density is quite clear in the inactive 2IPH

structure. In this region, the C-terminus of the protein possesses a VQ-AS sequence at the P_2-P_1-P_1'-P_2' positions, making it a less preferable cleavage site which may be cleaved during a relatively long-term proteolysis process, e.g. during crystallisation (Hussey *et al.*, 2011; Kankanamalage *et al.*, 2015). Whist the enzyme is inhibited by the MAPI in the 2IPH structure, leaving these residues un-cleaved. This may help to clarify the ambiguous question of where the enzyme cleaves the polyprotein between SV3CP and the RdRp.

The $P6_1$ C139A protein has approximately the same structure as the ligand-free WT SV3CP protein, whilst in the $P2_1$ mutant protein structure the *β*-hairpin has moved even further back to a distance of 7.8 Å, as shown in Figure 2.15d. This confirms the flexibility of this *β*-hairpin which undergoes conformational changes in order to accommodate the substrate or inhibitor which is not caused by crystal contact. Interestingly, although it has been proved that the $3CL^{pro}$ tend to form dimers or tetramers, the $P2_1$ structure does not seem to form polymers.

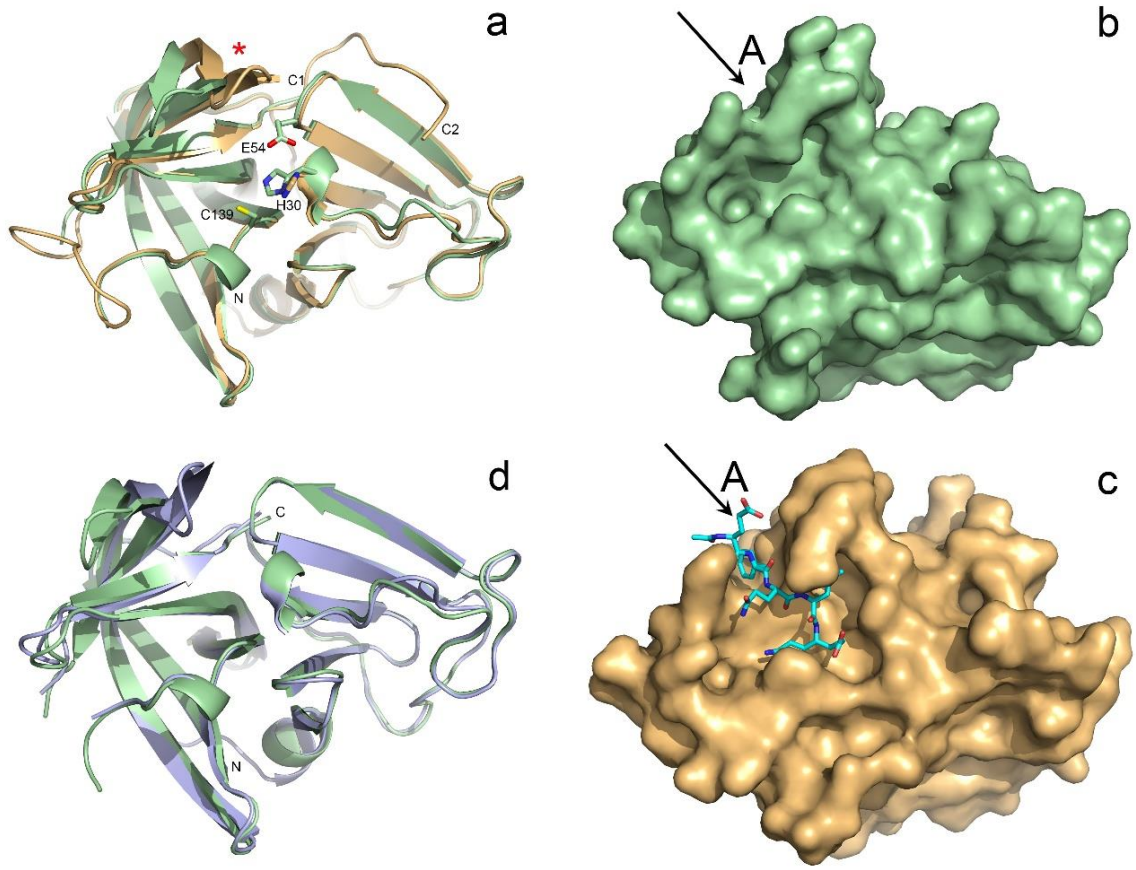


Figure 2.15 Comparison of different SV3CP structures. a) Superimposition for the ligand-free SV3CP and the 2IPH structures. The ligand-free SV3CP structure is coloured as pale-green with its C-terminal end labelling as C1. The 2IPH structure is coloured as light-orange with its C-terminus labelling C2. The asterisk indicates the 'moved' *β*-hairpin. b) The surface representation of ligand-free SV3CP with closed channel A. c) The surface representation of the 2IPH structure, in which channel A is opened in response to inhibitor binding. d) Structural superimposition of WT and $P2_1$ C139A proteins, the WT and C139A structures are coloured as pale-green and light-blue, respectively.

2.5.2 Fragments bind in the protease active site

As far as we know, all the noroviral $3C^{pro}$ inhibitors that have been reported are covalent inhibitors. In this experiment, two non-covalent fragments were identified that bind in the active site of the protease named as J01 and J02. Fragment J01 binds in one part of the long active site and forms several hydrogen bonds,

through the carboxylic group, with the side chains of the neighbouring residues Gln110, Ile109 and Arg112 including those mediated by a water molecule HOH457 (Figure 2.16a, b). These residues reside at the tip of the functionally important *β*-hairpin that is involved in substrate recognition. The ligand –NH group (N1) is also within hydrogen bonding distance of the hydroxylic group of the Thr134 residue. In addition, J01 forms many hydrophobic interactions with the active site residues including Cys139, especially through the 5-methyl-2-thienyl group. However, this group does not go deep into the binding pocket which is possibly occupied by a DMSO molecule, probably due to the higher affinity and smaller size of the DMSO molecule. The *β*-hairpin possesses the same conformation as the ligand-free SV3CP which seems to be held in place by the fragment. By holding the structure of the *β*-hairpin tightly by the fragment, the enzyme may not be able to adopt the 'open' conformation in order to accommodate the substrate, thus the activity is inhibited.

J02 binds in the other part of the long active site on the other side of the Cys139 residue. It does not seem to form any hydrogen bonds with the active site residues, as shown in Figure 2.16c and 2.16d. Instead, the benzoic acid ring is involved in a π-π interaction with the side chain of the His30 from the catalytic triad, and also forms a cation-π interaction with the Arg122 residue from the *β*-hairpin. In addition, several hydrophobic interactions are formed between the fragment and the adjacent residues in the active site including Glu54 from the triad, and Arg112, Leu113 and Val114 which are all from the *β*-hairpin. The same as J01, J02 stabilises the *β*-hairpin through the interactions which makes it a good candidate for developing inhibitors.

In conclusion, J01 and J02 bind at different sides of the Cys139 in the long active site and both of them interact with the surrounding residues including those from the functionally important *β*-hairpin. J01 forms hydrophobic interaction with catalytic Cys139 while J02 forms hydrophobic and π-π interactions with Glu54 and His30, which are also from the catalytic triad. Both J01 and J02 could potentially be developed into SV3CP inhibitors, however, a better ligand may be obtained by linking them together, given that the distance between the closest two atoms in them is 4.3 Å.

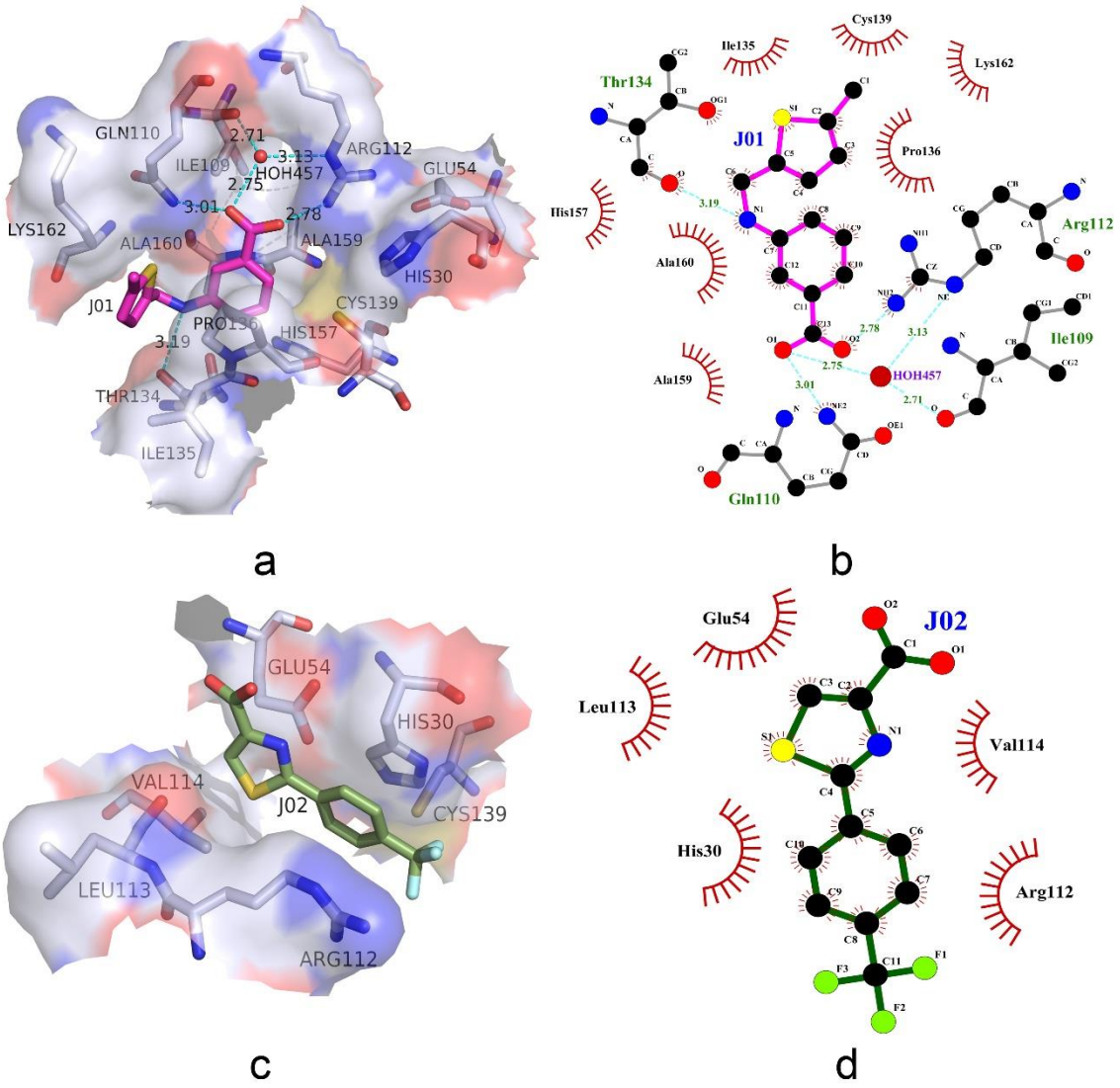


Figure 2.16 Interactions between SV3CP and the fragments J01 and J02. 3D (a, c) and 2D (b, d) representations of the interactions between fragments J01 (a, b), J02 (c, d) and the active site residues. The binding pocket is shown by a surface representation, fragments J01 and J02 are coloured as purple and dark green, respectively. Hydrogen bonds are indicated by dashed lines in cyan. Hydrophobic interactions are indicated by eyebrow-like icons coloured as brown.

2.5.3 Fragments bind in the putative RNA binding site

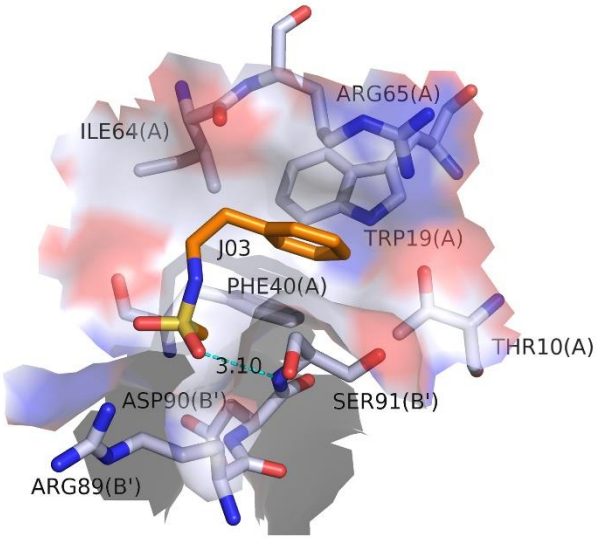
In addition to the protease activity, studies on viral 3C proteases from different organisms suggested that 3C proteases or their larger precursors can bind specifically to the 5'-terminal nucleotides of the viral RNA (Leong *et al.*, 1993a;

Nayak *et al.*, 2006). The interaction occurs only on the plus strand which forms a ribonucleoprotein (RNP) complex that is necessary for the initiation of the plus strand synthesis (Andino *et al.*, 1990). Viswanathan *et al*. (2013) suggested that SV3CP also has the ability to bind short RNAs and binding of RNA non-competitively inhibited the 'protease activity' with an IC_{50} of 3-5 μM (Viswanathan *et al.*, 2013). The RNA binding site has been studied by mutagenesis in other homologous 3C proteases, in which they identified a key arginine residue and a conserved sequence, KF/VRDI (F/V represents the amino acid could be an F or a V), which are required for the interaction with RNA (Bergmann *et al.*, 1997; Leong *et al.*, 1993b; Nayak *et al.*, 2006). Structural comparison of SV3CP with HRV $3C^{pro}$ (PDB ID: 5FX5) (Kawatkar *et al.*, 2016) and FMDV $3C^{pro}$ (PDB ID: 2J92) (Nayak *et al.*, 2006) identified the site around Arg65 to be the putative RNA binding site in SV3CP, which possesses a KIRPDL sequence that has similar properties with KF/VRDI. This putative RNA binding site for SV3CP is shown in Figure 2.13 (site B), which does not look like a typical RNA binding groove, however, the same trend is seen in the FMDV $3C^{pro}$ and only a shallow tiny groove is found in the HRV $3C^{pro}$. In addition, these sites are in crystal contact areas and form deep channels with the neighbouring symmetry-related molecules in HRV, FMDV and SV3CP $3C^{pro}$ or $3CL^{pro}$. Inhibitors binding in the RNA binding site have the potential to inhibit both the RNA binding ability and the protease activity, which will give excellent antiviral activity.

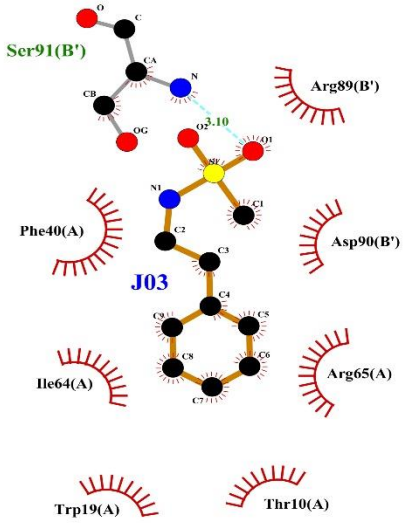
Interactions between J03-J06 and the residues in the putative RNA binding site are shown in Figure 2.17. All the fragments form hydrophobic contacts with Arg65 and some other residues from the KIRPDL sequence, which may confirm the importance of this sequence. While J03 and J06 are mainly involved in

hydrophobic interactions, J04 and J05 also form many hydrogen bonds with the neighbouring residues, possibly making them stronger binders. The carbonyl group (O1) of J04 is involved in three hydrogen bonds formed, directly or mediated by a water molecule, with Thr10, Lys11 and Ser91 (from the symmetry related molecule). The N1 atom forms two hydrogen bonds with Ser7 and Pro3 (also from the symmetry mate) with the participation of a water molecule. A weak hydrogen bond is also seen between the fluorine and the NE1 atom on the side chain of Trp19. J05 contributes to two hydrogen bonds, which are also mediated by a water molecule, formed with the symmetry related Ser91 and Glu93. Unlike the active site fragments which bind in different parts of the active site, these four fragments bind in the same position with their 'heads' overlapping at one place and their 'tails' pointing toward different directions. As it has been reported that binding of RNA also inhibits the protease activity of SV3CP (Viswanathan *et al.*, 2013), further analysis in solution are required to assess the binding ability and then inhibitory activity of these fragments which bind in an area that is involved in crystal contact in the complex structures obtained from the screening.

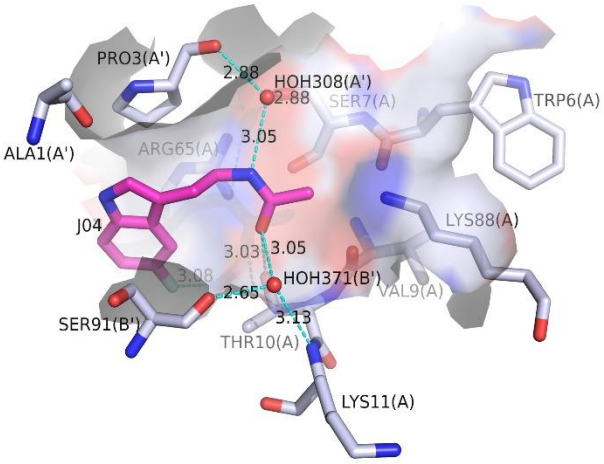
Fragment J07 was found to bind in both the putative RNA binding site (B) and site C. Similarly, the main forces between the fragment and the neighbouring side chains are hydrogen bonds and hydrophobic interactions which are illustrated below in Figure 2.18.


ARG65(A)
ILE64(A)
TRP19(A)
J03
PHE40(A)
THR10(A)
3.10
ASP90(B')
SER91(B')
ARG89(B')

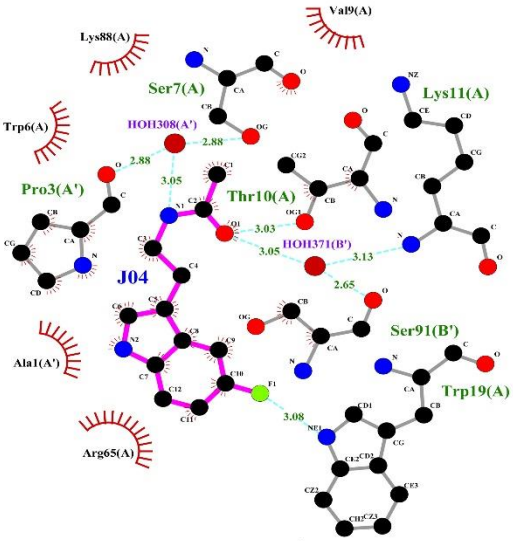
a


Ser91(B')
Arg89(B')
3.10
Phe40(A)
Asp90(B')
J03
Ile64(A)
Arg65(A)
Trp19(A)
Thr10(A)

b


PRO3(A')
2.88
HOH308(A')
2.88
SER7(A)
TRP6(A)
ALA1(A')
ARG65(A)
3.05
J04
LYS88(A)
3.03
3.05
3.08
2.65
HOH371(B')
VAL9(A)
SER91(B')
3.13
THR10(A)
LYS11(A)

c


Val9(A)
Lys88(A)
Ser7(A)
Lys11(A)
Trp6(A)
HOH308(A')
2.88
2.88
Pro3(A')
3.05
Thr10(A)
3.03
HOH371(B')
3.05
3.13
J04
2.65
Ser91(B')
Ala1(A')
Trp19(A)
3.08
Arg65(A)

d

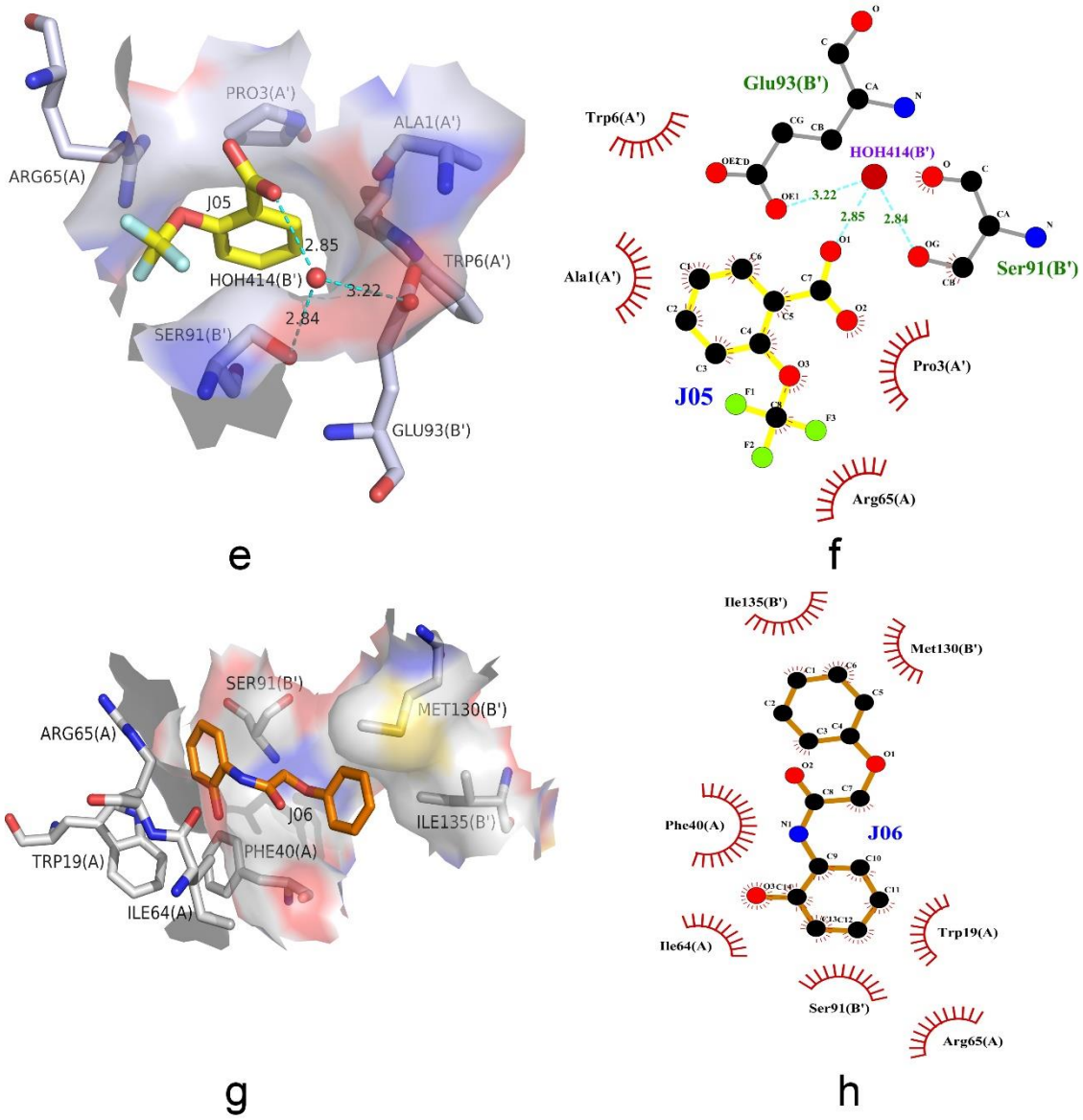


Figure 2.17 Interactions between SV3CP and the fragments J03-J06. 3D (a, c, e, g) and 2D (b, d, f, h) representations of the interactions between fragments J03-J06 and the residues in the putative RNA binding site. The binding pocket is shown by a surface representation. Hydrogen bonds are indicated by dashed lines as cyan. Hydrophobic interactions are indicated by eyebrow-like icons coloured as brown. Protein chain ID is indicated by the letters A and B in the brackets and those with a prime stand for the symmetry related chains.

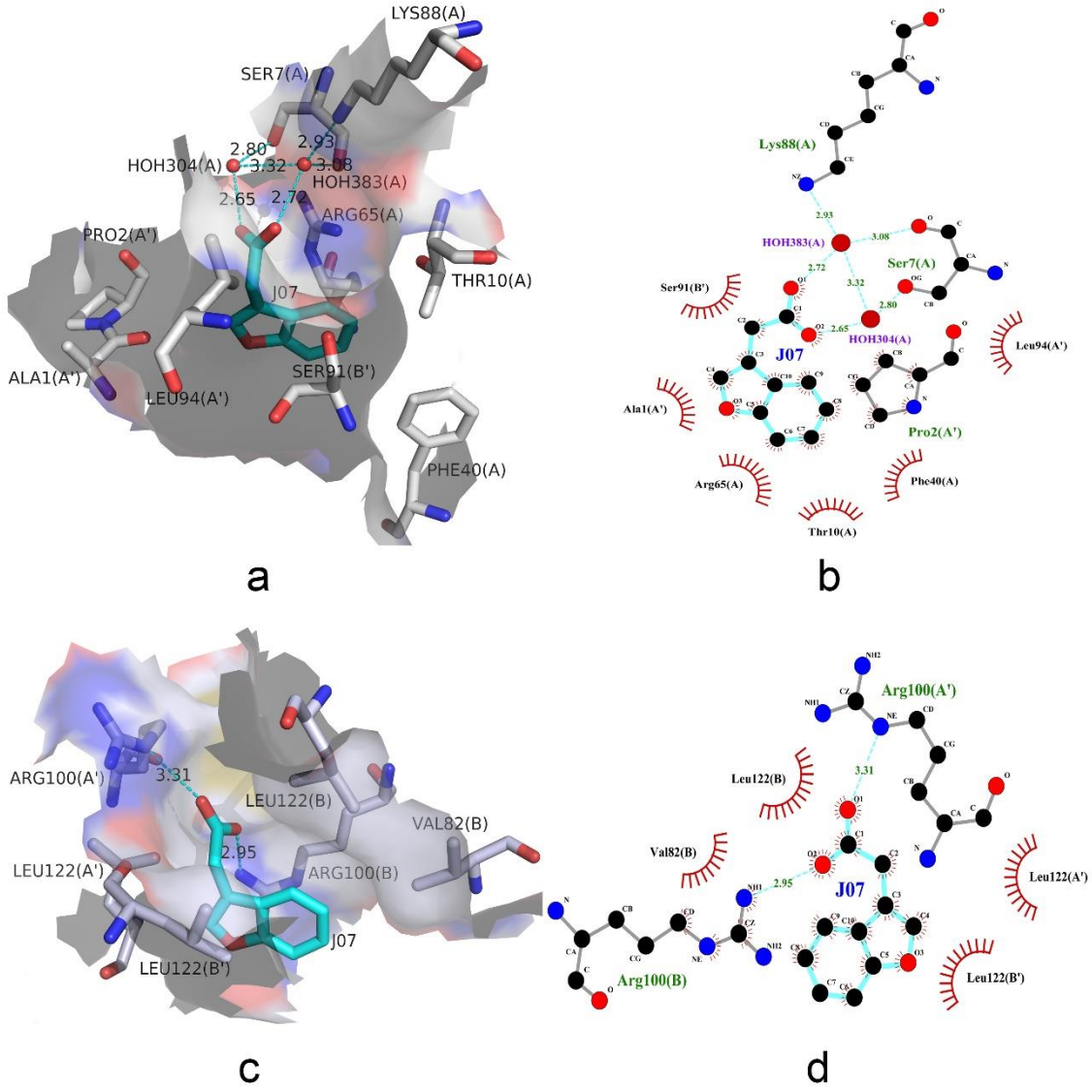


Figure 2.18 Interactions between SV3CP and the fragment J07. The same 2D and 3D representations of fragment J07 which binds in both the putative RNA binding site (a, b) and site C (c, d).

2.5.4 Other fragments

Interestingly, the peptide bonds in fragments J11 and J18 (Figure A of the appendices) are cleaved for some reason and the resulting cleaved parts containing the thiadiazole ring from both J11 and J18 bind in sites C and D, and the other part from J11 binds in a different place near Glu93. Since both of the fragments possess a 2-ethyl-1,3,4-thiadiazole group and a peptide bond in their

structures, it is likely that these features can be recognised by the enzyme and compounds possessing them can be cleaved. This gives some clues for inhibitor design, for example, inhibitors may be developed by modifying the peptide bonds of these fragments so that they can be recognised but cannot be cleaved by the enzyme. However, analysis is required to make sure they have not broken down automatically before further development is performed.

In addition, many other hits have been identified to bind in site C and some other sites, see section 2.4.3 for details.

2.6 Future work

1.The most important work in future is to evaluate the binding affinity and inhibitory activity of these fragments.

2.Given any good hits, computer- and activity-based analysis needs to be performed to identify better homologous compounds.

3.Chemistry as well as crystallography work is required to further develop the fragments into good inhibitors.

Chapter 3

Structural and functional studies of porphobilinogen deaminase from *Bacillus megaterium*

3.1 Introduction

3.1.1 Tetrapyrroles

Tetrapyrroles are a class of chemical compounds in which four pyrrole rings are linked together to form linear or cyclic molecules. These natural molecules are essential for living systems since they participate in key processes such as photosynthesis and respiration. Tetrapyrroles are only synthesised in small quantities and the synthetic pathways are tightly regulated.

There are two pathways for tetrapyrrole biosynthesis which are similar in all living organisms. The first common intermediate in these pathways is 5-aminolaevulinic acid (ALA). Higher plants and many prokaryotes adopt the C5 or glutamate pathway (Figure 3.1) in which 5-ALA is generated from the carbon skeleton of glutamate (Kannangara *et al.*, 1994; Kannangara *et al.*, 1988). Mammals and other eukaryotic organisms produce 5-ALA from glycine and succinyl-CoA along the C4 or “Shemin” pathway which requires the enzyme 5-ALA synthase (ALAS) (Leeper, 1985; Li *et al.*, 1989).

The next stage includes transformation of 5-ALA into the common precursor for tetrapyrroles, uroporphyrinogen III, in three enzymatic steps which are common in all living systems (Chadwick and Ackrill, 1994).

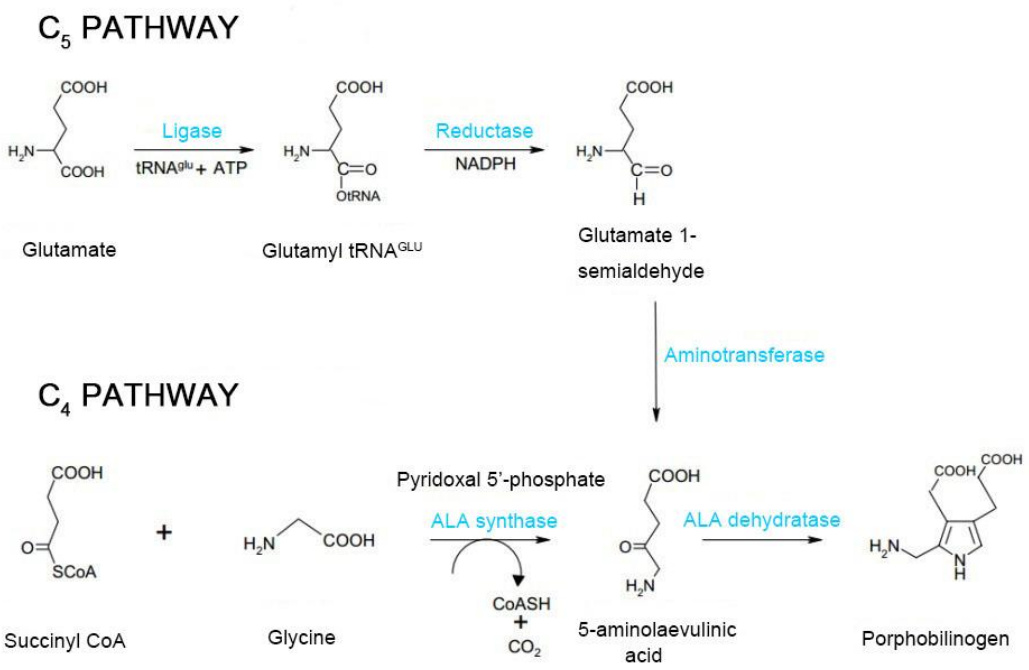


Figure 3.1 5-ALA synthesis pathways. Higher plants and prokaryotic organisms take the C5 pathway whilst eukaryotes and mammals adopt the C4 pathway.

In the first step, two molecules of 5-ALA are condensed by the enzyme 5-ALA dehydratase (ALAD, also known as porphobilinogen synthase, PBGS) to form the basic pyrrole building block porphobilinogen (PBG). In the next step, four molecules of PBG are condensed together to produce a highly unstable linear hydroxymethylbilane (HMB), which is also known as preuroporphyrinogen. This reaction is catalysed by the enzyme porphobilinogen deaminase (PBGD). The fourth pyrrole ring of the preuroporphyrinogen is then rearranged and connected with the first ring to form a uroporphyrinogen III in a reaction catalysed by the enzyme uroporphyrinogen synthase (UROS) (Warren *et al.*, 1995), as shown in Figure 3.2. Uroporphyrinogen III acts as a branch point in the pathway that many tetrapyrroles, such as chlorophyll, haem, sirohaem, F_{430} and vitamin B_{12}, are all derived from (Figure 3.3).

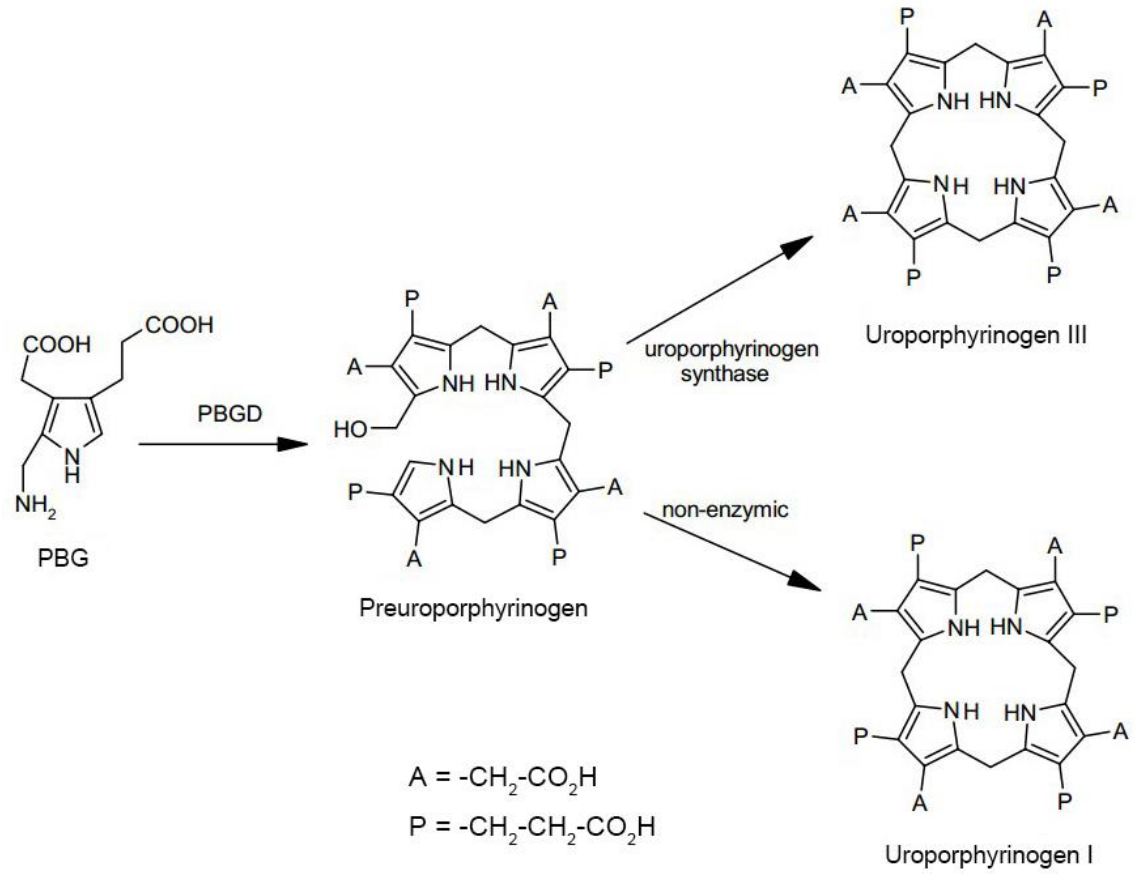


Figure 3.2 Synthesis of uroporphyrinogen I and III from PBG. Uroporphyrinogen III is synthesised from the unstable intermediate preuroporphyrinogen under the catalysis of uroporphyrinogen synthase. Uroporphyrinogen I is produced when uroporphyrinogen synthase is absent. [Figure generated based on (Chadwick and Ackrill, 1994)].

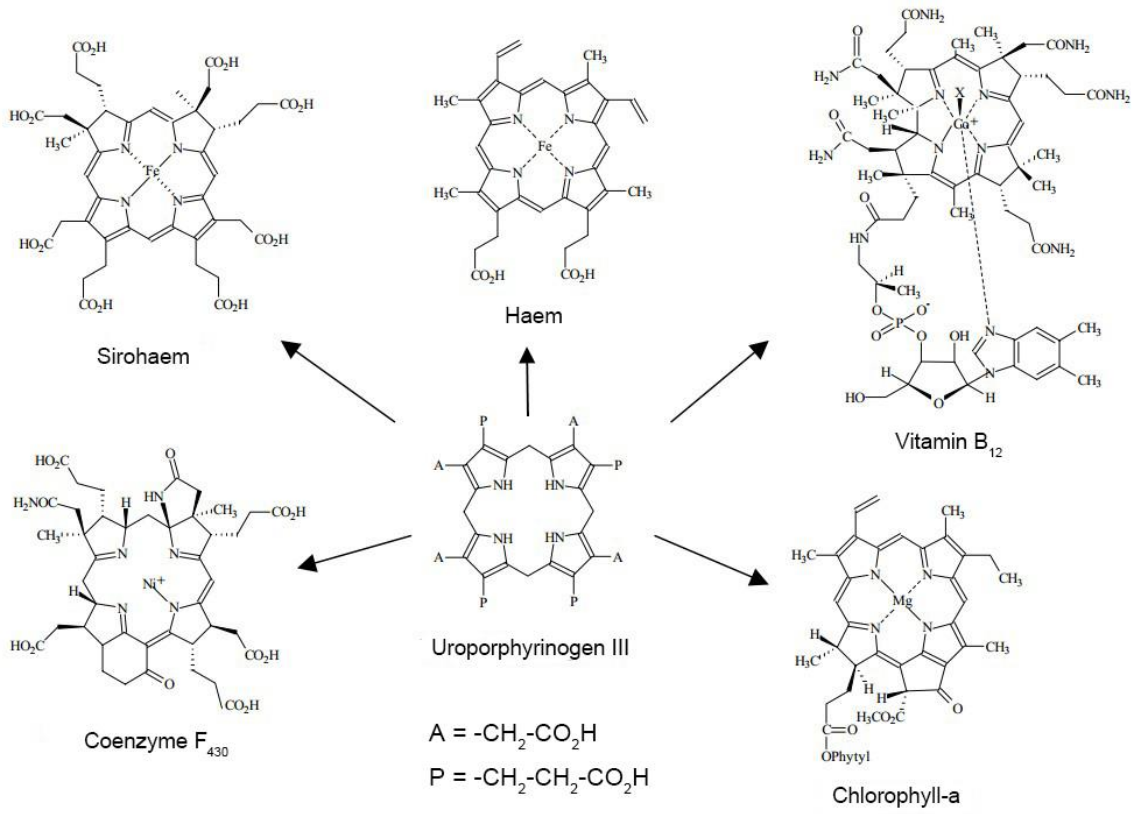


Figure 3.3 Tetrapyrroles formed from Uroporphyrinogen III.

3.1.2 Haem and haemoproteins

Haem is a prosthetic group of proteins like myoglobin and haemoglobin, and consists of a ferrous ion (Fe^{2+}) chelated in the centre of a large heterocyclic organic ring named a porphyrin. While the most common type of haem is *haem B* (Figure 3.4), other important types include *haem A* and *haem C*. As a regulatory molecule, haem mediates gene expression, translation, protein targeting, protein stability and differentiation (Chen and London, 1995; Karplus and Diederichs, 2012; Lathrop and Timko, 1993; Schmitt, 1999; Wang *et al.*, 1999). It also acts as a sink or source of electrons during electron transfer or redox chemistry. Haemoproteins are involved in many fields in oxidative metabolism including O_2 transport, O_2 sensing, oxidative stress response, oxidative phosphorylation and oxygenation reactions. They also participate in transportation of diatomic gases,

chemical catalysis, electron transfer and sensing of diatomic gases such as nitric oxide and carbon monoxide (Rodgers, 1999). Haemoproteins are coloured since the conjugated double bonds absorb visible light.

Most haem in humans is synthesised endogenously and each tissue produces haem to satisfy its own requirement. More than 70% of total haem in the body is synthesised in the bone marrow where haem is incorporated into haemoglobin for erythrocyte precursors. The liver, which is the second most important site in the human body for haem synthesis, accounts for about 15% of the total haem production. There is a high demand in the liver for haem to be incorporated into mitochondrial cytochromes as well as into cytochrome P_{450}, catalase and cytochrome b_5.

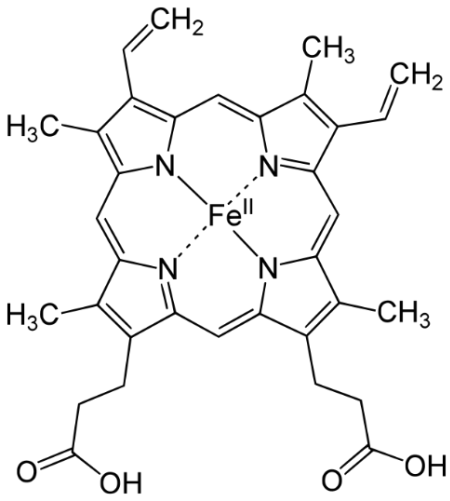


Figure 3.4 Structure of *haem B*. The most common type of haem is *haem B*, other types include *haem A* and *haem C*. A ferrous ion is located in the centre of the porphyrin molecule.

3.1.3 Porphobilinogen deaminase

Porphobilinogen deaminase (PBGD), also referred to as hydroxymethylbilane synthase, uroporphyrinogen I synthase or preuroporphyrinogen synthase,

catalyses one of the early steps in the synthesis of tetrapyrroles (Warren and Smith, 2009). Several PBGDs have been isolated from different organisms of both prokaryotic and eukaryotic sources including *Escherichia coli (*Bugg *et al.,* 2011; Jordan and Warren, 1987*)*, plants (Cooper *et al.*, 2014) and mammals (Coates *et al.*, 2006; Knight *et al.*, 2006). The primary amino acid sequences of this enzyme from different sources have a conservation of at least 32%. There are two isoforms of PBGD, the erythroid specific enzyme and the housekeeping enzyme. Both originate from a single PBGD gene on chromosome 11 and arise by alternate splicing of the primary transcript (Erskine *et al.*, 2006). PBGDs have molecular weights ranging from 34 to 44 kDa and pH optima in the range 8.0 - 8.5. These enzymes from most organisms have great thermostability which has been exploited during their purification. Most of the proteins can be heated to 60 °C or even higher temperatures for 10 to 120 minutes with no or very little loss of activity.

3.1.4 The structure of PBGD

A number of X-ray structures of PBGD have been reported including those from *Escherichia coli*, human, *Arabidopsis thaliana* and *Bacillus megaterium* (Azim *et al.*, 2014; Gill *et al.*, 2009; Louie *et al.*, 1992; Roberts *et al.*, 2013). *E. coli* PBGD was the first enzyme in the tetrapyrrole pathway to have its structure determined by X-ray crystallography (Louie *et al.*, 1992). The polypeptide itself is folded into three domains (1 - 3) each of approximately the same size. Both domain 1 and domain 2 are composed of five-stranded mixed *β*-sheets and have similar overall topology to the type II periplasmic binding proteins, which have been reported to adopt “open ” and “closed” states in response to ligand binding (Louie, 1993; Louie *et al.*, 1992). Domain 3, which possesses an open-faced anti-parallel *β*-

sheet of three strands and three *α*-helices, is folded completely differently from the other two domains (Figure 3.5).

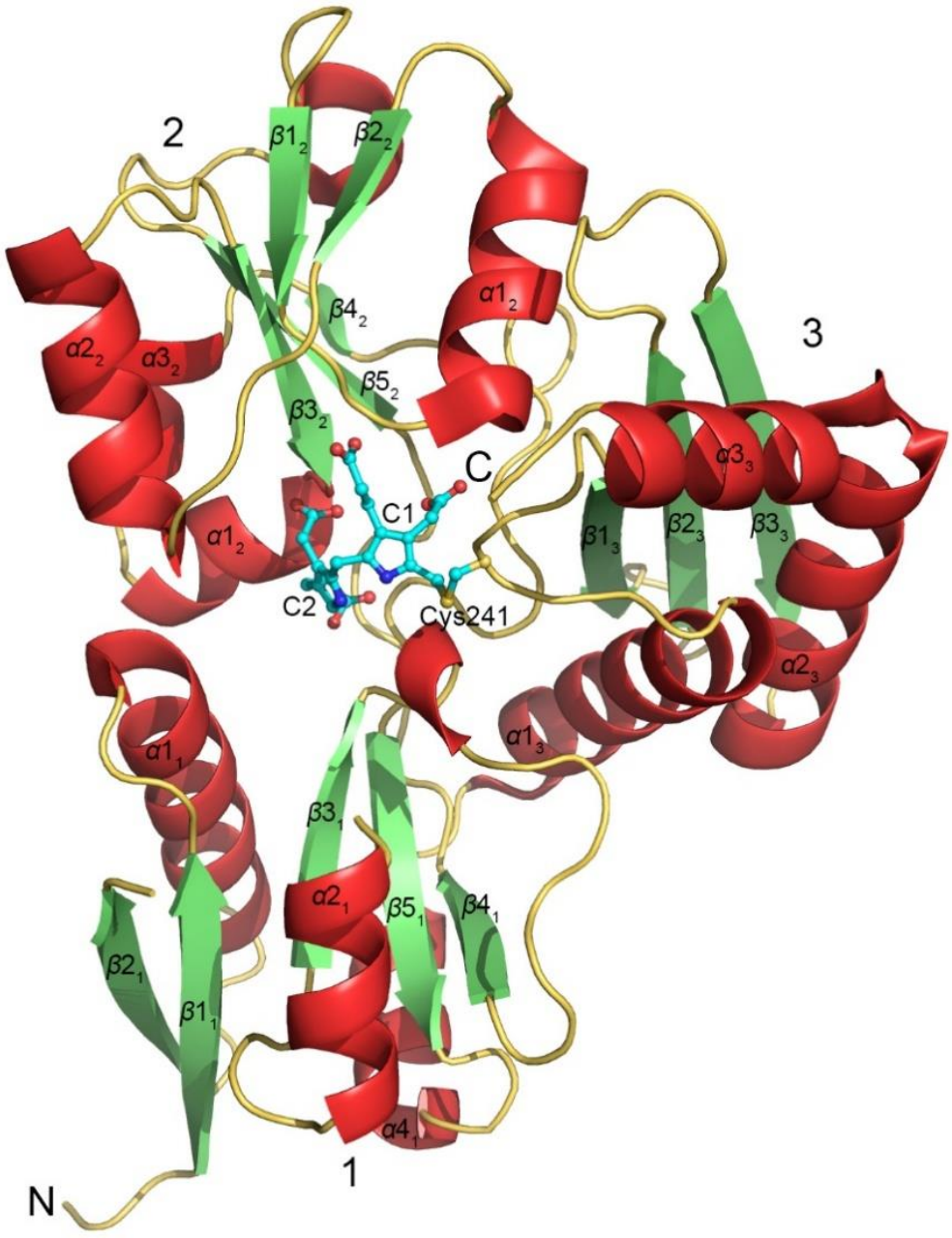


Figure 3.5 X-ray structure of the PBGD from *B. megaterium* (BPBGD, PDB ID: 4MLV). A dipyrromethane (DPM) cofactor is covalently bound to Cys241 in the active site. The domains, *α*-helices and *β*-sheets are labelled according to the nomenclature of Louie *et al* (1992). [Figure generated based on (Azim *et al.*, 2014), reproduced with permission of the International Union of Crystallography (http://journals.iucr.org)].

There is a DPM cofactor (Figure 3.6) that is covalently attached to a cysteine (Cys241 in *B. megaterium*) in a loop which is located in the active-site. The two pyrrole rings of the cofactor are named C1 and C2, with C1 being attached to the protein. The presence of the cofactor can be confirmed by the treatment of PBGD with Ehrlich's reagent which gives an absorbance at 565 nm and a subsequent shift to a λ_{max} of 495 nm after 15 min (Williams *et al.*, 2006). The cofactor can be generated in two different ways, 1) by slow assembly from two molecules of PBG or 2) by cleavage of the product, preuroporphyrinogen, which reacts rapidly with the apo-enzyme (Awan *et al.*, 1997; Shoolingin-Jordan *et al.*, 1997).

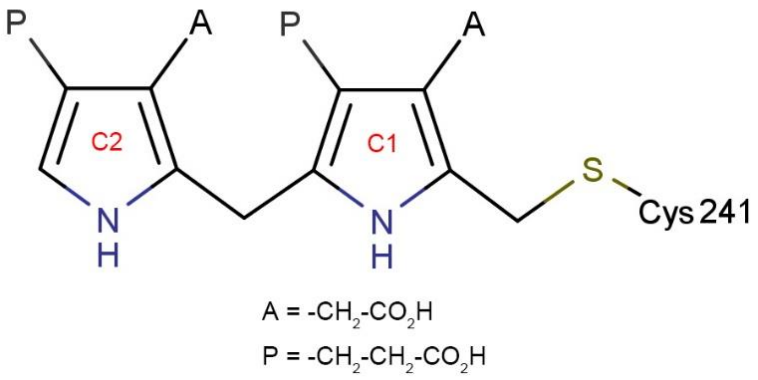


Figure 3.6 Structure of the DPM cofactor. DPM is covalently linked to Cys241 through a thioether linkage. Four molecules of PBG are linked sequentially to the cofactor via the free α-position of each pyrrole and the tetrapyrrole product is then released from the cofactor [Figure from (Guo *et al*., 2017a)].

The DPM cofactor plays two different roles in the catalysis. Firstly, it acts as the primer in the active site in order to initiate the enzymatic reaction and to connect the substrate moieties to the enzyme during the elongation stage. This has been confirmed by ^{14}C labelling experiments in which the enzyme containing a labelled

cofactor was incubated with unlabelled PBG. The final product contained no labelled compounds, indicating that the DPM cofactor was not affected by the catalytic turnover and remained permanently linked to the enzyme (Warren and Jordan, 1988). Secondly, the permanent connection between the cofactor and the enzyme may help to limit the maximum number of substrate molecules that can bind and make it precisely four.

The cofactor adopts different conformations depending on its oxidation state. The C1 ring varies little between different states, while the C2 ring in the oxidised state (in which the cofactor is called dipyrromethenone) tends to occupy a space which is thought to be the position for the incoming substrate PBG (Louie *et al.*, 1996). In this state, both rings of the dipyrromethenone factor are found to be coplanar whilst in the reduced dipyrromethane state, the cofactor adopts a conformation where the C2 ring occupies a more internal position in the active site cleft (Azim *et al.*, 2014; Louie *et al.*, 1996; Roberts *et al.*, 2013).

There is also an invariant aspartic acid (Asp82 in *B. megaterium*) that is located in a highly conserved region of the PBGD sequence, V-H-S-M-K-D-M-P, from residues 77 to 84 in the *B. megaterium* enzyme (Azim *et al.*, 2014). The aspartate side chain forms two hydrogen bonds with the NH groups of the DPM and is thought to catalyse the tetramerisation reaction.

3.1.5 Catalytic mechanism of PBGD

During the catalysis, four molecules of the substrate PBG bind to the enzyme in a stepwise head-to-tail manner through the free *α*-position of the DPM cofactor. The catalytic reaction starts with the binding of the first substrate PBG (ring A), which is deaminated and forms an azafulvene. The azafulvene then

nucleophilically attacks the free *α*-position of the C2 ring of the cofactor to form a new C-C bond (Figure 3.7). The following rings B, C and D are added sequentially in the same way (Jordan and Woodcock, 1991). In summary, the formation of the linear tetrapyrrole preuroporphyrinogen is achieved from the sequential binding, deamination and condensation of four substrate molecules through a few stable intermediates known as ES, ES_2, ES_3 and ES_4 (Figure 3.8) (Shoolingin-Jordan, 1998). This was firstly identified by incubating the human erythroid enzyme deaminase with [^{3}H]-PBG after which the labelled four complexes were detected (Anderson and Desnick, 1980). These intermediate complexes are more negatively charged than the original enzyme so that they could be separated by electrophoresis or ion exchange chromatography (Anderson and Desnick, 1980). ES_4, which is actually a hexapyrrole composed of the cofactor and the tetrapyrrole bilin product, is then hydrolysed and the unstable product (preuroporphyrinogen) is released from the enzyme (Jordan and Woodcock, 1991; Warren and Jordan, 1988). This is achieved by the protonation of the carbon in the *α*-position of the C2 ring followed by the cleavage of the C-C bond between the C2 ring and ring A, which frees up the cofactor for the next catalytic reaction.

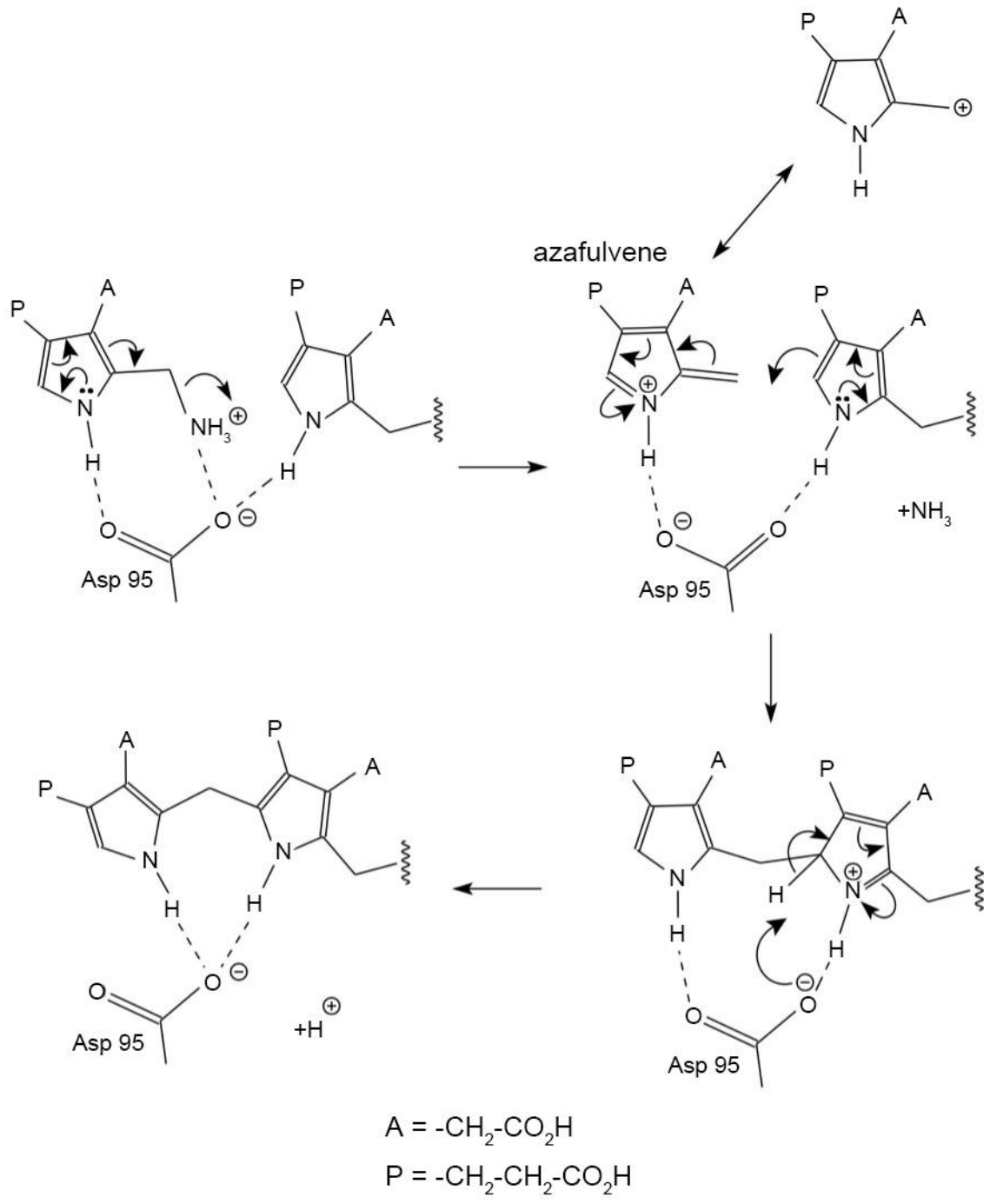


Figure 3.7 The mechanism for attachment of a PBG molecule. The PBG molecule can be attached to the cofactor or any of the subsequent intermediates through three steps known as: deamination of PBG, nucleophilic attack of the free *α*-position of the previous ring to be attached and deprotonation of the C^{α} atom. [Figure generated based on (Roberts *et al.*, 2013), reproduced with permission of the International Union of Crystallography (http://journals.iucr.org)].

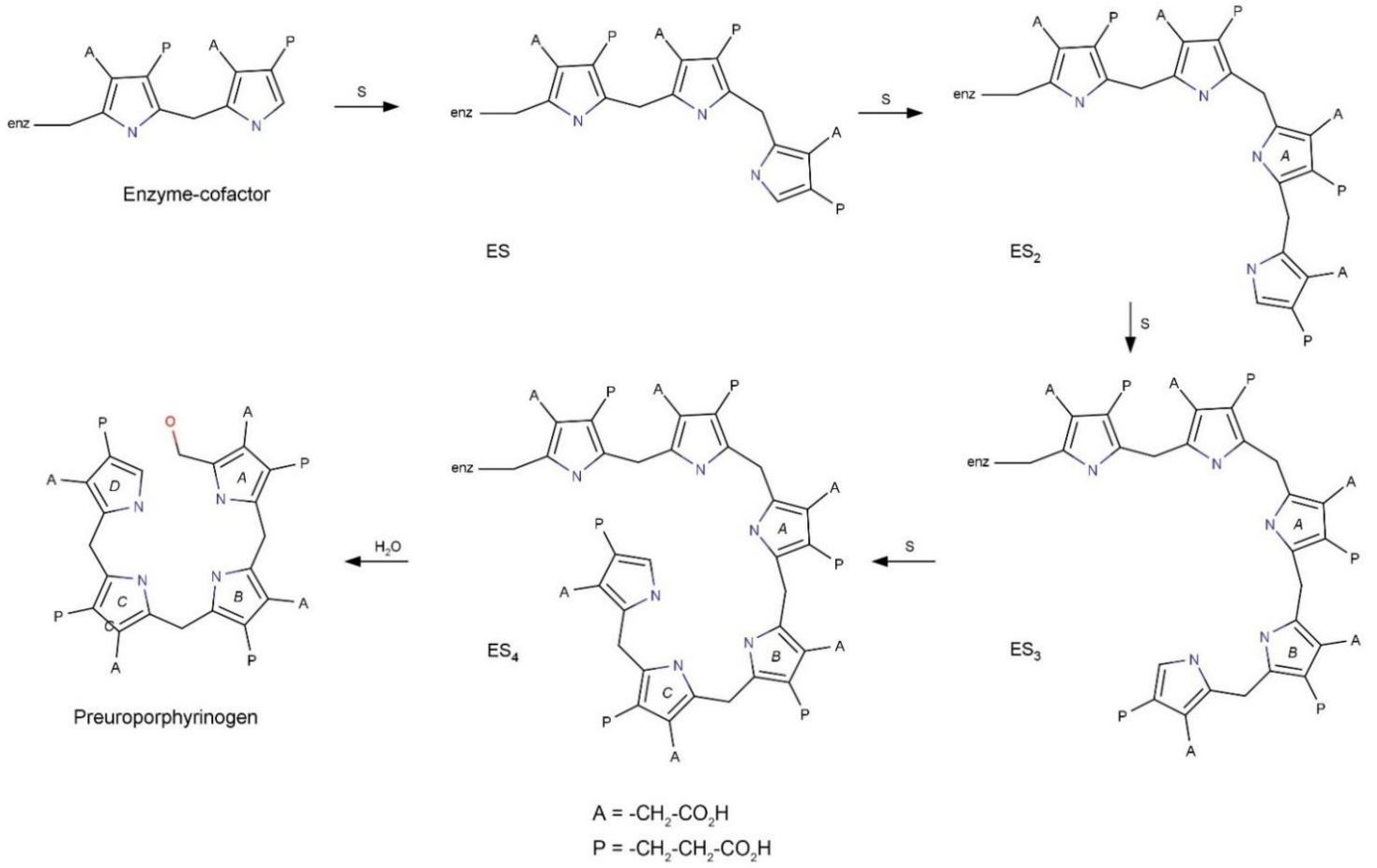


Figure 3.8 Formation of preuroporphyrinogen. The assembly of preuroporphyrinogen is formed from four molecules of PBG in a stepwise manner through the intermediates known as ES, ES_2, ES_3 and ES_4. [Figure generated based on (Jordan and Woodcock, 1991)].

3.1.6 Mutagenesis studies of PBGD

Comparison of the sequence of PBGDs among *E. coli*, human, *B. megaterium* and *A. thaliana* (Figure 3.9) identified that there are many highly conserved arginine residues at positions 11, 101, 131, 132, 149, 155, 176, 182, 206, 232 and 277 in *E. coli* PBGD, several of which bind to the acetate and propionate groups of the cofactor. Site-directed mutagenesis of these arginine residues revealed that they play important roles in the catalytic mechanism (Jordan and Woodcock, 1991). The enzyme lost the ability to assemble the cofactor with mutations at R131 or R132, resulting in inactive mutants, while mutations at R149 or R176 produced mutants with less activity and had the ability to accumulate intermediates. In addition, R11 or R155 mutated proteins were able to assemble the cofactor, but did not bind to the substrate PBG (Azim *et al.*, 2014; Jordan and Woodcock, 1991).

Figure 3.9 Sequence alignment of PBGDs. PBGD from different organisms including *B. megaterium*, *E. coli*, *A. thaliana* and human are compared. The secondary structural elements based on the BPBGD are shown at the bottom. The amino acid residues are coloured as follow: cyan, basic; red, acidic; green, neutral polar; pink, bulky hydrophobic; white, Gly, Ala and Pro; yellow, Cys. [Figure generated based on (Azim *et al.*, 2014), reproduced with permission of the International Union of Crystallography (http://journals.iucr.org)].

Site-directed mutagenesis was undertaken by Jordan and Woodcock (1991; 1994) in which the Asp84 in *E. coli* PBGD (Asp82 in *B. megaterium*) was mutated

to Glu, Ala and Asn. The D84E mutant had a significantly reduced activity of less than 1% of that of the WT enzyme while the other two mutants, D84A and D84N, were completely inactive. It was found that the cofactor of the D84E mutant was more sensitive to oxidation and one reason might be that the hydrogen bond between the carboxylic group of Asp84 and the NH group of the C2 ring was weaker or lost due to the mutation (Lambert *et al.*, 1994).

3.1.7 Acute intermittent porphyria

The porphyrias are a group of inherited disorders caused by enzyme defects in haem biosynthesis and the resulting accumulation of phototoxic intermediates. Acute intermittent porphyria (AIP) is an autosomal dominant metabolic disease resulting from diminished erythrocyte activity of PBGD. Onset of AIP typically occurs during puberty or later and symptoms include severe abdominal pain, vomiting, peripheral, central neuropathy and proximal motor weakness, which may well be due to the toxicity of excess 5-ALA that has a similar structure and may behave as an analogue of the inhibitory neurotransmitter *γ*-aminobutyric acid (GABA). The attacks may be precipitated by circumstances including the use of drugs (e.g. barbiturates and sulphur-containing drugs), alcohol and heavy metals as well as stress, poor diet, infections and hormonal changes (Wood *et al.*, 1995). The reduced ability of the enzyme to fold and function is caused by many of the point mutations in PBGD, which affect the amino acid residues (mainly arginine residues in the active site) of the enzyme, as mentioned by before. In addition, some other mutations probably destabilise the three-dimensional fold by steric and electrostatic effects (Delfau *et al.*, 1990; Jordan and Woodcock, 1991; Lander *et al.*, 1991; Scott *et al.*, 1989).

Biochemical tests can help the diagnosis of the AIP through estimating the PBGD activity in erythrocytes and/or cultured transformed lymphocytes, though 5% to 10% of affected individuals exhibit normal erythrocyte PBGD activity. Another way is to measure the PBG and 5-ALA in urine, however, both of them are most clearly disturbed only during the acute attacks. In addition, some patients with AIP show normal red cell enzyme activity, but depletion in other tissues. Since it is an inherited disorder, family members of a patient with AIP should take suitable tests as well. It is important to undertake the diagnosis as soon as possible because acute episodes of AIP can be fatal (Moore and McColl, 1989). Treatment with large doses of haem can be effective through inhibition of the first reaction of the pathway, shutting down the flux and providing haem for the patient. Avoidance of precipitating factors is also useful in preventing the symptoms of AIP.

3.2 Project aim

The aim of this investigation was to use a combination of mutagenesis, X-ray crystallography, enzymology, computational analysis to explore how PBGD works, in terms of the substrate binding and domain movements during catalysis. Structural information gained from X-ray crystallography of mutants would be used to determine the domain movements and enzymatic assays would be performed to establish the kinetic behaviour and the role of the cofactor.

3.3 Methods

3.3.1 Mutagenesis and DNA transformation

The WT BPBGD gene in a pET-14b expression construct, as previously reported by Azim (Azim *et al.*, 2013), was applied to the QuickChange mutagenesis kit

(Agilent Technologies, Cheshire, UK) to introduce the required base changes. The primers were ordered from Yorkshire Bioscience (York Science Park, York, UK) with sequences as follows:

D82A (GAT to GCT):

Forward: 5'-GGCCGTTCATAGTATGAAAGCTATGCCGGC-3'

Reverse: 5'-GCCGGCATAGCTTTCATACTATGAACGGCC-3'

D82E (GAT to GAA):

Forward: 5'-GGCCGTTCATAGTATGAAAGAAATGCCGGC-3'

Reverse: 5'-GCCGGCATTTCTTTCATACTATGAACGGCC-3'

D82N (GAT to AAT):

Forward: 5'-GATATGGCCGTTCATAGTATGAAAAATATGCCGGCTG-3'

Reverse: 5'-CAGCCGGCATATTTTTCATACTATGAACGGCCATATC-3'

The base changes are coloured in red.

Each mutagenesis reaction was undertaken using a Techne Progene thermal cycler (Techne, Staffordshire, UK) following the protocol provided in the kit. The nicked plasmid DNA was then transformed into 'subcloning efficiency DH5α' competent cells (Invitrogen, Thermo Fisher Scientific, Dartford, UK) for nick repair and DNA amplification, following the protocol described in Method i of the appendices. The plasmid DNA was then extracted and purified by use of an AxyPrep Plasmid Miniprep Kit (Axygen, Union City, California, USA).

DNA sequencing, carried out at DNA Sequencing & Services (University of Dundee, Dundee, UK), confirmed that the base changes were introduced successfully (Figure B of the appendices) which could also be seen in the X-ray

structures determined later. Each of the mutated genes was transformed into Rosetta (DE3) pLysS *E. coli* cells (Novagen, Darmstadt, Germany) for protein expression.

3.3.2 Protein preparation

All the mutant proteins were expressed by using the heat shock method described in Method ii of the appendices. The His-tagged proteins were then purified initially by a HisTrap FF column, followed by the cleavage of the tag using thrombin. A HiTrap benzamidine column was then used to remove the thrombin and the proteins were finally desalted and purified by a Superdex75 gel-filtration column. All the columns used were purchased from GE Healthcare (Buckinghamshire, UK).

3.3.3 Protein crystallisation

Screening for crystallisation conditions for all the BPBGD mutant proteins was undertaken by use of the hanging-drop method with the Structure Screens 1 & 2 kit from Molecular Dimensions (Suffolk, UK). 5 µl of each mutant protein (5 mg/ml) was mixed with 5 µl of the corresponding well solution on a siliconised coverslip and the plates were stored at 21 °C for crystallisation. Showers of small yellow crystals for the D82N protein started to appear after three days in Screen 1 condition 2 (0.2 M ammonium acetate, 0.1 M sodium acetate pH 4.6, 30% PEG 4,000). Subsequent optimisation revealed that better crystals for all the mutants could be obtained reproducibly in 0.2 M ammonium acetate, 0.1 M sodium acetate pH 3.5 - 4.0, 22 - 28 % PEG 4,000 (Figure 3.10). These crystals grew as clusters formed by thin plates and there was no significant difference among

crystals of the mutants. It seemed that all of the mutants preferred crystallising at a ‘final’ pH of around 5.7 in each hanging droplet.

In order to get protein-substrate complex structures, co-crystallisation was undertaken for all BPBGD mutants with the substrate PBG ranging from 0.14 to 1.4 mM (1-10 times in molar excess of the protein) using the same crystallisation condition. In addition, co-crystallisation of the mutants with the WT enzyme and PBG was also carried out with a ratio of WT:mutant:PBG = 1:50:500 (in molar excess). The idea behind this experiment was that the WT enzyme might make sufficient product which would then bind to the inactive mutants, given the evidence that the product is actually the precursor of the cofactor (Jordan and Warren, 1987; Mauzerall and Granick, 1956; Pluscec and Bogorad, 1970).

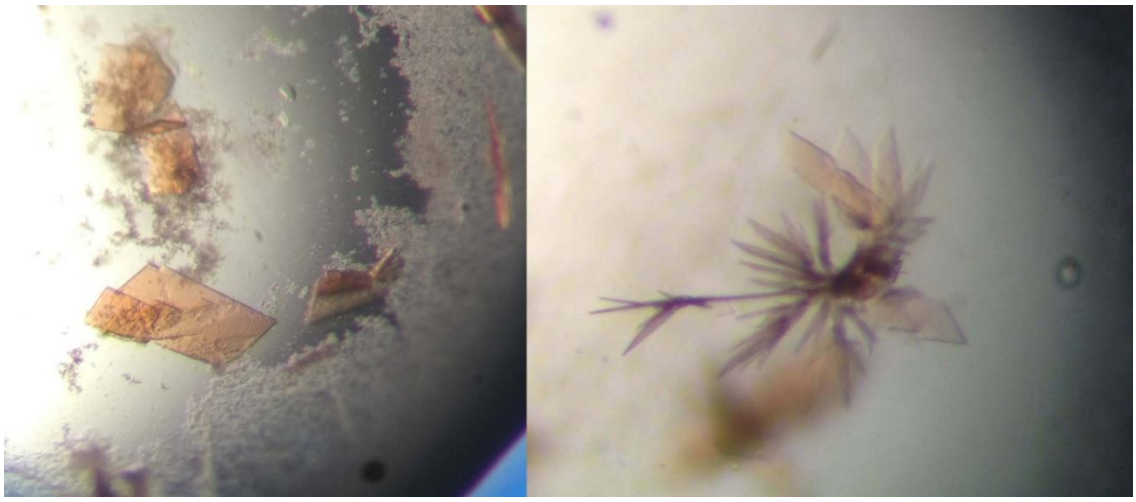


Figure 3.10 Crystals of BPBGD D82N mutant from *B. megaterium*. A single fragment used for data collection was approximately 400 μm in its longest dimension and 100 μm wide.

3.3.4 Crystal freezing, data collection and data processing

Selected crystals were mounted in loops with approximately 30% glycerol as the cryo-protectant before flash-cooling in a nitrogen gas cryostream (Oxford

Cyrosystems Ltd., Oxford, UK) at 100 K or immersion in liquid ethane. Crystals were then stored in pucks in liquid nitrogen prior to exposure to X-ray beam.

X-ray data collection was performed at stations I02, I03, I04 and I04-1 at DLS. 190 degrees of data were collected for every crystal which yielded good diffraction with 1 s exposure times, 1° oscillation and 20% transmission. Data collection revealed that all the crystals for the mutants were orthorhombic and belonged to the space group $P2_12_12_1$ which is the same as that of the WT protein (Azim *et al.*, 2013) and the unit-cell dimensions are somewhat smaller for the mutants.

Data processing was achieved through two different ways, either by use of the automatic processed data from *xia2* (Winter, 2010) or by manual processing. For manual data processing, *iMosflm* (Battye *et al.*, 2011; Leslie, 2006) was used to integrate the diffractions spots and produce the reflection .mtz file before *Scala* (Evans, 2006) was used to scale all the equivalent symmetry related reflections together. *Ctruncate* (Dauter, 2006; French and Wilson, 1978) was then used to obtain the structure factor amplitudes from the diffraction intensities.

3.3.5 Structure determination, model building, refinement and validation

Structure determination was undertaken by molecular replacement using *Molrep* (Vagin and Teplyakov, 2010) with the WT enzyme structure as the search model (PDB ID: 4MLV). Following one round of refinement of the solutions, manual rebuilding and the introduction of the mutant residues Ala, Glu and Asn, to replace Asp82 in the WT structure, was accomplished by use of the program *Coot* (Emsley and Cowtan, 2004). The structures were then refined by restrained refinement with *Refmac5* (Murshudov *et al.*, 1997; Winn *et al.*, 2001; Winn *et al.*,

2003). All the statistics for data collection, data processing, structure determination and refinement are shown in Table 3.1.

Table 3.1 X-ray statistics for all the three mutant structures. Values in parentheses are for the outer resolution shell [Table from (Guo *et al*., 2017a)].

	D82A	D82E	D82N
Beamline	I03 (DLS)	I03 (DLS)	I02 (DLS)
Wavelength (Å)	0.9763	0.9763	0.9795
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit-cell parameters (Å)			
a (Å)	49.1	49.2	49.0
b (Å)	62.5	62.7	62.7
c (Å)	91.4	91.8	91.3
Mosaic spread (°)	0.37	0.24	0.53
Resolution (Å)	29.56-2.69 (2.76-2.69)	91.83-1.81 (1.86-1.81)	36.89-1.87 (1.92-1.87)
R_{merge} (%)	12.8 (48.3)	4.8 (62.9)	6.9 (54.1)
R_{meas} (%)	15.1 (58.5)	5.3 (69.7)	8.2 (65.5)
$CC_{½}$ (%)	99.0 (79.7)	99.9 (76.6)	99.6 (78.8)
Completeness (%)	99.6 (97.4)	99.9 (100.0)	98.9 (96.5)
Average $I/\sigma(I)$	11.8 (3.2)	20.7 (2.6)	16.9 (4.0)
Multiplicity	6.4 (5.6)	6.7 (5.8)	6.2 (5.7)
No. of observed reflections	52,592 (3,255)	178,498 (11,002)	194,239 (12,783)
No. of unique reflections	8,221 (579)	26,586 (1,907)	23,650 (1,679)

Wilson plot *B*-factor (Å^2)	51.8	24.0	24.0
Solvent content (%)	36.9	37.5	36.8
R_{factor} (%)	20.9	18.3	21.0
R_{free} (%)	23.9	22.6	25.3
RMSD bond lengths (Å)	0.002	0.002	0.003
RMSD bond angles (°)	0.479	0.560	0.515
No. of reflections in working set	8,186	26,585	23,589
No. of reflections in test set	394	1,310	1,164
Mean protein *B*-factor (Å^2)	38.7	32.6	28.7

All the mutant structures were validated by the program *MolProbity* (Chen *et al.*, 2010).

3.3.6 Ehrlich's reaction

Modified Ehrlich's reagent:

4-dimethylaminobenzaldehyde	1g
Glacial acetic acid	42 ml
Perchloric acid (70%)	8 ml

The reagent was made fresh every time and was kept in the dark.

Ehrlich's reaction to determine the status of the DPM cofactor in BPBGD WT and mutant enzymes was carried out as described by Jordan and Warren (1987; 1988). In brief, an aliquot (500 µl) of each enzyme solution (1 mg/ml) was mixed with 500 µl of modified Ehrlich's reagent at room temperature and any precipitate was removed by centrifugation. The absorbance for each sample was monitored by a wave scan between 380 nm and 600 nm after 1 and 15 min using an

Ultrospec 3000 UV/Visible Spectrophotometer (GE Healthcare, Buckinghamshire, UK). 500 µl of protein buffer (50 mM Tris, 100 mM NaCl, pH 7.3) was mixed with an equal volume of modified Ehrlich's reagent as a blank. Observation of a peak at λ_{max} = 565 nm which subsequently shifted to a λ_{max} = 495 nm after 15 min confirmed the presence of the cofactor.

3.3.7 Determination and classification of domain movements

Relative domain movements were analysed using the *DynDom* website (Hayward and Berendsen, 1998; Hayward and Lee, 2002). The hinge and shear classification of the domain movements was determined using the Dynamic Contact Graphs (DCG) method described by Taylor *et al.* (2013; 2014). This method is based on the contact changes between residues from different domains as a result of the domain movements. There are five types of contact changes named as: 'no contact change', 'new contact change', 'maintained contact change', 'exchange-partner contact change' and 'exchange-pair contact change'. *PyMOL* (The *PyMOL* Molecular Graphics System, Schrödinger, LLC) was used to visualise residues have contact changes and the Bioinformatics Toolbox in *MATLAB* (*MATLAB 2014b*, The MathWorks Inc, Natick, MA) was used to create the DCG when determining the classifications. The value of y(N) can be calculated from equation 3.1:

$$y(N) = \frac{1}{1+e^{\alpha}} \tag{3.1}$$

Where y(N) is the logistic regression of the numbers of the four different types of contact changes and $\alpha = -0.2387N_{maint} - 0.0356N_{exchpart} + 0.4249N_{exchpair} + 0.2122N_{new} + 0.1467$.

The classification was determined as follow:

Shear motion: $0 \leq y \leq 0.45$

Mixed motion: $0.45 < y < 0.55$

Hinge motion: $0.55 \leq y \leq 1.0$

3.3.8 Kinetic assay

BPBGD kinetic assay was conducted for WT and all mutant proteins according to the method described by Jordan *et al* (1997). A 25 µl aliquot of the enzyme (0.5 mg/ml) was pre-incubated with 75 µl of 20 mM Tris-HCl buffer, pH 7.2, at 37°C in a water bath for 2 min. The reaction was initiated by the addition of 50 µl of a pre-warmed PBG solution ranging from 0.05 mM to 2.5 mM. 65 µl of 5 M HCl was added to stop the reaction after 5 min at 37°C, followed by the addition of 25 µl of fresh benzoquinone solution [0.1% (w/v) in methanol] in order to oxidise the porphyrinogens to porphyrins. Each sample was then stored on ice in the dark for 20 min after which 50 µl of saturated sodium metabisulphite was added to decolourise any remaining benzoquinone. The sample was then diluted 10-fold with 1 M HCl and any precipitate was removed by centrifugation. The absorbance was determined at 405.5 nm (E_M = 548000 M^{-1} cm^{-1}) using a Ultrospec 3000 UV/Visible Spectrophotometer (GE Healthcare, Buckinghamshire, UK) and the kinetic calculation and figure plotting were achieved using *OriginPro* 9.1 (OriginLab, Northampton, MA).

3.4 Results

3.4.1 Protein preparation

Successful expression of each protein was suggested by a clear band slightly larger than 35 kDa on the SDS-PAGE gel, see Figure 3.11a for the D82N mutant. Heat shock at 42°C followed by incubating at 16°C was found to be necessary for producing soluble proteins. The initial purification was undertaken by nickel affinity chromatography which allowed his-tagged proteins to be purified under gentle and non-denaturing conditions. All the fractions from the column including the loading flow-through (L), the wash-through (W) and the eluted sample (E), were collected for SDS-PAGE analysis (Figure 3.11b). The strong band observed in the eluted sample indicated successful purification of the desired protein. The His-tag was then removed by thrombin digest and the cleaved protein (lane T) was found to be slightly smaller in molecular mass compared with the one in the eluted sample, suggesting that the His-tag plus a few residues (2 kDa) had been removed from the protein. Gel filtration was undertaken to further purify the proteins and to exchange the high salt wash buffer to Tris buffer (50 mM Tris, 100 mM NaCl, pH 7.3). The purified proteins were concentrated to 5 mg/ml using a Vivaspin centrifugal concentrator (GE Healthcare, Buckinghamshire, UK) before being used in crystallisation trials.

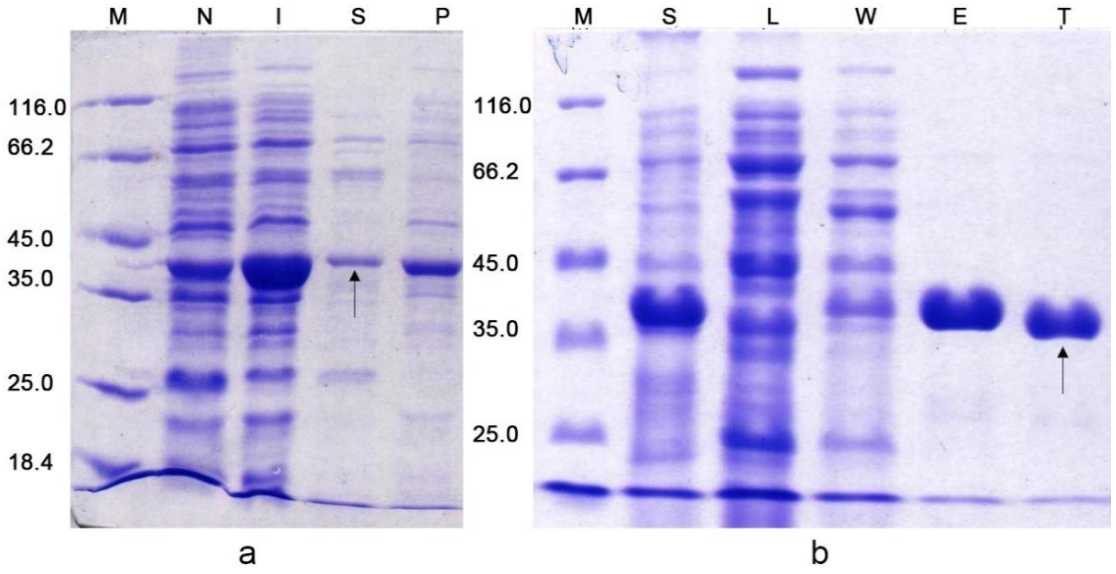


Figure 3.11 SDS-PAGE for D82N expression and purification. Lanes M, N, I, S, P, L, W, E and T correspond to the marker, non-induced, induced, supernatant, pellets, loading flow-through, wash-through, eluted and thrombin-digested samples. The D82N mutant of BPBGD protein is indicated by the arrows.

3.4.2 Protein crystallisation

It proved to be fairly difficult to crystallise the PBGD mutant proteins from different sources. Different screening kits, temperatures, additives and methods like proteolysis (Keegan *et al.*, 2014), lysine methylation to reduce surface entropy (Sledz *et al.*, 2010) were tried, but no crystal-like objects were obtained. By chance, a few small crystal-like needles of D82N mutants were obtained in condition 2 of the Structure Screen 1 kit (0.2 M ammonium acetate, 0.1 M sodium acetate pH 4.6, 30% PEG 4,000) from Molecular Dimensions and a thick yellow crystal (Figure 3.12) was obtained in condition 14 of the same kit (0.2 M ammonium sulphate, 0.1 M sodium cacodylate, pH 6.5, 30% PEG 8,000). Subsequent optimisation revealed that better crystals of all the mutant proteins could be obtained reproducibly in 0.2 M ammonium acetate, 0.1 M sodium acetate pH 3.5 - 4.0 and 22 - 28 % PEG 4,000. However, condition 14 was quite

difficult to reproduce. The optimisation of condition 2 showed that the mutants prefer crystallising at final pH values of around 5.7 in each drop.

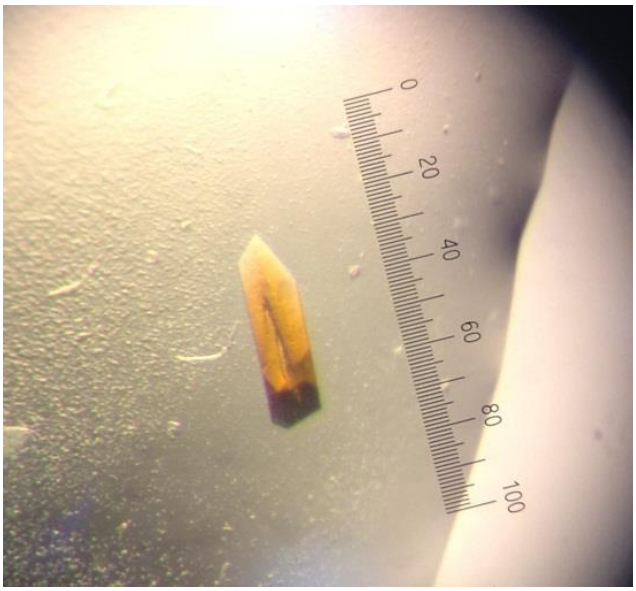


Figure 3.12 An unusual form of D82N crystal. The crystal was obtained in 0.2 M ammonium sulphate, 0.1 M sodium cacodylate, pH 6.5, 30% PEG 8,000, which was quite difficult to reproduce. One small unit on the scale is 10 microns.

3.4.3 Data collection, data processing, structure determination and refinement

All of the three mutant proteins diffracted reasonably well during data collection, but there were some differences. While D82E diffracted the best of all with a high resolution limit of 1.81 Å, D82A had the lowest resolution of 2.76 Å, leaving D82N in the middle with a resolution of 1.87 Å which was close to that of D82E. The same trend could be seen with the other parameters for the data collection, data processing and refinement such as the mosaic spread, R_{merge}, R_{meas} and $I/\sigma I$. This might be because there was a big difference in the status of the cofactor. It was found that, by comparing the structures solved later, the D82E mutant has the cofactor bound tightly while the D82A has nothing observable in the active site.

In contrast, the D82N has only the C1 ring present (Figure 3.13). Residues from 40 to 60 in D82A, D82E and from 40 to 59 in D82E were missing in the structure which might be due to proteolytic digest during the treatment with thrombin or the flexibility of these residues themselves.

It was mentioned by Woodcock and Jordan that the mutant PBGD enzymes could bind to the substrate PBG to form stable enzyme-substrate complexes depending on the PBG concentration and these complexes could be separated by ion-exchange chromatography (Woodcock and Jordan, 1994). All the mutant enzymes were co-crystallised with different concentrations of the PBG but, unfortunately, no crystal of any of the intermediate complexes was obtained.

Table 3.2 The Ramachandran statistics for BPBGD mutant structures.

	D82E	D82N	D82A
Ramachandran favoured (%)	97.5	97.6	99.3
Ramachandran outliers (%)	0	0	0

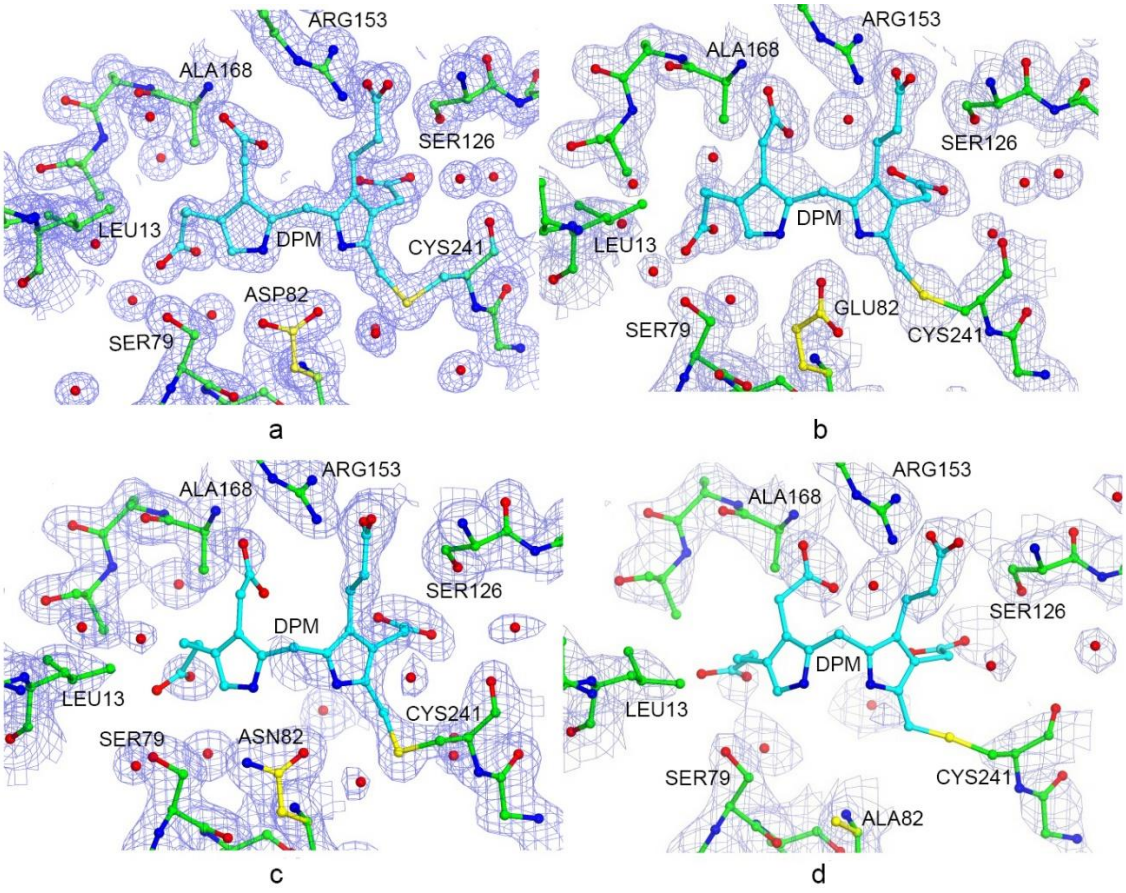


Figure 3.13 The electron density map for the DPM cofactors and a selection of the surrounding active-site residues in all the mutant structures. There is almost no density for the cofactor in D82A and for the C2 ring of the cofactor in D82N, while the D82E mutant shows good density for both rings of the cofactor [Figure from (Guo *et al*., 2017a)].

3.4.4 Ehrlich's reaction

The free *α*-position of the DPM cofactor of PBGD can react with Ehrlich's reagent which initially gives a dark purple colour to the solution and has a λ_{max} at 565 nm. This is because of the formation of a double bond between the *para*-dimethylaminobenzaldehyde (DMAB) and the free *α*-position of the DPM cofactor, as shown in Figure 3.14. The colour changes to orange after 15 min which gives a λ_{max} at 495 nm due to the formation of the conjugated bonds (Jordan and Warren, 1987).

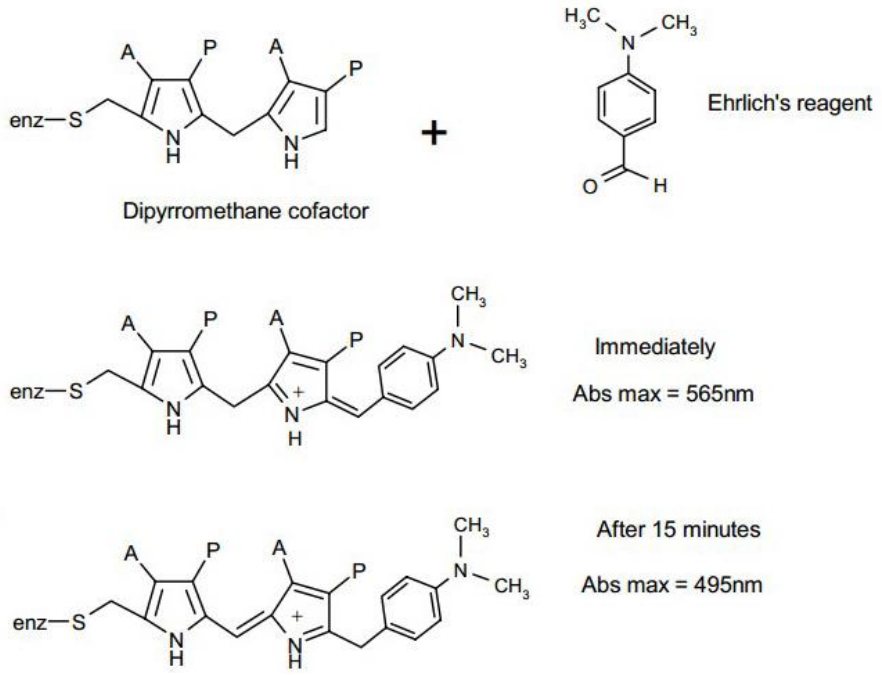


Figure 3.14 The reaction of the DPM cofactor with Ehrlich's reagent [Figure from (Guo *et al*., 2017a)].

All the WT and mutant BPBGD enzymes were treated with Ehrlich's reagent and the absorbance of each sample was monitored at both 1 min and 15 min after initiation of the reaction. As shown in Figure 3.15, the WT BPBGD had a distinct peak at 565 nm at 1 min which shifted to 495 nm after 15 min, confirming the existence of the cofactor. The D82E mutant protein had the same peaks which also proved the presence of the cofactor. In contrast, the peaks were much lower in the D82A sample, which indicated that the occupancy of the cofactor was probably quite low in this mutant. The loss of the C2 ring in the D82N mutant actually didn't affect the binding of the DMAB which can also bind to the C1 ring of the cofactor, as indicated by the peak at around 565 nm. However, the lower peak at 495 nm may indicate that less conjugated DMP is being formed because of the lower occupancy of the C2 ring.

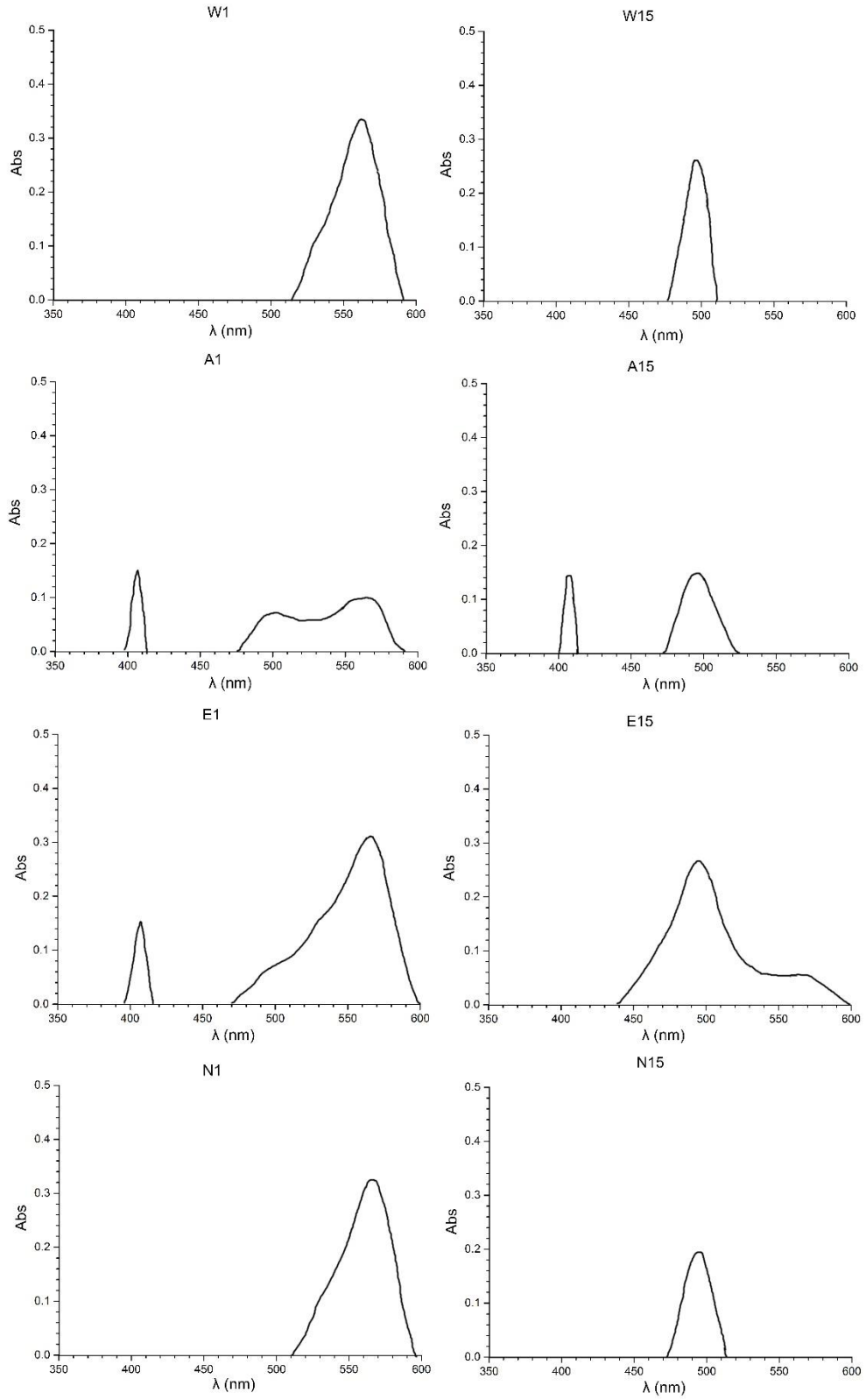


Figure 3.15 Absorbance spectrum of the products of the Ehrlich's reactions. The column on the left indicates the spectrum after 1 minute and the column on the right corresponds to the spectrum after 15 minutes. W, E, N and A stand for the WT, D82E, D82N and D82A with the numbers after them representing 1 min and 15 min [Figure from (Guo *et al*., 2017a)].

3.4.5 Domain movements

The *DynDom* website was used to determine the domain movements, the screw axis and the bending residues during the domain movements. It was confirmed that domains 2 and 3 tend to move together as a rigid body except that there is an additional local movement of the $\alpha 2_2$-helix and the $\alpha 3_2$-helix in domain 2 (Figure 3.16b). Superposition of domains 1 and 3 of the WT BPBGD structure individually with the counterparts in all the mutant proteins gave *RMSD* values of around 0.4 Å, which were not significant. However, there were bigger differences in domain 2 with *RMSD* values of all above 0.85 Å. *RMSD* values of all above 1.0 were observed when considering the whole protein as a rigid body in the superpositions (Figure 3.16a, Table 3.3).

When considering domains 2 and 3 as a rigid part, domain 1 in the mutants was found to rotate about 7° with negligible translation when compared with the WT. Thus a screw axis representing the rotation and translation movements between each pair of structures (WT and D82E, WT and D82N, WT and D82A) was generated using *DynDom* (see Figure 3.17 as an example for the pair between WT and D82A). The contact changes were determined using *PyMOL* and Dynamic Contact Graphs were produced using *MATLAB* (Figure 3.18 as an example of D82A). According to equations 3.1, D82A has a y(N) value of 0.921 indicating that its domain movements can be classified as a pure hinge motion. D82N has a hinge motion as well, whilst D82E seems to have a mixed type with both hinge and shear motions as indicated by a y(N) value of 0.456. However, this is likely to be caused by the flexibility of a few side chains which are not well defined due to the poor electron density and thus affect the determination of the contact changes.

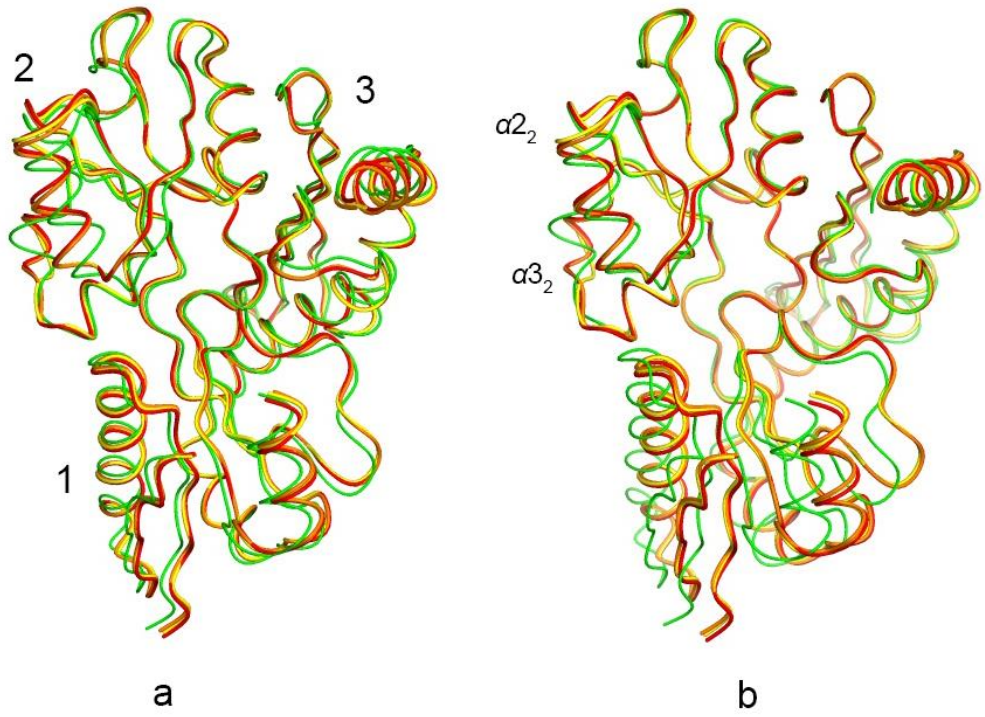


Figure 3.16 Superposition of BPBGD WT and mutant structures. a) Secondary structure superposition of the WT (green), D82A (red), D82N (yellow) and D82E (orange) BPBGD structures. b) Least-squares superposition of all the structures by domains 2 and 3 indicating a clear movement of domain 1 and some local movements of $\alpha 2_2$-helix and $\alpha 3_2$-helix in domain 2 [Figure from (Guo *et al*., 2017a)].

Table 3.3 Superposition *RMSD* values, rotation angles and translation distances between equivalent domains of the WT and mutant BPBGDs [Table from (Guo *et al*., 2017a)].

Mutant	*RMSD* (Å)				Rotation (°)	Translation (Å)
	All C$^{\alpha}$	Domain 1	Domain 2	Domain 3		
D82A	1.12	0.430	0.87	0.49	7.4	0.2
D82E	1.04	0.37	0.86	0.49	7.2	0.1
D82N	1.07	0.42	0.89	0.44	7.2	0.2

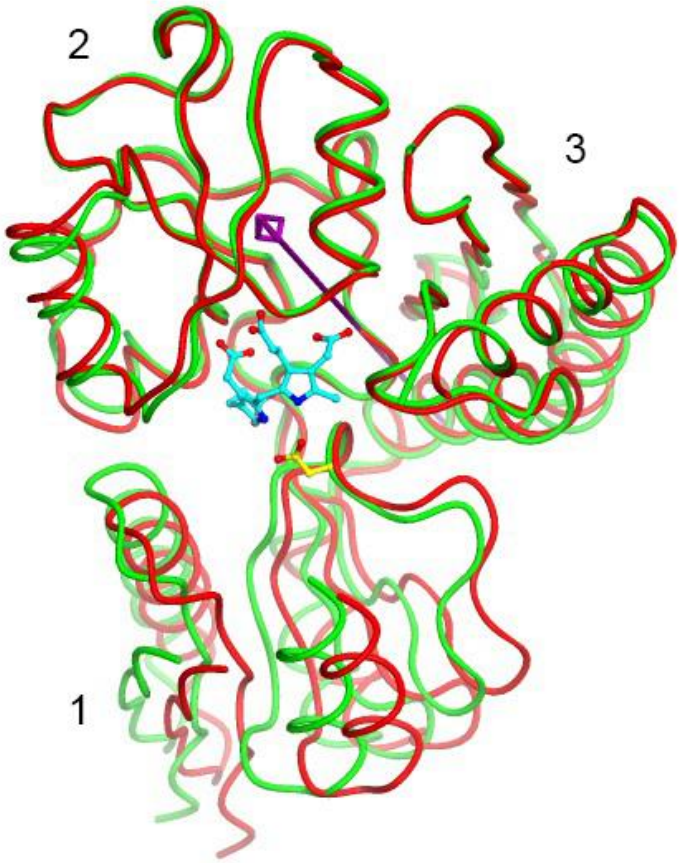


Figure 3.17 The inter-domain screw axis. Domain 1 rotates and translates along this axis which results in the 'open' (D82A, red) and 'closed' (WT, green) conformations of the protein [Figure from (Guo *et al*., 2017a)].

In addition, there is no structural difference between the D82N crystal obtained at a lower pH (5.7 in droplet) and the one obtained at a higher pH (7.0 in droplet) which suggests that the domain movement is probably caused not by the lower pH, but by the mutation itself. This is consistent with Awan *et al.* (1997) who suggested that the *E. coli* PBGD crystal structure determined at pH 5.1 is likely to be similar to the structure at physiological pH values.

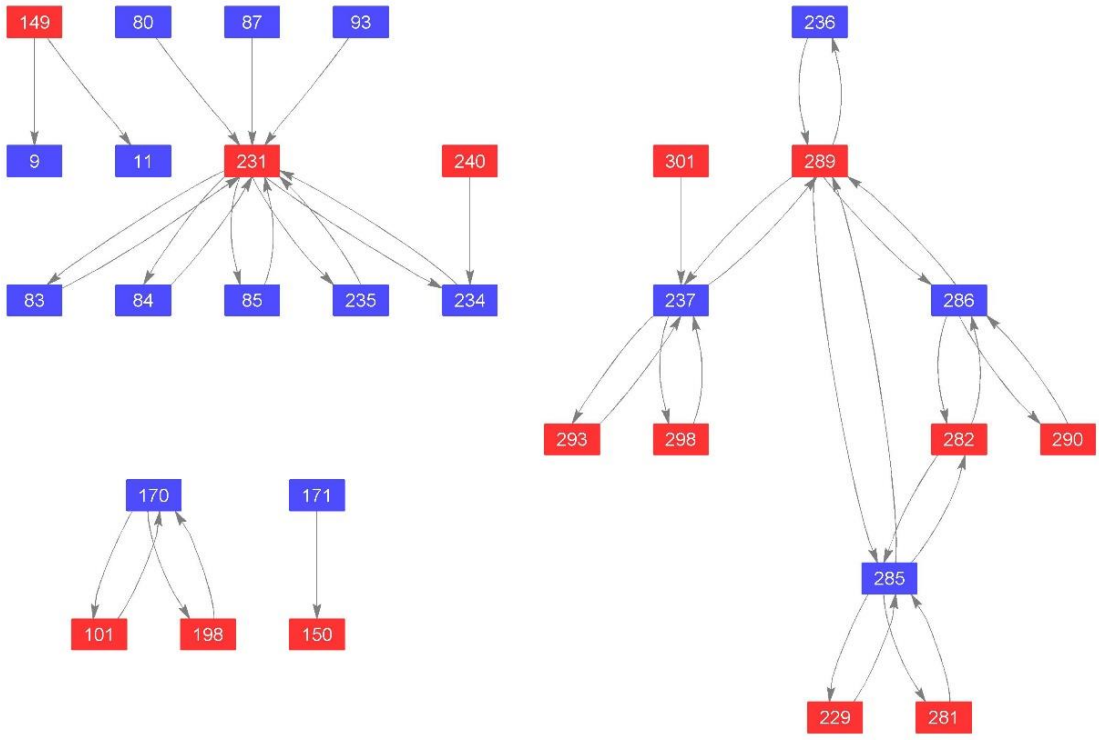


Figure 3.18 The Dynamic Contact Graphs for the WT and D82A BPBGD structures. A blue square corresponds to a residue in domain 1 and a red square corresponds to a residue in domains 2&3 with the residue number labelled in the middle.

3.4.6 BPBGD kinetic assay

The substrate PBG is converted to preuroporphyrinogen by PBGD at a suitable pH. The product preuroporphyrinogen is highly unstable and will be converted to uroporphyrinogen I in the absence of the next enzyme uroporphyrinogen III synthase or converted to uroporphyrinogen III in the presence of this enzyme. Both uroporphyrinogen I and III can be oxidised to form uroporphyrin by the addition of benzoquinone. Uroporhyrin has a λ_{max} at 405.5 nm and produces a pink fluorescence under UV light (Figure 3.19) (Shoolingin-Jordan *et al.*, 1997).

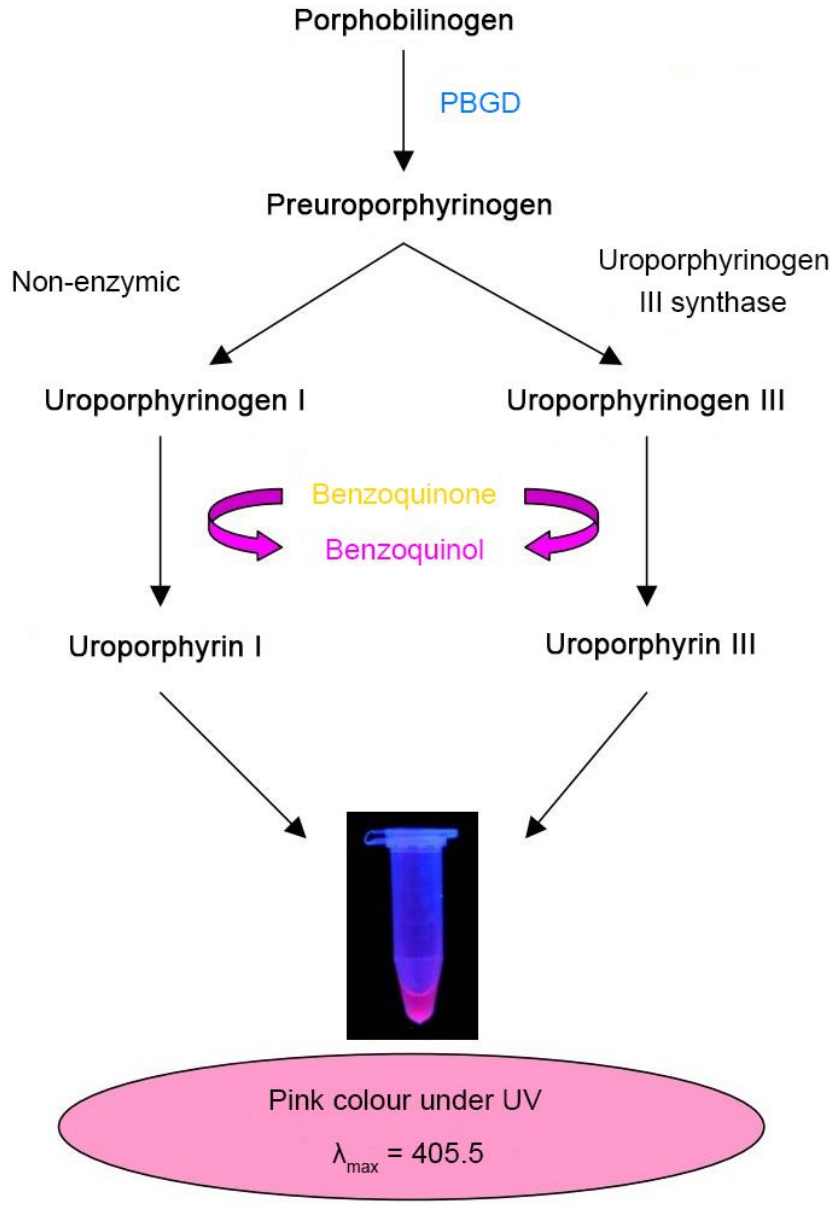


Figure 3.19 Mechanism of the PBGD kinetic assay. Without the presence of uroporphyrinogen III synthase, the assay follows the pathway shown on the left, in which the preuroporphyrinogen is converted to uroporphyrinogen I which is then oxidised to uroporphyrin I that has a λ_{max} at 405.5 nm. The reaction follows the pathway on the right if uroporphyrinogen III synthase is added.

A range of PBG concentrations from 0.05 to 2.5 mM were assayed with both the WT and the mutant enzyme. Samples with higher concentrations of PBG turned pink after incubation with benzoquinone on ice in the dark for 20 min. The colour of the mutant samples disappeared after the addition of sodium metabisulphite while the colour of the WT samples did not, indicating the existence of porphyrin. The concentration of product was calculated from the $OD_{405.5}$ values according to the Beer-Lambert law (equation 3.2) and the K_M was calculated according to the Michaelis – Menten equation (equation 3.3). Figure 3.20 shows the kinetic curves for the WT and the D82E enzymes, which have K_M values of 1.38 μM and 7.71 μM and k_{cat} values of 5.8×10^{-2} s^{-1} and 6.57×10^{-5} s^{-1}, respectively. In conclusion, the D82E mutant showed a significantly reduced activity while the D82A and the D82N mutants were completely inactive.

$$a = \varepsilon c l \tag{3.2}$$

$$V = \frac{V_{max}[S]}{K_M + [S]} \tag{3.3}$$

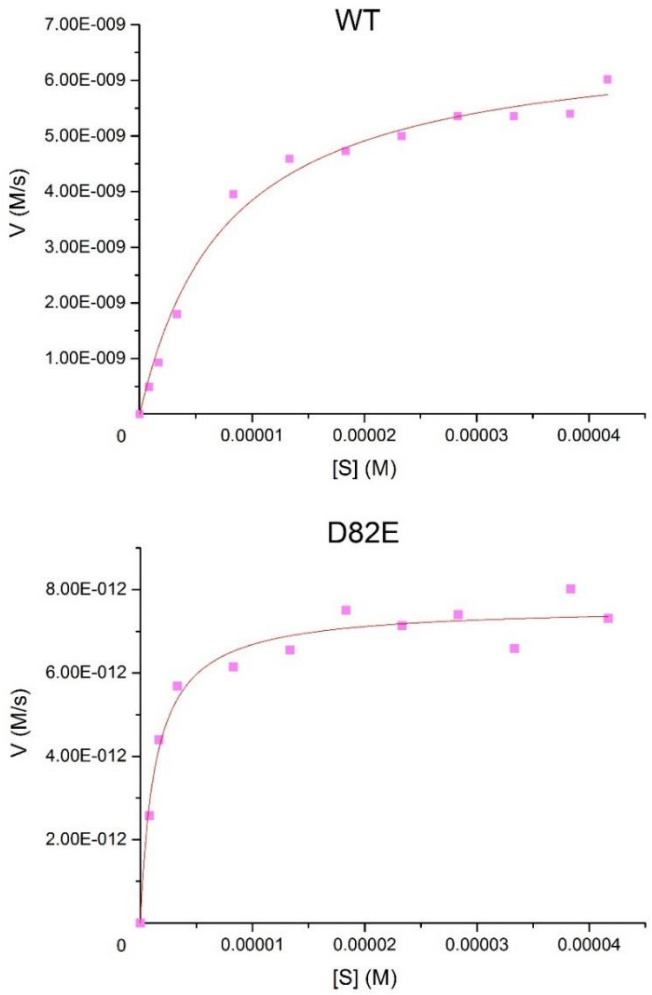


Figure 3.20 Michaelis–Menten kinetic curves for the WT and D82E BPBGDs.

3.5 Discussion

3.5.1 Domain movements

Domains of multi-domain proteins are often linked by flexible regions that allow them to move relative to one-another. Since the catalytic mechanisms of many enzymes can be dependent on domain movements, static information that can be obtained by comparison of domains in open and closed crystallographic structures, such as those of ligand-bound and ligand-free forms or WT and mutant structures, is of great importance. Characterisation of domain movements

is important for understanding how a protein folds or reorganises its domains to attain its functional state and how it functions once it is there.

When considering individually, superposition of WT BPBGD with the three mutant structures shows that there are small differences in domains 1 and 3 with *RMSD* values around 0.4 Å. In contrast, there are larger differences between them in domain 2 with *RMSD* values of all above 0.85 Å. This could be because of the local movements in the second and third *α*-helices ($\alpha 2_2$ and $\alpha 3_2$) (Roberts *et al.*, 2013) which can be found in all of the mutant structures and affect the shape of the active site cleft (Figure 3.16b). All of the *RMSD* values are appreciably lower than those obtained when the crystallographic structures are superposed as a rigid compilation of the three domains, which give *RMSD* values of all above 1.0 Å (Figure 3.16a, Table 3.3).

By comparison of domains 2 and 3 in the *A. thaliana* and *E. coli* PBGD structures, it revealed that the two domains move in a concerted manner with respect to domain 1 (Roberts *et al.*, 2013). The same effect could also be seen when comparing the structures of WT and mutant *B. megaterium* enzyme. In addition, these mutant structures exhibit a marked local movement of *α*-helices $\alpha 2_2$ and $\alpha 3_2$ which further opens the active site cleft. The differences in the relative domain orientation are emphasized most clearly when the WT and mutant structures are superposed by domains 2 and 3, as shown in Figure 3.16b. This is also confirmed by a separate analysis using the protein domain motion website, *DynDom*, which automatically defines domains 2 and 3 as a fixed unit and domain 1 as the moving partner. Specifically, domain 1 has a rotation of more than 7.0 degrees about the inter-domain axis in all the mutant structures compared with the WT enzyme which leads to an opening of the active site cleft (Figure 3.21a). To date, as far

as we are aware, this is the largest domain rotation that has been reported for this enzyme family. It is stated by Chasles' theorem that any rigid body displacement is a screw motion in which there is a translation along an axis followed by a rotation about the axis. In this case, there is quite a small translation of only 0.1 Å or 0.2 Å in the mutants which means that the domains move mainly in a rotation mode or as a hinge motion (Gerstein *et al.*, 1994). Compared with more complicated shear or mixed motion, hinge movements are relatively simple and allow parts of the respective domains to approach each other perpendicular to the plane of the domain interface. This hinge motion of BPBGD has also been confirmed quantitatively using Dynamic Contact Graphs (Taylor *et al.*, 2013; Taylor *et al.*, 2014). Residues 99, 100 and 198, which reside in the linker region between domain 1 and 2, are recognised by *DynDom* as bending residues and therefore their flexibility could be key to the opening and closure of the domains during catalysis.

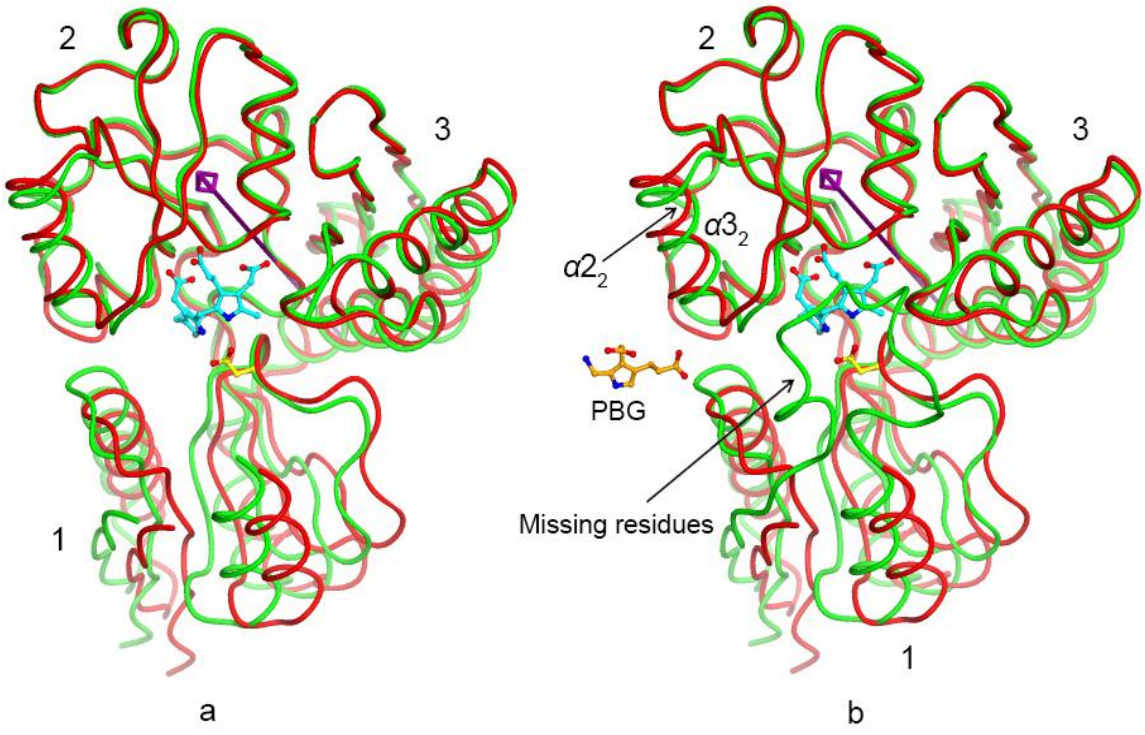


Figure 3.21 BPBGD domain movements. a) Superposition of the WT (green) and D82A (red) structures via domains 2 and 3 emphasises the hinge movement of domain 1 about the screw axis which is shown in purple. b) A similar superposition made using a model of the missing residues in the active site flap of BPBGD which were built according to the structure of the *A. thaliana* enzyme. This suggests a possible route for substrate (shown as ball-and-stick model on the left) to gain access to the active site when the structure is in the more open state that is apparently adopted by the mutants [Figure from (Guo *et al*., 2017a)].

The amino acid residues from 40 to 60 are missing in the BPBGD WT and mutant structures. In the structure of the *A. thaliana* enzyme, these residues form two *α*-helices and loop regions which altogether cover the active site of the protein (Roberts *et al.*, 2013). The absence of this region in the *B. megaterium* enzyme could be due to proteolysis caused by the addition of thrombin to remove the His-tag or by protease impurities during storage of the enzyme or crystallisation. It is reasonable to assume that this region of the enzyme is more ordered *in vivo* since it is highly conserved and forms many contacts with the DPM cofactor. Hence,

we modelled the missing residues of the *B. megaterium* enzyme based on the plant PBGD structure, as shown in Figure 3.21b. If this model is correct, then the movement of domains 2 and 3 away from domain 1 (which is also enhanced by the local movement of the helix $\alpha 2_2$ in domain 2) could be critical to let the substrate gain access to the active site. Once the substrate has gained entry, it would be quite close to both the free α-position of the C2 ring of the cofactor and the catalytic residue Asp82 of the enzyme, so that condensation can occur. Further substrate moieties can then get into the active site and bind in a large cavity close to the catalytic aspartate such that they become connected, one by one to ultimately form a hexamer.

Although the structures of the three mutants are similar to each other, they do have some slight differences. In general, when they are superposed by domains 2 and 3, there are small movements in domain 1. The D82E mutant has the most closed conformation while the D82A mutant has the most open one, leaving D82N as the intermediate of the three mutant structures.

3.5.2 Active site and cofactor electron density

The DPM cofactor is covalently attached to Cys241 in domain 3 in the WT BPBGD by a thioether link. It has been shown that there are a number of conserved amino acid residues, mainly arginine residues, which bind the acidic side chains of the cofactor through many ionic interactions (Wood *et al.*, 1995). Additionally, the catalytic residue, Asp82 in domain 1, forms hydrogen bonds with both pyrrole N atoms. Hence the cofactor is tightly held by these residues in the WT structure.

Inspection of the electron density in the active site of the mutant structures shows that there is no electron density for the whole of the cofactor in the D82A mutant. In contrast, there is good density for only the C1 ring of the cofactor in the D82N mutant, but in the D82E mutant there is good electron density for both of the cofactor rings (Figure 3.13). This suggests that the whole cofactor in the D82A mutant and the C2 ring of the cofactor in the D82N mutant are missing, or at least highly flexible. This was also consistent with the result of Ehrlich's reaction, which can be used to determine the DPM cofactor status in the active site of the enzyme.

In the D82E mutant structure, the active site arginine residues are still in almost the same position as in the wild structure, keeping the cofactor firmly in place. In the D82N mutant structure, the arginine residues (129, 130 and 153) which are close to the side chain of the C1 ring can still form strong interactions (donor-acceptor distances around 2.75 Å) with the C1 ring, but those which are close to the C2 ring in the WT structure are all moved back to some distance (> 3.15 Å), perhaps making it difficult for the C2 ring of the cofactor to bind firmly, as its lack of electron density would suggest. In addition to this, there are two conserved water molecules in the WT and D82E structures which mediate 4 hydrogen bonds between Arg174, Gln17 and the side chain of the C2 ring of the cofactor. These two water molecules are missing in the D82N structure which makes it impossible to form these hydrogen bonds. A similar trend can be seen in the D82A structure which completely lacks the cofactor and has the most open of all the mutant structures.

Enzyme kinetic studies showed that the D82A and D82N mutants were completely inactive. In contrast, the D82E mutant enzyme was found to react much slower compared with WT enzyme. This mutant has a K_M of 7.71 μM and

a k_{cat} of 6.57×10^{-5} s^{-1} which compares with values of 1.38 μM and 5.8×10^{-2} s^{-1}, respectively, for WT BPBGD. This is consistent with the structural information that has been obtained showing that the D82E mutant has the whole cofactor bound in a well-ordered manner, while the others either do not assemble the complete cofactor or are not able to bind it sufficiently tightly to allow the electron density for both rings to be resolved.

3.6 Future work

1. Woodcock *et al.* (1994) found that the mutant PBGD enzymes could bind to the substrate PBG to form stable enzyme-substrate complexes. These complexes could be separated by ion-exchange chromatography (Woodcock and Jordan, 1994). Previous studies on co-crystallisation of the substrate and mutant enzymes failed to get crystals of any of the complexes. It would be worth trying to get stable complexes first and then crystallise them directly to see if their structures could be obtained.

2. In the substrate analogue *α*-bromoporphobilinogen, the reactive *α*-position of porphobilinogen is blocked by a bromine atom (Figure 3.22). Hence this compound acts as a chain terminator and suicide substrate in the polymerisation reaction. It has been used to investigate the catalytic mechanism of PBGD during the tetrapolymerisation reaction. When *α*-bromoporphobilinogen was incubated with PBGD, not only was the enzyme inactivated, but the Ehrlich's reaction was also inhibited. It reacts with all of E, ES, ES_2 and ES_3 and produces EBr, ESBr, ES_2Br and ES_3Br, respectively (Figure 3.23), where Br indicates bromoporphobilinogen. All of these complexes can also be isolated (Warren and Jordan, 1988). Thus, future work on this project includes synthesis of *α*-

bromoporphobilinogen as well as *α*-iodoporphobilinogen and trying to get the complex structures to understand how the substrate moieties bind to the active site in the different intermediates.

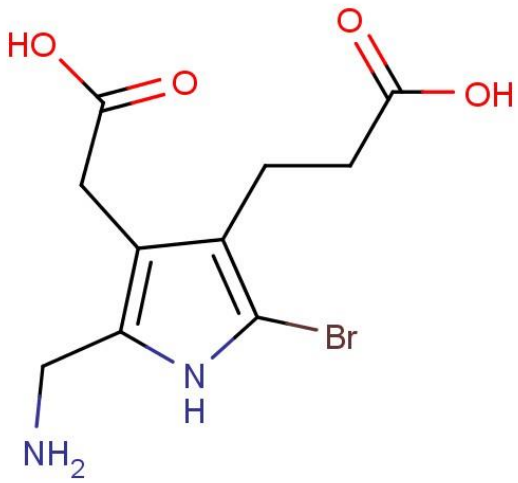


Figure 3.22 Structure of *α*-bromoporphobilinogen. The otherwise free *α*-position of the pyrrole is occupied by a bromine atom which blocks the subsequent binding of the next PBG molecule.

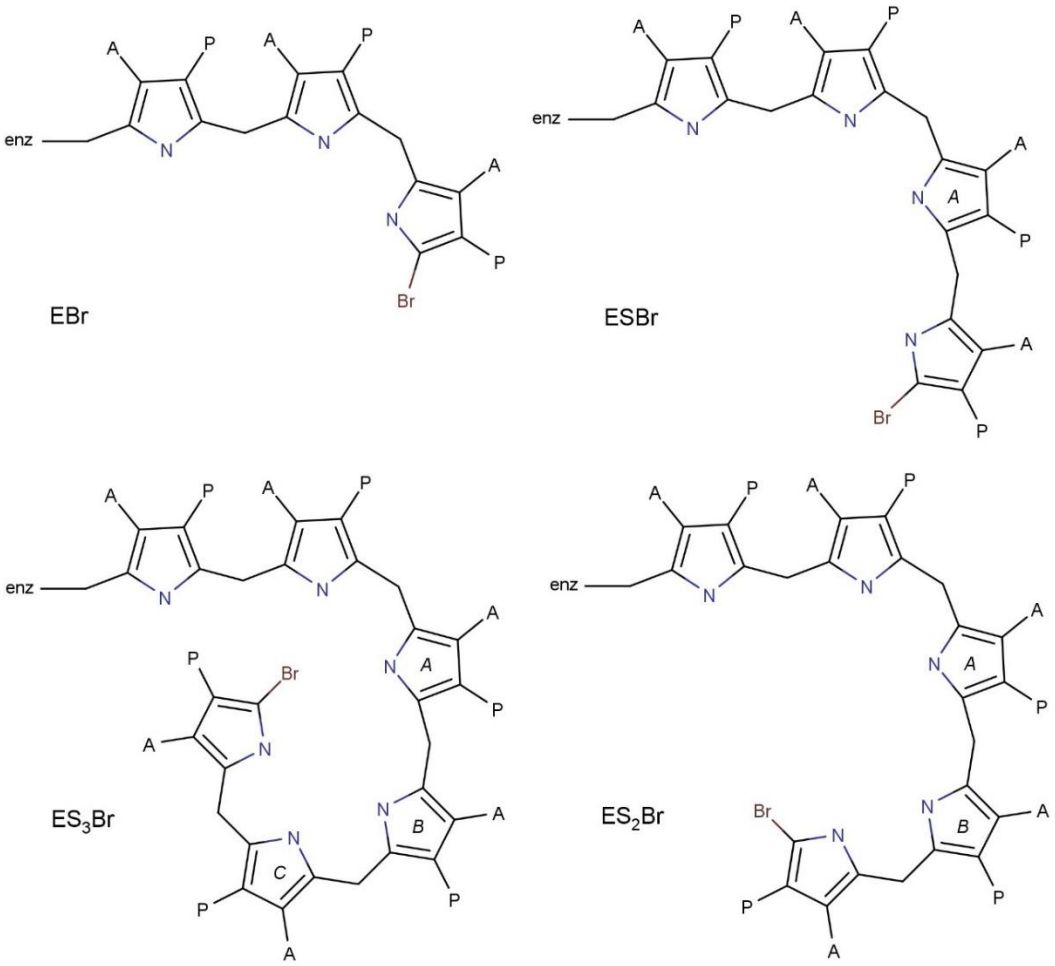


Figure 3.23 Complexes formed between PBGD and *α*-bromoporphobilinogen. EBr, ESBr, ES_2Br and ES_3Br are formed through inhibition of the reaction of PBGD with the substrate PBG by *α*-bromoporphobilinogen at different stages.

Chapter 4

Structural studies of a Kunitz-type potato cathepsin D inhibitor

This chapter is based on the work published by Guo *et al.* (2015).

4.1 Introduction

4.1.1 Protease inhibitors

Proteases regulate the synthesis, activation and turnover of proteins. Uncontrolled proteolysis is harmful to cellular functions and is related to many diseases such as cancer (Bell-McGuinn *et al.*, 2007), Alzheimer's disease (De Strooper, 2010), emphysema (Greene and McElvaney, 2009) and pancreatitis (Kitagawa and Hayakawa, 2007). Nature has developed many strategies to control proteolysis, one of which is to produce polypeptides known as protease inhibitors (PIs). PIs inhibit the activity of their target proteases by forming protease-inhibitor complexes through their reactive-site loops and the active site of the target proteases (Rakashanda *et al.*, 2012). PIs work as antibiotics and pesticides by targeting against hydrolytic enzymes of pests and microbes to prevent their growth, which is adopted by many plants (Satheesh and Murugan, 2011). PIs are particularly abundant in plant storage tissues including fruits, seeds and tubers, for example, 50% of the total weight of the soluble proteins in potato juice is contributed by PIs (Pouvreau *et al.*, 2001) and their expression is increased under environmental stress or during wound healing. PIs prevent plant seeds from being digested in a predator's gut which acts as a defensive mechanism. In addition, it has been suggested that they also have anti-tumour effects in colon, breast, skin and prostate cancers (Correa, 1981; Li *et al.*, 2009).

4.1.2 Classification of PIs

Based on the amino acid sequence homology, structure, reactive-site location, disulphide bridges and the proteases they inhibit, PIs can be classified into eight

distinct families known as: Kunitz, cereal, squash, potato, mustard, cystatin, Kininogen, and Bowman-Birk (Habib and Fazili, 2007). Kunitz-type protease inhibitors are mostly active against serine proteases but can also inhibit aspartic and cysteine proteases (Oliva *et al.*, 2010). These inhibitors usually have a molecular mass of around 20 kDa and have a low cysteine content forming two disulphide bridges. The 3-dimensional structure is characterised with a *β*-trefoil composed of 12 antiparallel *β*-strands with long protruding loops, some of which act as the reactive sites for proteases (Figure 4.1). Single-headed Kunitz-type inhibitors (Khamrui *et al.*, 2005) have only one reactive site while double-headed inhibitors (Azarkan *et al.*, 2006) have two reactive sites which can bind two target molecules (Figure 4.1). Kunitz-type protease inhibitors bind tightly to their targets in competitive and non-competitive ways and the complexes dissociate fairly slowly (Migliolo *et al.*, 2010; Ritonja *et al.*, 1990). The reactive site of Kunitz-type trypsin inhibitors often possesses a Lys or Arg residue.

Protease inhibitors can also be placed into four classes based on the targets they inhibit, known as serine protease inhibitors, aspartic protease inhibitors, cysteine protease inhibitors and metallocarboxy protease inhibitors. The serine protease inhibitor (SPI) family is the largest one among all the four families. Most SPIs have a molecular mass between 3 and 25 kDa and can inhibit trypsin and/or chymotrypsin. The two best characterised SPI families are the Kunitz-type and Bowman-Birk type inhibitors. They have different molecular masses, cysteine contents and numbers of reactive sites compared with other families (Richardson, 1977) and their structures are stabilised mainly by hydrophobic interactions of short stretches of hydrogen bonded sheets and/or disulphide linkages.

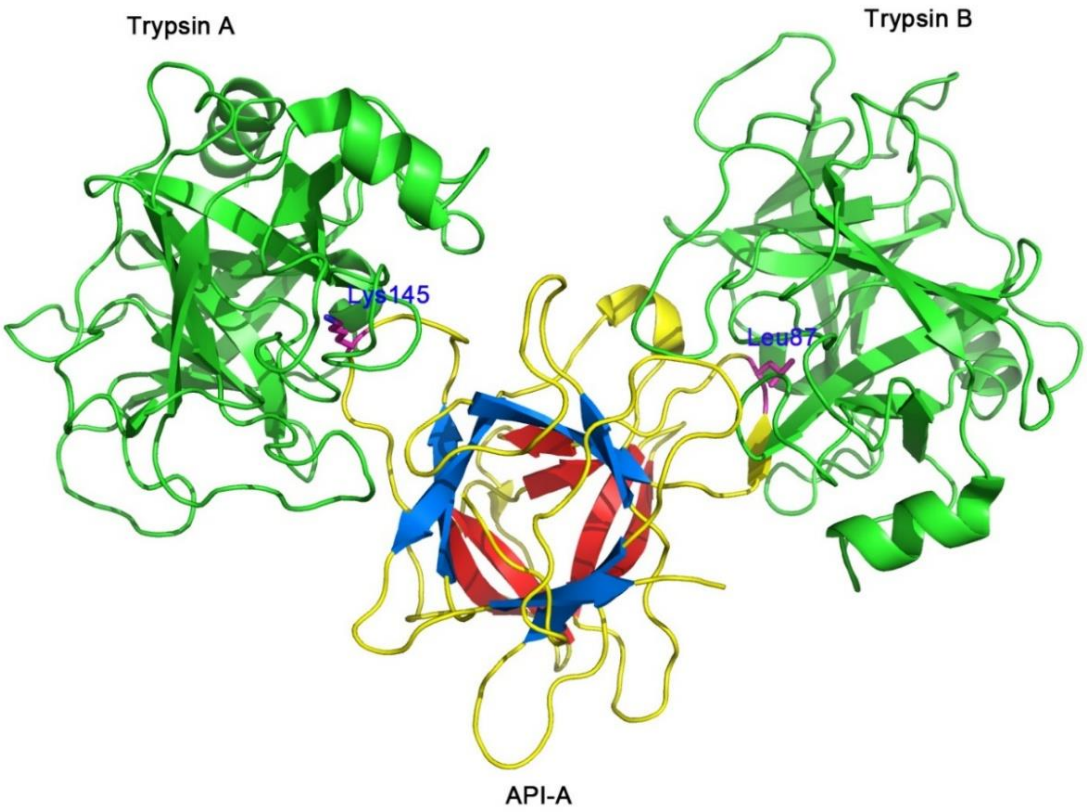


Figure 4.1 Structure of a PI-protease complex. The tertiary structure of a Kunitz-type double-headed arrowhead protease inhibitor A (API-A) complexed with two trypsin molecules (PDB ID: 3E8L) (Bao *et al.*, 2009). API-A adopts a *β*-trefoil composed of 12 antiparallel *β*-strands, 6 of which form the 'bottom' barrel (blue) covered by a 'cap' (red) composed by the other 6 strands. API-A binds to the two trypsin molecules through the two reactive site residues Lys145 and Leu87.

Aspartic protease inhibitors (APIs) contain only the Kunitz-type family. They are characterised by a molecular mass of 20-22 kDa and two disulphide bridges. It has been proved that APIs inhibit cathepsin D and may inhibit trypsin and/or chymotrypsin as well (Park *et al.*, 2005). APIs are rare in nature and have been identified in plants such as tomato (*Solanum lycopersicum*) (Werner *et al.*, 1993), potato (*Solanum tuberosum*) (Guo *et al.*, 2015) and wheat (*Triticum aestivum*) (Galleschi *et al.*, 1993).

4.1.3 Cathepsin D

Cathepsin D (EC 3.4.23.5) belongs to the aspartic protease family and is ubiquitously distributed in lysosomes. Like other aspartic proteases, the catalytic process of cathepsin D depends on the protonation of the catalytic Asp residues in the active site and it can accommodate up to 8 amino acid residues in the active site. It is involved in the degradation of intracellular and internalized proteins, activation and degradation of polypeptide hormones and growth factors, it also participates in brain antigen processing and regulates programmed cell death (Benes *et al.*, 2008).

4.1.4 Potato cathepsin D inhibitor

Potato cathepsin D inhibitor (PDI), isolated from *Solanum tuberosum*, is one of the natural cathepsin D inhibitors. It belongs to the Kunitz-type protease inhibitor family and has been reported to inhibit both bovine cathepsin D and trypsin with K_i values of 3.8×10^{-7} M and 8.6×10^{-9} M, respectively (Keilova and Tomášek, 1976a; Keilova and Tomášek, 1976b; Mareš *et al.*, 1989). It is highly homologous to many known serine protease inhibitors including potato serine protease inhibitor (PSPI, PDB ID: 3TC2) (Meulenbroek *et al.*, 2012) and soybean trypsin inhibitor (STI, PDB ID: 1AVX) (Song and Suh, 1998). Previous studies showed that PDI is a single polypeptide consisting of 188 amino acids with a molecular mass of 20.6 kDa, and has high pH- and thermal-stability (Keilova and Tomášek, 1976a; Keilova and Tomášek, 1976b). PDI is expressed in flower buds and potato tubers, and the expression is induced during wounding and treatment of leaves with the plant hormones abscisic acid and jasmonic acid (Hannapel, 1993; Herbers *et al.*, 1994). Transgenic plants, which have been modified to over-

express protease inhibitors, have shown improved resistance to pests, presumably due to the anti-nutritional effects on predatory insect larvae (Schluter *et al.*, 2010; Srinivasan *et al.*, 2005).

There are six cysteine residues in PDI which form three disulphide bridges (Mares *et al.*, 1989). It is a glycoprotein which contains 2 $GlcNH_2$ residues and the $-NH_2$ group of the side chain of Asn19 acts as the attachment point for the carbohydrate moiety to form an Asn-linked oligosaccharide (Kornfeld and Kornfeld, 1985).

Several isoforms of PDI have been identified from the genome sequence study of potato (Consortium, 2011; Hirsch *et al.*, 2014) and they share a high sequence identity of above 90%. The isoform reported here is encoded by the aspartic protease inhibitor 8 gene, which also encodes an N-terminal pre-sequence (31 residues) composed of a predicted endoplasmic reticulum (ER) secretion signal and a putative N-terminal vacuolar sequence determinant (NLIDL) separated by a cleavage site between residues 18 and 19. It is likely that the Kunitz-type protease inhibitors are primarily stored in the storage vacuoles of potato tuber cells (Jørgensen *et al.*, 2011) and are finally secreted into the apoplast or extracellular space, where they inhibit proteases produced by predatory microbes in order to protect the plant cell wall (Jashni *et al.*, 2015). The closest homologue of PDI, which shares 93% amino acid sequence identity, is encoded by the aspartic protease inhibitor 5 gene (Ritonja *et al.*, 1990). It shows a much stronger inhibition of yeast aspartic protease A than cathepsin D (Cater *et al.*, 2002).

4.2 Project Aim

The aim of the project was to determine the crystal structure of PDI and then identify the reactive sites for cathepsin D and trypsin based on the structure.

4.3 Methods

4.3.1 Crystallisation

A freeze-dried powder of PDI was a gift of Prof. Vladimir Kostka (Institute of Organic Chemistry and Biochemistry, Prague, Czech Republic) and had been purified as described by Keilova and Tomášek (1976b). It was dissolved in 50 mM Tris, 100 mM NaCl pH 7.3 to a final concentration of 5 mg/ml. Screening of crystallisation conditions was performed with the same kits as described in section 2.3.1 by use of the hanging-drop method with a TTP Labtech Mosquito crystal screening robot (TTP Labtech, Hertfordshire, UK). All the 96-well plates were stored at 21 °C and crystals started to appear after 7 days. Two crystals were harvested for data collection with crystal 1 (Figure 4.2a) being obtained in 2.0 M NaCl and 10% PEG 6,000 and crystal 2 (Figure 4.2b) being obtained in 0.2 M $MgCl_2 \cdot 6H_2O$, 0.1 M Tris pH 8.5 and 20% PEG 8,000. These crystals were mounted in loops with 30% glycerol as a cryo-protectant and were flash-cooled by a nitrogen gas cryostream (Oxford Cyrosystems Ltd., Oxford, UK) prior to data collection.

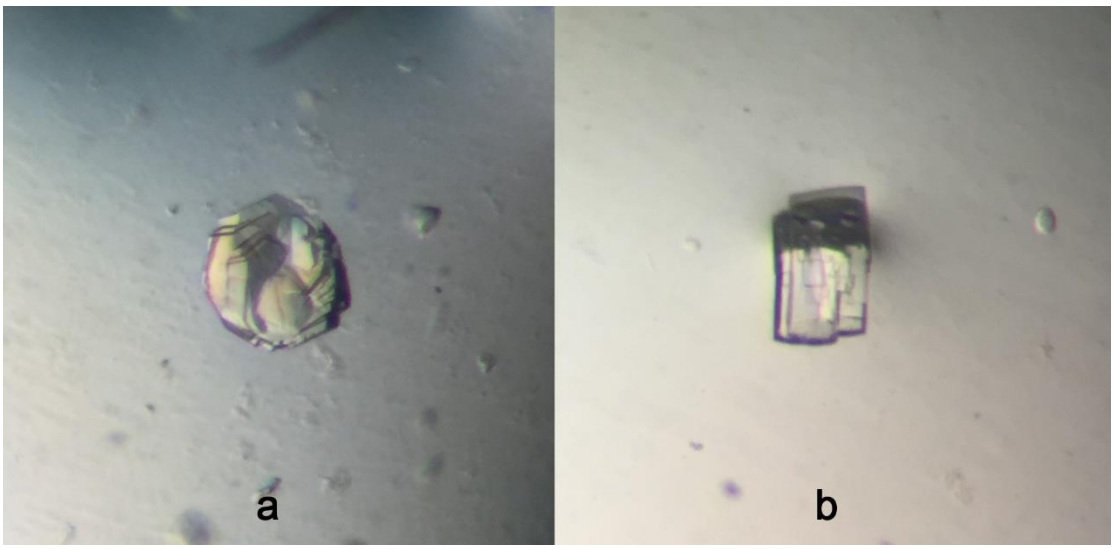


Figure 4.2 Crystals of PDI. a) The cluster containing crystal 1 which belonged to the space group *C*2. b) The cluster containing crystal 2 which belonged to the space group $C222_1$. Both of the crystals were separated into single pieces for data collection. [Figure originally from (Guo *et al*., 2015)].

4.3.2 Data collection, data processing, structure determination and validation

X-ray data collection was undertaken at station I04-1 at DLS. Crystals 1 and 2 diffracted to a resolution of 2.1 Å and 2.8 Å, respectively. Automatic data processing, performed at DLS, using *xia*2 (Winter, 2010) identified that crystal 1 was monoclinic and belonged to the space group *C*2 while crystal 2 was orthorhombic and belonged to the space group $C222_1$. The solvent contents, calculated with *Matthews_coef* (Matthews, 1968), suggested that both crystals had two molecules in the ASU. The recently solved structure for PSPI (PDB ID: 3TC2, 73% sequence identity with PDI) (Meulenbroek *et al.*, 2012) allowed the determination of the *C*2 structure by molecular replacement using the program *Molrep* (Vagin and Teplyakov, 2010) in the CCP4 suite (Winn *et al.*, 2011). The refined *C*2 structure was then used as the search model for the analysis of the

$C222_1$ structure by molecular replacement. Manual rebuilding and introduction of sugar moieties were carried out by use of the program *Coot* (Emsley and Cowtan, 2004). The structures were refined using *Refmac*5 (Murshudov *et al.*, 2011) and validated with *MolProbity* (Chen *et al.*, 2010). All the statistics for data processing and refinement are shown in table 4.1.

Table 4.1 X-ray statistics for the two PDI structures. Values in parentheses are for the outer resolution shell [Table from (Guo *et al*., 2015)].

	Crystal 1	Crystal 2
Beamline	I04-1 (DLS)	I04-1 (DLS)
Wavelength (Å)	0.9173	0.9173
Space group	*C*2	$C222_1$
Unit-cell parameters (Å)		
a (Å)	75.5	70.9
b (Å)	124.8	119.5
c (Å)	37.9	131.1
β (°)	95.1	90.0
Mosaic spread (°)	0.33	0.15
Resolution (Å)	27.90-2.11 (2.17-2.11)	29.70-2.83 (2.90-2.83)
R_{merge} (%)	7.3 (96.9)	15.6 (78.2)
R_{meas} (%)	10.0 (132.8)	18.2 (91.5)
$CC_{½}$ (%)	99.6 (31.3)	99.3 (83.8)
Completeness (%)	96.0 (70.8)	99.5 (95.2)
Average $I/\sigma(I)$	8.3 (0.8)	8.8 (2.9)

Multiplicity	3.3 (2.5)	6.9 (6.9)
No. of observed reflections	63,269 (2,588)	93,388 (6,614)
No. of unique reflections	19,145 (1,024)	13,609 (963)
Wilson plot *B*-factor (Å^2)	42.1	36.6
Solvent content (%)	44.6	64.6
R_{factor} (%)	19.4	19.3
R_{free} (%)	27.0	26.8
RMSD bond lengths (Å)	0.013	0.013
RMSD bond angles (°)	1.704	1.785
No. of reflections in working set	19,858	13,652
No. of reflections in test set	985	732
Mean protein *B*-factor (Å^2)	47.5	42.0

4.4 Results and discussion

4.4.1 Quality of the model

These two crystals diffracted to medium and low resolutions and have been well refined as indicated by reasonable *R*-values and *RMSD* values. Most residues in the models fit well in the electron density map contoured at 1.0 *RMSD* while residues 145-155 of chain B in crystal 1, which are in a loop region, are missing probably due to the flexibility. The Ramachandran plot showed that the models of crystal 1 and 2 have 97.5% and 98.1% of the total residues in the allowed region. Similar to the other Kunitz-type protease inhibitor structures (Krauchenco *et al.*, 2003; Song and Suh, 1998), all the loop regions including the reactive-site loops are relatively more flexible, which is indicated by high *B*-factors. This inherent high flexibility allows the loops to access the active sites of the target

proteases more easily. The 40.3 Å mean protein *B*-factor of crystal model 1 is slightly higher than expected for a resolution of 2.1 Å since it has been reported that PDI is very stable (Keilova and Tomášek, 1976a). This is presumably due to the less-than-ideal crystal quality as shown in Figure 4.2.

4.4.2 Overall structure

Each of the two structures analysed in different crystal forms contain two molecules in the ASU. The four molecules in both structures are quite similar in the core region, which is composed of 12 *β*-strands. Around 95% of the residues have an *RMSD* value of the C^{α} atoms of around 0.2 Å in any pair-wise superposition. In contrast, the *RMSD* values of C^{α} atoms of the loop regions between any two molecules are all above 0.8 Å, with the highest value of 1.1 Å between chain A of crystal 1 and chain A of crystal 2. In fact, some of the loops, such as the one from Arg154-Phe158, adopt different conformations in different structures and different chains.

The overall structure of PDI is shown in Figure 4.3 which shows a typical *β*-trefoil fold common to several other Kunitz-type protease inhibitors such as PSPI and STI. The PDI structure consists of 12 antiparallel *β*-strands that form six pairs of double-stranded *β*-hairpins. Three of these hairpins (*β*1-*β*12, *β*4-*β*5 and *β*8-*β*9) form a short antiparallel *β*-barrel that is closed by a triangular-like “lid” at one side consisting of the remaining three hairpins (*β*2-*β*3, *β*6-*β*7 and *β*10-*β*11). The 12 *β*-strands are connected by long protruding loops. The *β*-trefoil is characterised by three structural repeats around a pseudo-threefold axis of symmetry which coincides with the barrel axis. Each repetition unit is a four-stranded motif that contains about 60 amino acid residues. There are three disulphide bridges in PDI

structure known as: S1, Cys48-Cys93, S2, Cys142-Cys159 and S3, Cys150-Cys153. The S1 and S2 disulphide bridges are also found in STI, which lacks S3. The S2 and S3 disulphide bridges apparently reduce the flexibility of the large loop from Cys142 to Ala160, which is much smaller and has fewer residues in other Kunitz-type inhibitors. There is a 3_{10}-helix at positions 91 to 96 which has also been found in STI (84-86) (Song and Suh, 1998), but not in this region in winged bean chymotrypsin inhibitor (WCI) (Ravichandran *et al.*, 2001) and E*rythrina* trypsin inhibitor (ETI, PDB ID: 1TLE) (Onesti *et al.*, 1991). This 3_{10}-helix protrudes from the centre of PDI and STI and can be cleaved by subtilisin, which is a serine protease, through the bond between Met84 and Leu85 in STI (Laskowski *et al.*, 1974). It is possible that this region is the reactive-site loop for serine proteases like trypsin. One molecule of N-acetylglucosamine (NAG) has also been found connecting to residue Asn19 in all chains in the two crystal structures except for chain B in crystal 1. One molecule of carbonate was fitted in the map between Asp35 and Gln186 in chain A, crystal 1.

The overall structure of PDI is similar to the Kunitz-type serine protease inhibitors. The structure of several of these inhibitors have been reported including PSPI, RASI (rice bifunctional alpha-amylase subtilisin inhibitor, PDB ID: 2QN4, unpublished work), STI and WCI. PSPI has the highest amino acid sequence similarity of 73% to PDI. The other three protease inhibitors have sequence similarities of 32%, 26% and 26%, respectively. Figure 4.4 shows the sequence alignment of PDI with these protease inhibitors. Their amino acid sequences are more similar in the skeleton part, which is composed of the 12 *β*-strands, compared with the loop regions which contribute to the specificity of the inhibitors. Structural superposition also shows that the similarity is mostly in the *β*-strand

scaffold, with the loops being different among each other. There is statistical evidence that the formation of the *β*-trefoil fold is due to a double gene duplication which makes it possible for these Kunitz-type protease inhibitors to have a common ancestor (Ponting and Russell, 2000).

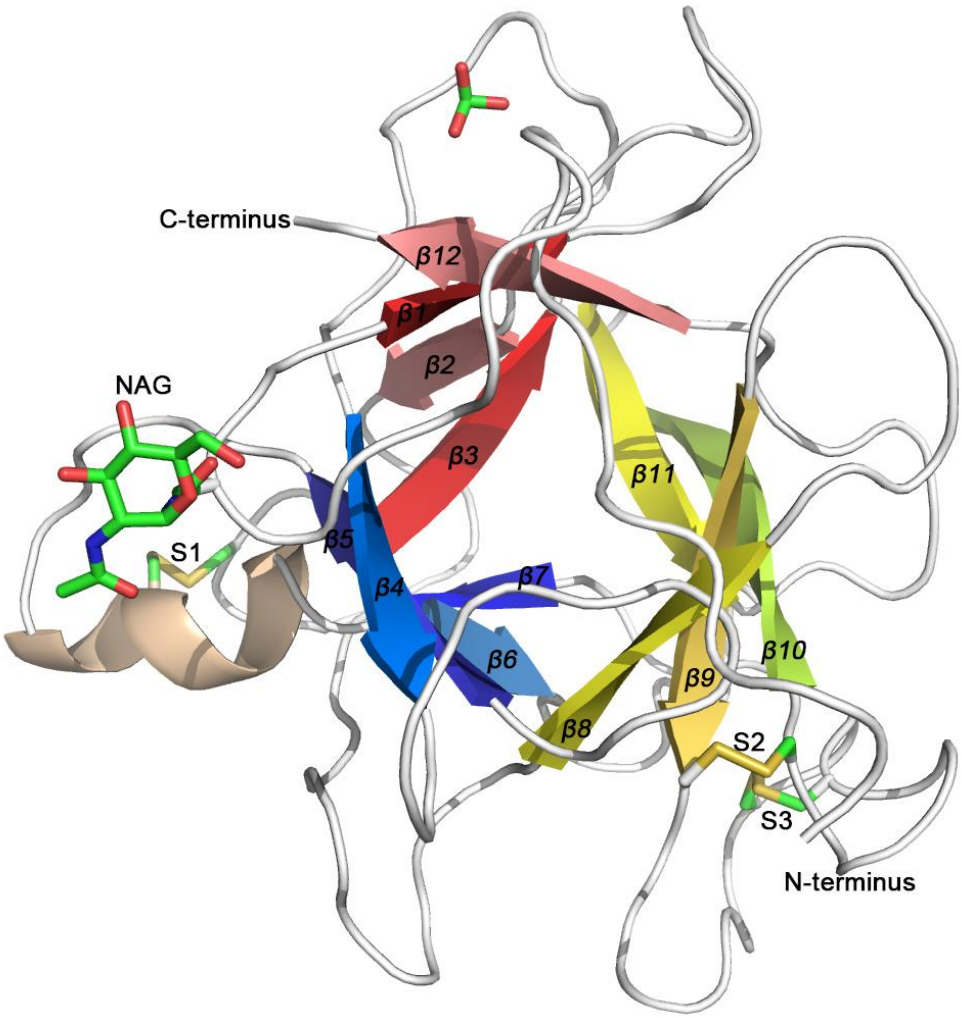


Figure 4.3 The overall structure of PDI. The cartoon diagram shows the *β*-trefoil view from the bottom side to the lid. The three disulphide bridges are labelled as S1, S2 and S3 and the disulphide bonds are coloured as yellow. [Figure originally from (Guo et al., 2015)].

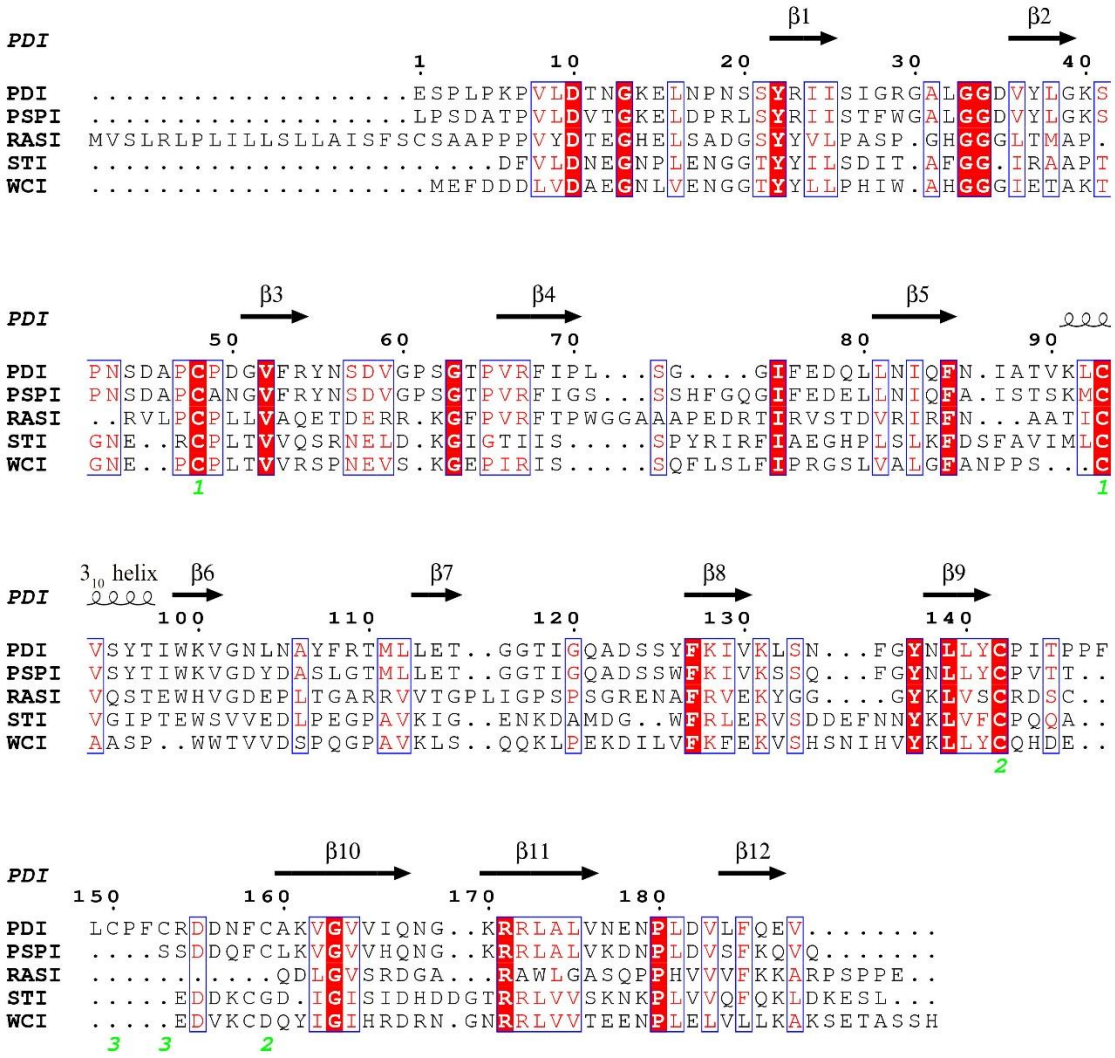


Figure 4.4 Sequence alignment and the secondary structure characteristics of PDI with several other Kunitz-type serine protease inhibitors. Beta-strands are labelled according to the PDI structure and are indicated by black arrows. All the conserved residues are boxed and the fully conserved residues are coloured in white with a red background, while the not fully conserved residues are coloured in red. The three disulphide bridges are labelled in green numbers. Alignment was performed using *ESPript* 3.0 website (Gouet *et al.*, 2003; Robert and Gouet, 2014). [Figure originally from (Guo et al., 2015)].

4.4.3 Identification of reactive-site loops

4.4.3.1 Reactive-site loops for trypsin

The putative reactive-site of PDI for trypsin, as described in *UniProt* (accession number P16348), is around residues Arg67 and Phe68 which was predicted by sequence alignment with STI. However, inspection of PDI structure indicated that these two residues are located in the middle of a *β*-strand which is not exposed, suggesting that other reactive-sites need to be considered. In order to find out possible reactive-site loops for trypsin, the PDI structure was superposed with several other Kunitz-type serine protease inhibitor-trypsin complexes including STI, WCI and API-A (Bao *et al.*, 2009) using *PyMOL*.

Two possible loops were identified with the first one being the loop from residues Leu71 to Leu80 (Figure 4.5 region 1), which is the most common reactive-site loop for Kunitz-type serine protease inhibitors. Arg63 and Ile64 have been discovered to be the P1 and P1' residues in this loop in the structure of STI. Similarly, Leu65 and Ser66 are the P1 and P1' residues in the equivalent loop in WCI (PDB ID: 1EYL). By comparison of several Kunitz-type trypsin inhibitors, it has been shown that the P1-P1' residues are often a combination of two residues from Lys, Arg, Leu, Ile and Ser in the following pairs: Lys-Ile (API-A), Arg-Ile (STI), Leu-Ile (API-A), Leu-Ser (WCI) and Arg-Ser (ETI). Although there is Leu71-Ser72 in the 71-80 loop of PDI, it is close to the *β*4-strand, making it less likely to access the active site of trypsin. However, the Gly74-Ile75 were identified as the P1 and P1' residues because it has been reported in other Kunitz-type trypsin inhibitors that the relatively conserved Arg or Lys were replaced by a Gly or Glu which gave a pair of Gly-Ile or Glu-Ile (Srinivasan *et al.*, 2005). These two residues reside in

the middle of the loop that adopts the canonical conformation of the Kunitz-type serine protease inhibitors. This would facilitate its recognition by a target protease.

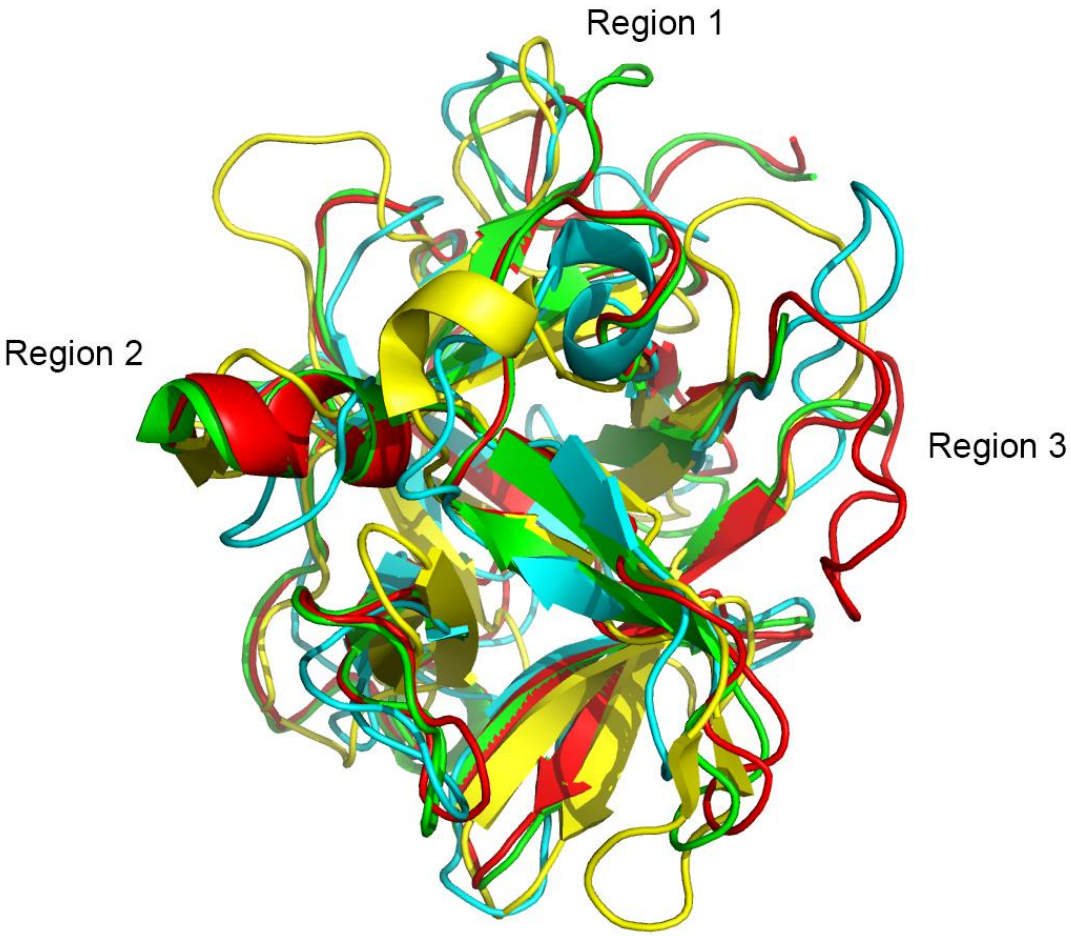


Figure 4.5 Structural superposition. Superposition of PDI structure (red) with the structures of PSPI (green), WCI (cyan) and API-A (yellow). The three predicted reactive-sites for trypsin and/or cathepsin D are labelled as regions 1, 2 and 3.

A second reactive-site loop for trypsin was identified by structural superposition with API-A. Despite the fact that API-A shares a sequence similarity of only 11.8% with PDI, their overall structures are very similar including the 3_{10}-helix (Figure 4.5 region 2) and the three disulphide bridges are at equivalent positions including the 3rd one which is not common in Kunitz-type serine protease inhibitors. API-A is a double-headed serine protease inhibitor which can bind two trypsin molecules

at the same time. Superimposing of PDI and API-A revealed the second possible reactive-site loop which is from residues Phe85 to Ile98 in PDI. The P1 and P1' residues are Lys91 and Leu92 which have their equivalent residues as Leu87 and Ile88 in API-A. As described before, Lys is one of the common residues in the P1 position while Leu is not common in P1'. However, it is still reasonable to have Leu87 in this position because of its similarity to Ile, which is common in P1' position. This loop lacks a typical canonical conformation of Kunitz-type serine protease inhibitors, but the large positively charged Lys91 is protruding out into the solvent and would fit well in the trypsin active site.

4.4.3.2 Reactive-site loops for cathepsin D

Since no structures of aspartic protease inhibitors have been reported, it is difficult to directly identify the reactive-site loops for cathepsin D by structural superposition. However, by comparison with other Kunitz-type protease inhibitors, the loop from Cys142-Ala160 (Figure 4.5 region 3) in PDI is found to be much larger than its equivalent parts in other protease inhibitors. This loop is too small to be a reactive-site loop in some other protease inhibitors such as WCI and ETI. It contains two disulphide bridges in PDI, which are S2 (142-159,) and S3 (150-153) as shown in Figure 4.3, with S3 not existing in most Kunitz-type protease inhibitors. So it is very likely that this loop is the reactive-site loop for cathepsin D. Since aspartic proteases tend to cleave the peptide bonds with hydrophobic and *β*-methylene groups, two possible pairs, Phe148-Leu149 and Phe152-Cys153, are identified as the P1-P1' residues which are all located in the middle of the loop. Actually, Phe152-Cys153 may be a better option because it can be recognised and cleaved by aspartic proteases but possibly cannot be released by the enzyme since Cys153 forms a disulphide bridge with Cys150. It has also

been found by sequence alignment that Phe152-Cys153 are highly conserved in all the aspartic protease inhibitors, whilst Phe148-Leu149 are less conserved. None of these residues are conserved in serine protease inhibitors.

A second possible reactive-site loop could also be the one from Phe85 to Ile98 containing the 3_{10}-helix and the Lys91 residue (Figure 4.5 region 2). Li *et al*. (2000b) reported that an aspartic protease was inhibited by a potent small protein IA_3 in which the Lys18 residue plays an important role in the inhibition mechanism. The positively charged $-NH_3^+$ group of IA_3 is within hydrogen bonding distance of the carboxyl oxygen of Asp32 of the aspartic protease and forms interactions with it. The inhibitor protein IA_3 also forms a near-perfect amphipathic *α*-helix during the inhibitory process (Li *et al.*, 2000a). The precursors of aspartic proteases have also been proved to possess a conserved Lys residue which occupies the position of the catalytically important water molecule and inhibits their activity by interacting with both of the Asp residues (Richter *et al.*, 1999). For cathepsin D specifically, inhibition by a Lys residue based on a similar mechanism was also found in a catalytically inactive form. The N-terminal part, which contains the Lys8 residue, is inserted into the active site cleft of cathepsin D at pH 7.5 and the enzyme activity is inhibited (Lee *et al.*, 1998). Figure 4.6 shows the docking result of PDI with human cathepsin D using the website *ClusPro* without restraints (Comeau *et al.*, 2004a; Comeau *et al.*, 2004b). The highly unconstrained docking did give a good result with the loop containing Lys91 going deeply in the active site of cathepsin D in one of the top solutions. The $-NH_3^+$ group of Lys91 is within good hydrogen bonding distance with the four oxygen atoms of the two aspartate groups (all around 2.6 to 2.8Å). Structural superposition of the complex with

human cathepsin D showed that the $-NH_3^+$ group occupies the position of the catalytic water molecule as well.

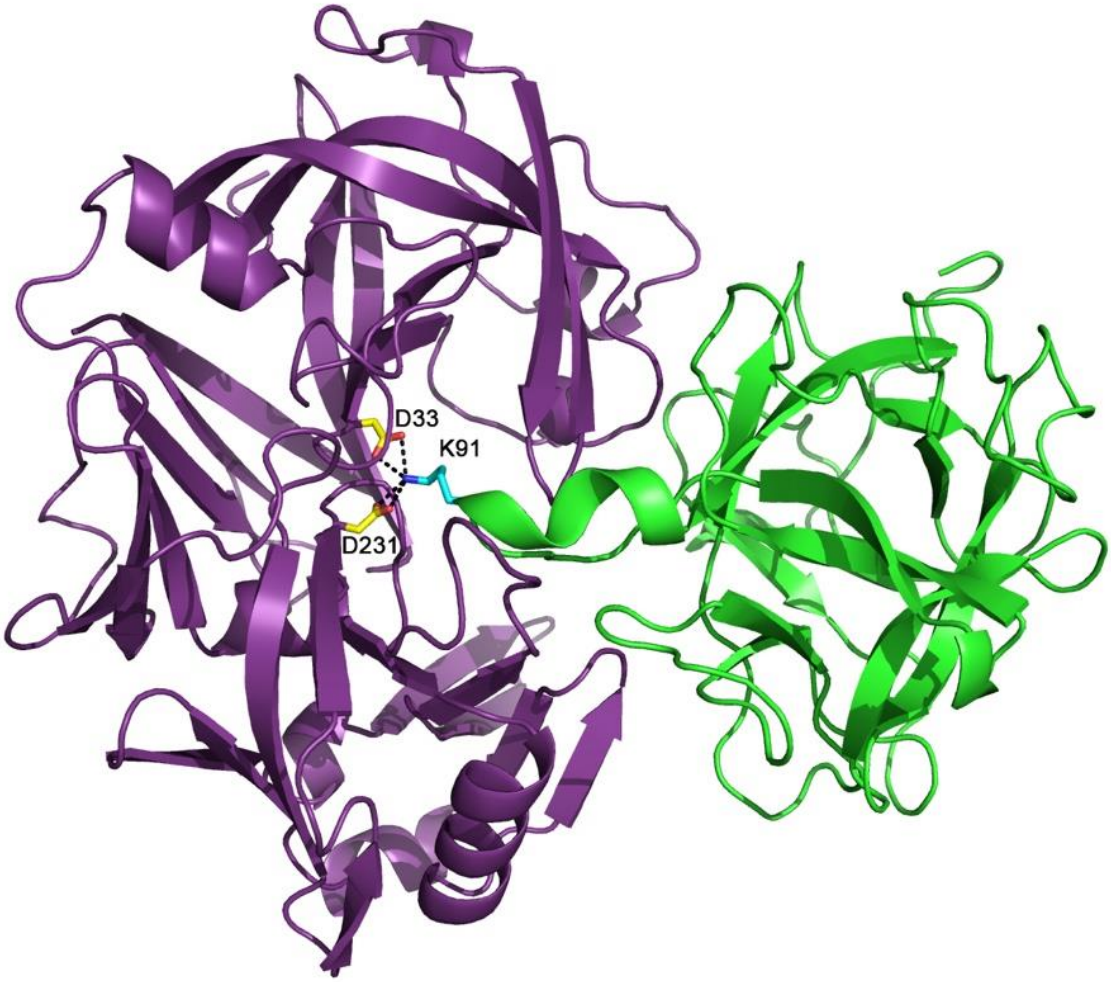


Figure 4.6 A docked model of PDI with human cathepsin D. PDI and cathepsin D are coloured as green and purple, respectively. The Lys91 (cyan) residue at the end of the 3_{10}-helix is sticking out from PDI core and going deeply into the active site of the cathepsin D. The $-NH_3^+$ group (blue) of Lys91 is within hydrogen bonding distance to the four oxygen atoms (red) of the two Asp residues (yellow). [Figure originally from (Guo et al., 2015)].

4.5 Summary

A very large proportion of the soluble proteins in plant tubers such as potatoes are protease inhibitors. These inhibitors provide a key defence mechanism

against pathogens and their expression is increased during wounding or other stress. Potato cathepsin D inhibitor belongs to the Kunitz-type protease inhibitor family and inhibits proteases in the aspartic- and serine-protease families. There are several isoforms of PDI in potato which share amino acid sequence identities as high as 90%.

The crystal structure of PDI shows that it adopts a *β*-trefoil fold with three disulphide bridges formed between cysteines 48-93, 142-159 and 150-153. The third one is very unusual for Kunitz family protease inhibitors, which indicates that it may be involved in the functioning of this molecule as a dual serine- and aspartic-protease inhibitor. Although the structure is mainly composed of *β*-strands, there is a short helical region from residues 91 to 96 which is a known reactive region in other serine protease inhibitors. PDI contains a likely recognition site for trypsin in this region which has been verified structurally in the API-A structure. The unique disulphide bridge in the long proline-rich loop indicates its possibility to be the reactive-site for cathepsin D, which is confirmed by the fact that this is a verified reactive site for a second trypsin molecule in the arrowhead inhibitor, the presence of several aromatic residues in this region, which is favourable for cathepsin D binding. However, the projecting Lys91 residue in the loop (residues 91-96) suggests its possibility to be another attractive candidate in view of the structurally-verified ability of the aspartic protease active site to interact electrostatically with polypeptides containing this basic residue. Docking studies demonstrate that this is a plausible model as well.

The 2.1 Å structural model and reflection file have been deposited in the PDB with the accession code: 5DZU.

Chapter 5

Structure of the family B DNA polymerase from the hyperthermophilic archaeon *Pyrobaculum calidifontis*

This chapter is based on the work published by Guo *et al* (2017c).

5.1 Introduction

5.1.1 *Pyrobaculum calidifontis* and Archaea

Archaea are prokaryotes like bacteria, but they also share similarities with eukaryotes. Many proteins involved in information storage and processing encoded in their genomes are eukaryotic-like, whilst their operational proteins are mostly homologous to bacterial proteins (Forterre *et al.*, 2002; Jain *et al.*, 1999; Koonin *et al.*, 1997). They are found in a broad range of habitats but some of them exist in extremely hot, salty, acidic, alkaline or high pressure environments.

Pyrobaculum calidifontis is a hyperthermophilic and facultative aerobic archaeon that was first isolated from a hot spring in Laguna, the Phillipines (Amo *et al.*, 2002). The Latin word *Pyrobaculum* means ‘fire stick’ and *calidifontis* means ‘hot spring’. The *P. calidifontis* cells are rod-shaped and are usually 1.5 to 10 μm long and 0.5 to 1 μm wide (Amo *et al.*, 2002). The organism has an optimal growth temperature between 90 °C and 95 °C, and an optimal pH of around 7 in aerobic conditions.

5.1.2 DNA polymerases and their classification

DNA polymerases (EC 2.7.7.7) function not only to replicate DNA but also to ensure the replication is accurate and correct. Following an annealed primer, DNA elongation is undertaken in the 5’ to 3’ direction with the catalysis by DNA polymerases. In addition to the polymerase activity, they can have 3’-5’ and/or 5’-3’ exonuclease activities which allow them to correct the wrong incorporation of nucleotides.

Based on the sequence homology and structures, DNA polymerases are classified into seven families including A (e.g. *E. coli* Pol I), B (e.g. *E. coli* Pol II), C (e.g. *E. coli* Pol III), D (e.g. *Euryarchaeota* Pol II), X (e.g. human Pol *β*), Y (e.g. human Pol η) and RT (e.g. reverse transcriptase). The C and D families are mainly replicative polymerases while the others are replicative and repair polymerases. Family B was given its name because it shares sequence homology with *E. coli* polymerase II which is encoded by the gene *pol*B. Family B DNA polymerases are found in bacteria, eukarya, archaea, viruses, plants and even in bacteriophages. They are highly accurate and perform 3'-5' proofreading of newly synthesised DNA so that any replication errors are corrected.

DNA polymerases, especially those from thermophilic organisms, are used extensively in biotechnological methods including PCR, gene sequencing, site-directed mutagenesis, DNA labelling and diagnostic purposes. The well-known *Taq* DNA polymerase (*Taq Pol*) was the first one to be characterised (Chien *et al.*, 1976) and was then applied in biotechnology. The *Taq Pol* lacks 3'-5' exonuclease activity which makes those with the ability, such as archaeal DNA polymerases, more popular of late because of their higher fidelity. The archaeal DNA polymerases are considered to represent the replication mechanism in higher organisms because they are in the same family as many of eukaryotic DNA polymerases.

5.1.3 The DNA polymerase from *P. calidifontis*

The gene encoding the DNA polymerase from *P. calidifontis* (Pc-polymerase) was cloned into a pET-21a plasmid and the protein was expressed and purified, as reported previously by Ali *et al*. (2011). The enzyme has a molecular mass of

approximately 90 kDa and is composed of 783 amino acids. It has an optimal temperature of 75 °C, a half-life of 4.5 hours at 95 °C and an optimal pH of 8.5. Thus it has greater thermostability compared with the widely-used *Taq* DNA polymerase. In addition, Pc-polymerase requires magnesium ions for activity and has been found to be inhibited by potassium and ammonium ions, while the *Taq* enzyme is dependent on monovalent cations. It is able to PCR-amplify larger DNA fragments of up to 7.5 kb. Similar to the DNA polymerases from *P. furiosus* (*pfu*) and *T. kodakarensis* (KOD1), Pc-polymerase is of high-fidelity, due to its 3'-5' exonuclease activity, which may have great commercial applications in future.

5.1.4 The structure of DNA polymerases

Many structures are available for DNA polymerases which show high conservation between them. The typical five-domain structure can be described as a 'right hand' with palm, fingers and thumb domains following the N-terminal domain which is followed by the 3'-5' exonuclease domain (Figure 5.1) (Brautigam and Steitz, 1998). The exonuclease domain is composed of anti-parallel *β*-strands with several catalytic carboxylate groups. The palm domain, which has a more conserved structure than the fingers and thumb domains, generally consists of 4 to 6 *β*-strands, two *α*-helices and three catalytic carboxylate residues. The fingers and thumb domains do not share much homology among different families. The fingers domain contains several residues whose side chains interact with the incoming deoxyribonucleoside triphosphate (dNTP). The thumb domain mainly constitutes parallel or anti-parallel *α*-helices and at least one of them makes contacts with the minor groove of the bound duplex. The thumb domain also has the capability of binding proteins known as processivity factors that encircle duplex DNA and stabilise the complex, thus

significantly improving the processivity of DNA polymerases. For example, the processivity of phage T7 polymerase is increased by at least 100 times with the help of the thioredoxin from the *E. coli* host cells, which acts as a processivity factor and sterically or electrostatically hinders the dissociation of the DNA (Brautigam and Steitz, 1998).

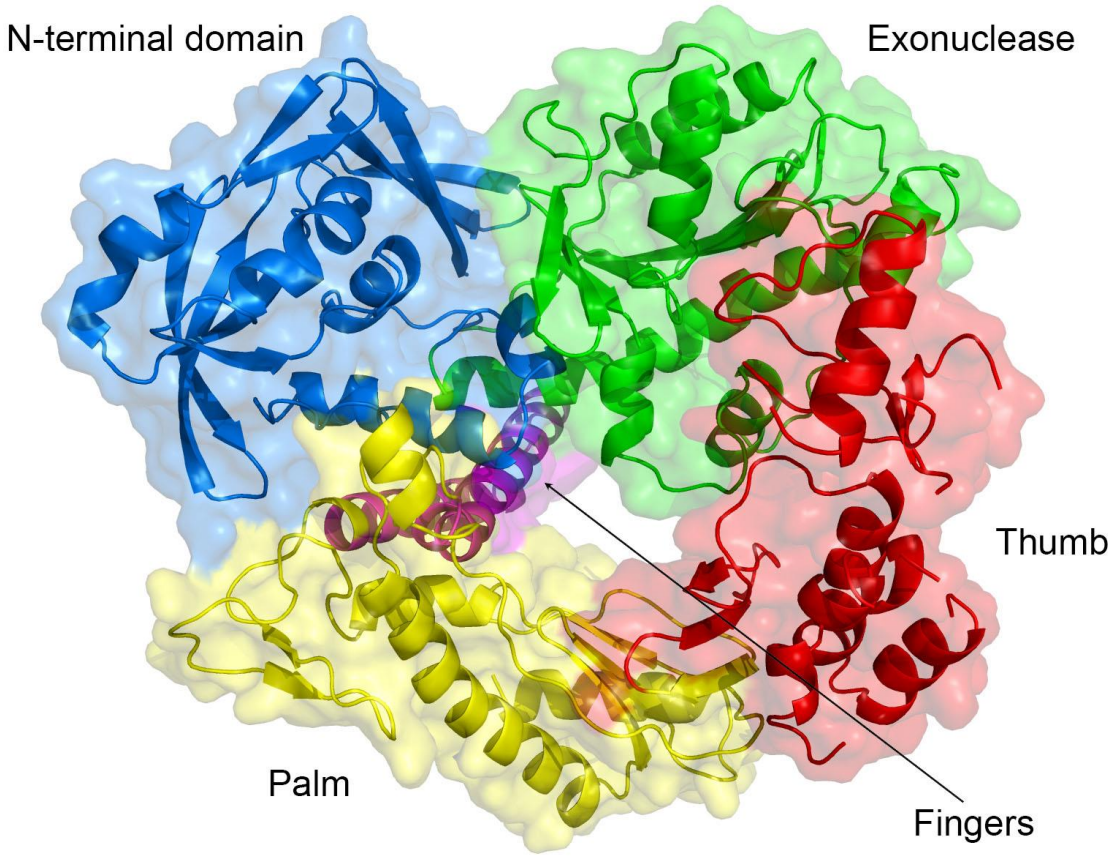


Figure 5.1 Crystal structure of the *Thermococcus* sp. 9 °N-7 DNA polymerase (PDB ID: 1QHT) (Rodriguez *et al.*, 2000). It is composed of five domains known as the N-terminal, 3'-5' exonuclease, palm, fingers and thumb domains which are coloured as blue, green, yellow, purple and red, respectively.

5.1.5 DNA polymerases in disease and as drug targets

Infrequent clonal somatic mutations have been identified in DNA polymerase genes in human tumours and some of them are deleterious. Mutations in a gene

that codes for DNA polymerase η have been found in patients with xeroderma pigmentosum variant (XPV) (Masutani *et al.*, 1999). Mutations of Pol *β* and γ genes were identified in colon adenocarcinoma (Starcevic *et al.*, 2004) and progressive external ophthalmoplegia (PEO) (Longley *et al.*, 2006), respectively.

Although DNA polymerases have a common catalytic mechanism and share extensive homology at the active site(s), several of them have been successfully targeted as drug targets. Nucleotide analogues, such as pyrimidine, purine and cyclic analogues, preferentially inhibit viral DNA polymerases and are used in treatment of infections caused by e.g. HIV (Argyris *et al.*, 2006) and cytomegalovirus (Cristofoli *et al.*, 2007). One of the important strategies in cancer treatment is to inhibit DNA polymerases that are involved in DNA repair to increase the efficiency of chemotherapeutic agents which damage DNA. For example, DNA synthesis may be effectively inhibited by drugs targeting the exonuclease proofreading or template switching involving the exonuclease and polymerase active sites.

5.2 Project aim

The aim of the project was to determine the crystal structure of Pc-polymerase, find out the unique structural features that distinguish it from other polymerases and investigate the reasons for its high thermostability.

5.3 Methods

5.3.1 Crystallisation

Expression and purification of the Pc-polymerase was described by Ali *et al*., (2011). Crystal screening was conducted by use of the hanging-drop method with

an enzyme concentration of 15 mg/ml in 20 mM Tris buffer pH 8.2. The same crystal screening robot and screening kits were used as described in section 2.3.1. The plates were stored at both 4 °C and 21 °C for the crystal growth. After four weeks, two crystals (Figure 5.2) were obtained in condition F4 of the Structure Screen 1+2 kit (0.2 M potassium thiocyanate, 0.1 M Bis-Tris propane pH 6.5, 20% PEG 3,350). However, following attempts for optimisation failed to reproduce any crystals at all. So the original two crystals were mounted in loops with 30% glycerol as the cryo-protectant and were flash-cooled before data collection.

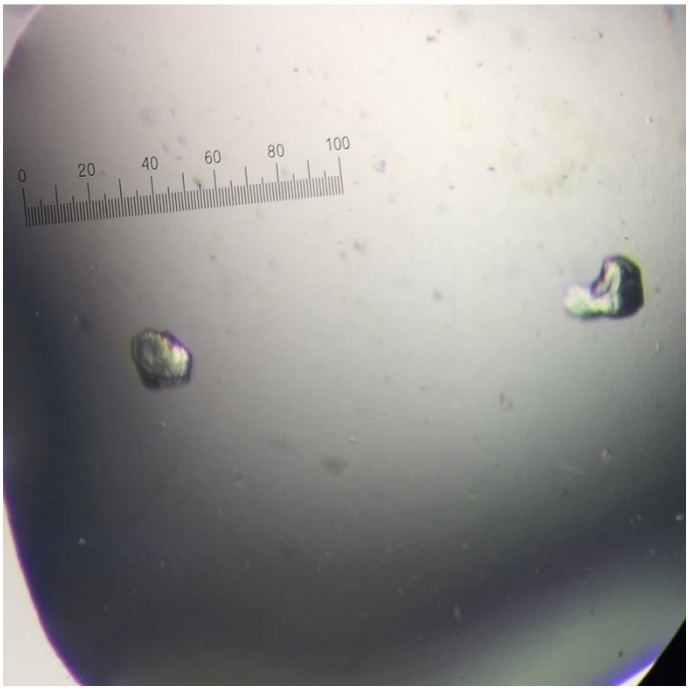


Figure 5.2 Pc-polymerase crystals. The crystals were obtained in the F4 condition from Structure Screen 1+2. They are approximately 200 microns in two dimensions and 50 microns thick. One small unit on the ruler is 10 microns.

5.3.2 Data collection and data processing

X-ray data collection was carried out at station I02, DLS. Data integration using the program *DIALS* (Waterman *et al.*, 2013) revealed that the crystals belonged

to the space group $P2_1$, which was confirmed by the program *Pointless* (Evans, 2006; Evans, 2011). Scaling and data reduction were performed using *Aimless* (Evans and Murshudov, 2013a), which showed that the better crystal diffracted to 2.8 Å. No twinning or translational NCS was present, as indicated by the program *Phenix.xtriage* (Zwart *et al.*, 2005). The solvent content was 50.7% with 2 molecules in the ASU, as estimated by *Matthews_coef* (Matthews, 1968).

5.3.3 Structure determination, refinement and further analysis

The initial raw structure was determined by molecular replacement with the *BALBES* website (Long *et al.*, 2008), which gave a top solution with a Q-factor of 0.67 and suggested that it was 99% likely to be a correct solution. The corresponding R_{factor} and R_{free} values were 37.6% and 44.1%, respectively. Four homologous structures were used as search models in the molecular replacement process. They were DNA polymerases from *T. gorgonarius* (PDB ID: 2XHB, 35.2% overall sequence identity) (Killelea *et al.*, 2010), *P. furiosus* (PDB ID: 3A2F, 36.0% overall sequence identity, to be published), *T. kodakarensis* (PDB ID: 1WN7, 35.2 overall sequence identity) (Kuroita *et al.*, 2005) and *Sulfolobus solfataricus* (PDB ID: 1S5J, 35.3% overall sequence identity) (Savino *et al.*, 2004). Since the raw model at this stage had a lot of residues fitted poorly in the electron density, many rounds of manual rebuilding were conducted by use of the program *Coot* (Emsley and Cowtan, 2004). This included the relocation of many 'shifted' residues that was guided by the electron density for large aromatic side chains. The following refinement using *Phenix.refine* (Afonine *et al.*, 2012) brought the R_{free} value down to 35.8%, which indicated that the model had been improved after the rebuilding. Following this, NCS refinement with torsion-angle restraints using *Phenix.refine* revealed more electron density in the last domain

of chain A and gave an R_{free} value of 32.7%. Further manual rebuilding, refinement and the introduction of metal ions and the 47 water molecules gave final R_{factor} and R_{free} values of 24.5% and 28.8%, respectively. All the statistics for data collection, data processing and refinement are shown in table 5.1. The *VADAR* (Willard *et al.*, 2003) and *ESBRI* (Costantini *et al.*, 2008) online services were used to analyse hydrogen bonds and salt-bridges and the sequence alignment was prepared with Alscript (Barton, 1993).

Table 5.1 X-ray statistics for the Pc-polymerase structure. Values in parentheses are for the high resolution shell.

Beamline	I02
Wavelength (Å)	0.9795
Space group	$P2_1$
Unit-cell parameters	
a (Å)	74.2
b (Å)	100.7
c (Å)	119.3
β (°)	94.7
Resolution (Å)	118.93-2.80 (2.90-2.80)
R_{merge} (%)	10.3 (77.0)
R_{meas} (%)	12.1 (91.2)
$CC_{½}$ (%)	99.1 (58.8)
Completeness (%)	97.5 (96.9)
Average $I/\sigma(I)$	6.8 (1.8)
Multiplicity	3.6 (3.7)

No. of observed reflections	154,369 (16,405)
No. of unique reflections	42,313 (4,403)
Wilson plot *B*-factor (Å^2)	66.8
Solvent content (%)	50.7
R_{factor} (%)	24.5
R_{free} (%)	28.8
RMSD bond lengths (Å)	0.004
RMSD bond angles (°)	0.818
No. of reflections in working set	42,206 (4,123)
No. of reflections in test set	2,132 (236)
Mean protein *B*-factor (Å^2)	73.5

5.4 Results and discussion

Pc-polymerase has several distinct features compared with other known DNA polymerase structures. Figure 5.3 illustrates the sequence alignment of Pc-polymerase with those from *T. gorgonarius*, *T. Kodakarensis* and *P. furiosus*. Although the secondary structure is conserved in all of them, Pc-polymerase has a relatively low sequence identity of approximately 37% with the other three, which have 80% to 90% sequence identity with each other. Indeed *P. calidifontis* belongs to the order *thermoproteales*, whilst the others are classified in the order *thermococcales*.

Figure 5.3 A structure-based sequence alignment of Pc-polymerase with the *T. gorgonarius* (*Tg*), KOD (*Tk*) and *pfu* (*Pf*) enzymes. The amino acid residues are coloured as follows: pale blue for basic; red for acidic; green for neutral polar; purple for hydrophobic; yellow for Cys; and white for the structurally important residues Gly, Ala and Pro. The catalytic metal-binding carboxylate residues are shaded grey in the exonuclease domains and are boxed red in the polymerase domain. The bottom row shows the secondary structure elements and are coloured blue, green, yellow, purple and red for the N-terminal, exonuclease, palm, fingers and thumb domains, respectively.[Figure from (Guo *et al*., 2017c), reproduced with permission of the Internation Union of Crystallography (http://journals.iucr.org)].

5.4.1 Structure of Pc-polymerase

The N-terminal domain of Pc-polymerase consists approximately of residues 1-163 and 360-390. It shows a bilobal feature which has an extra *β*-hairpin composed of 28 residues, compared with the *T. gorgonarius*, *pfu* and KOD DNA polymerases, inserted in the region which connects the exonuclease domain (Figure 5.4). A 7-stranded *β*-barrel is formed by the addition of this *β*-hairpin. The exonuclease domain extends from residues 164 to 360 that form a 7-stranded antiparallel *β*-sheet enclosed by *α*-helices. It has one fewer *β*-strand compared with the other enzymes, which are formed of 8 strands, due to the deletion of approximately 8 residues at the exposed outer edge of the protein. Several loop regions in this domain are quite different, with the largest being 6 Å apart from each other in superposition. There are four catalytic residues residing in this domain, known as Asp169, Glu171, Asp236 and Asp336, which are involved in binding two magnesium ions. Asp169 and Glu171 are located in the same *β*-strand which are closer to the magnesium ion, while the other two aspartates are located in helical segments.

The exonuclease domain is connected with the palm domain through a linker region (residues 391-404, coloured as grey in Figure 5.4) which is on the surface of the enzyme. It adopts a different conformation compared with other archaeal DNA polymerases which have an *α*-helix in this region. Structural superposition shows that it is approximately 10 Å apart from the equivalent parts in other archaeal enzymes. It is likely that this difference is due to the flexibility and exposed nature of this region which would form unfavourable contacts with symmetry related molecules if the conformation found in the homologous structures is adopted.

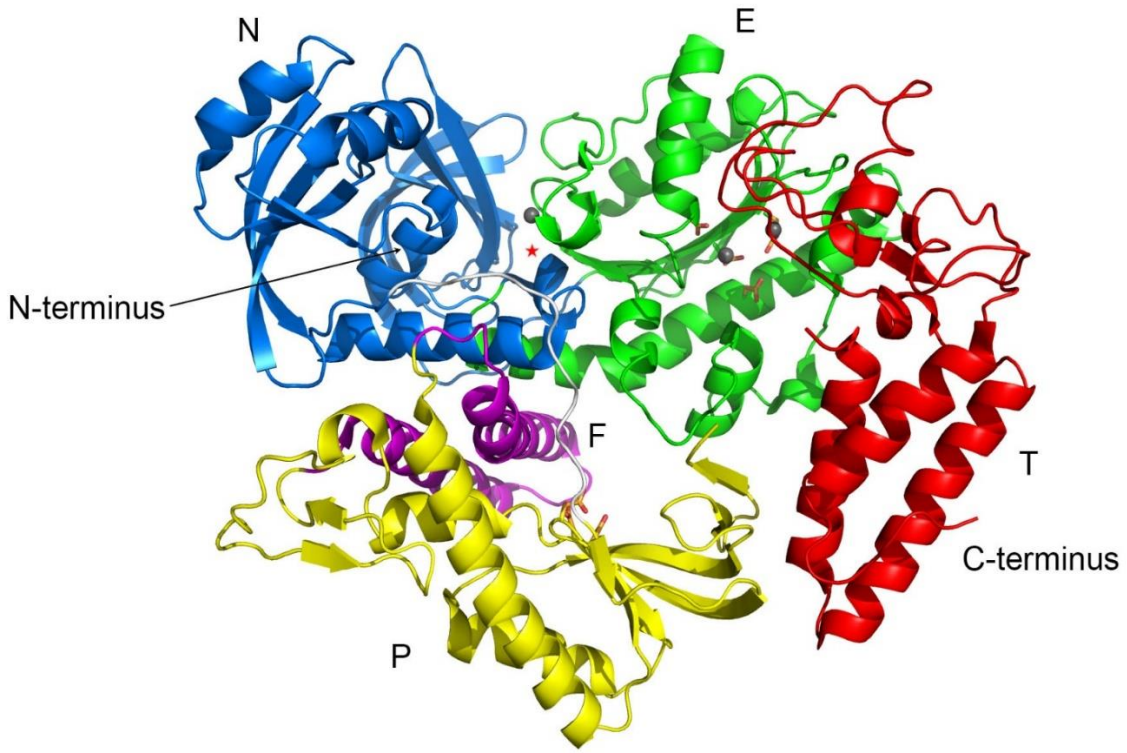


Figure 5.4 The overall structure of Pc-polymerase. The N-terminal, exonuclease, palm, fingers and thumb domains are coloured in blue, green, yellow, purple and red, respectively. The catalytic carboxylate side chains are coloured in orange and are represented as ball-and-stick with the three bound magnesium ions showing as dark grey balls. The small helix which plugs in the gap between the lobes of the N-terminal domain is indicated by a red star and the linker region, between the N-terminal domain and the palm domain, is coloured as grey.

The palm domain, which contains the active site for polymerase activity, is composed of two parts formed by residues 405 to 465 and residues 525 to 620 and are separated by the fingers domain. This first part contains a few small helical segments and *β*-strands while the second one consists of two large helices and a cup-shaped *β*-barrel formed by seven strands. The ‘triad’ formed by the three aspartates (Asp420, Asp560 and Asp562) was thought to be essential for catalysis in homologous structures (Delarue *et al.*, 1990), however, mutagenesis showed that the equivalent residue of Asp560 was not required by human Pol *α*

for catalysis (Copeland and Wang, 1993). It has also been proved by Wang *et al*. (1997) that only two of the three aspartates are invariant which correspond to Asp420 and Asp562 in Pc-polymerase. Both catalytic aspartates reside in adjacent strands at the same side which look like two teeth in the opening mouth of the cup. The electron density for the aspartate residues are well defined although there is no density for the catalytically essential magnesium ions.

The fingers domain (residues 466 to 524) is simply a helix-loop-helix structure which leans over the active site of the palm domain. The final thumb domain contains residues 625 to 766 and is composed by three helical features with several loops.

5.4.2 Comparison with *pfu* DNA polymerase

The Pc-polymerase has several significant structural differences compared with the *pfu* DNA polymerase. A structural superposition of both enzymes based on their secondary structures is illustrated in Figure 5.5. As mentioned in the previous section, there is an insertion of a *β*-hairpin at the region between the first two domains, which is not close to the active site. The largest difference in the exonuclease domain is the loop region formed by residues 173 to 186 which adopts a different conformation approximately 10 Å apart from that in *pfu* structure. There is short insertion in a *β*-hairpin formed by residues 244-247 in the *pfu* structure, which points towards the active site and may affect the catalytic properties. Also, the loop region from residues 391 to 404 in Pc-polymerase is largely different from its equivalent part in *pfu*, which has a helical segment in that region. The loops between residues 440 and 450, 490 and 494 in Pc-polymerase adopt different conformations as well.

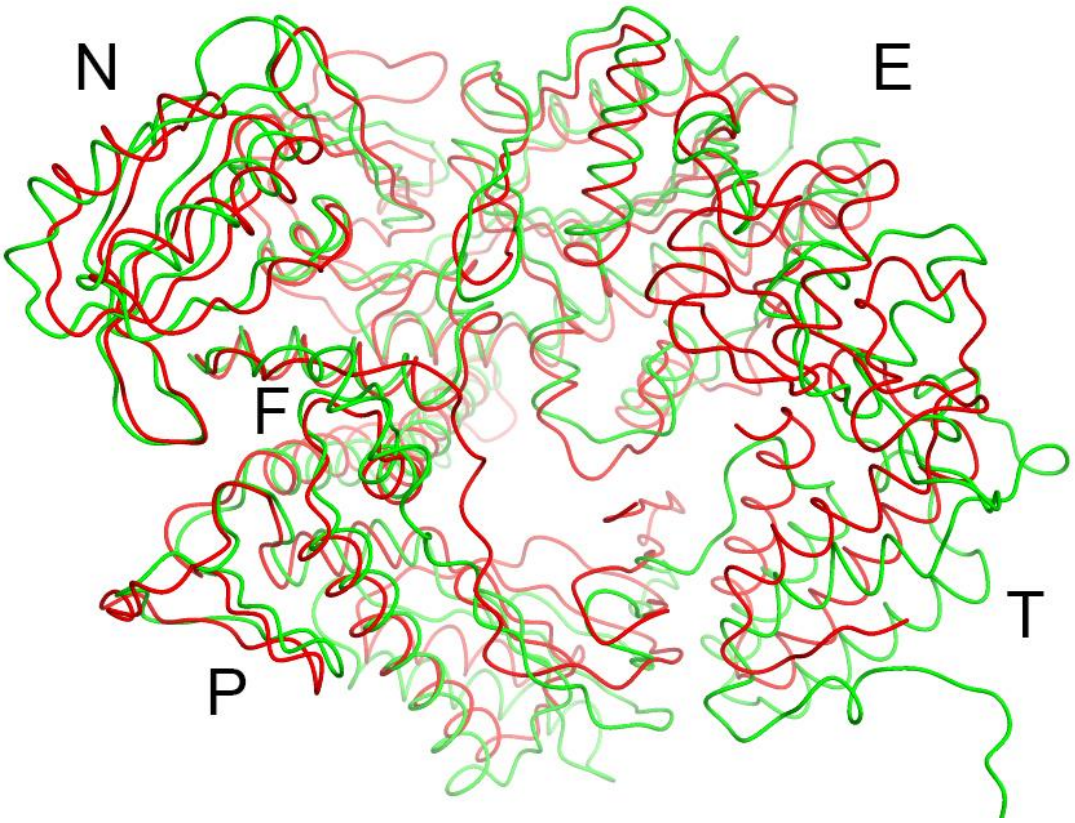


Figure 5.5 Secondary structure superposition of Pc-polymerase with the *pfu* enzyme. The Pc-polymerase structure is coloured red while the *pfu* structure is coloured green. The N-terminal, exonuclease, palm, fingers and thumb domains are labelled as N, E, P, F and T, respectively. The thumb domain has apparently moved a lot in response to DNA binding.

5.4.3 Modelling the structure with DNA

The overall shape of DNA polymerases is like a disk with a hole in the centre. There are three deep grooves emanating from the hole formed by the five domains. One of these deep grooves, known as cleft D, binds duplex DNA while another, referred to as cleft T, binds single stranded template DNA. The third cleft is the editing channel that leads to the active site of the 3'-5' exonuclease. When there is a mismatched nucleotide, it is directed towards the exonuclease active site through this channel. In response to DNA binding, the fingers and the thumb

domains tend to move relative to the palm domain which acts as a clamp to hold the DNA.

A protein-DNA complex of Pc-polymerase was constructed by fitting it to the DNA-bound structure of *pfu* polymerase (PDB ID: 4AIL). The DNA-*pfu* complex was obtained by co-crystallisation of its mutant-type protein with a primer-template duplex of DNA. Since there are domain movements on binding to DNA, the structures were first fitted by the N-terminal and the exonuclease domains, followed by fitting the other domains separately with their equivalent parts. The modelled complex structure is shown in Figure 5.6. The largest domain movement affects the thumb domain which, compared with the original model, rotates by 7.6° with its centroid moving by 10.0 Å. As a result of this domain movement, its helical hairpin region moves away from the DNA whilst the small *β*-sheet region moves closer to it. The palm and the fingers domains rotate by 15.5° and 14.1° and shift by 4.0 Å and 1.4 Å, respectively. The 3' end of the primer strand is positioned close to the catalytic residues Asp420 and Asp562 in the active site. Interestingly, the linker region composed of residues 391-404 seems to partially block the active site which suggests that it is likely to adopt a different conformation on DNA binding.

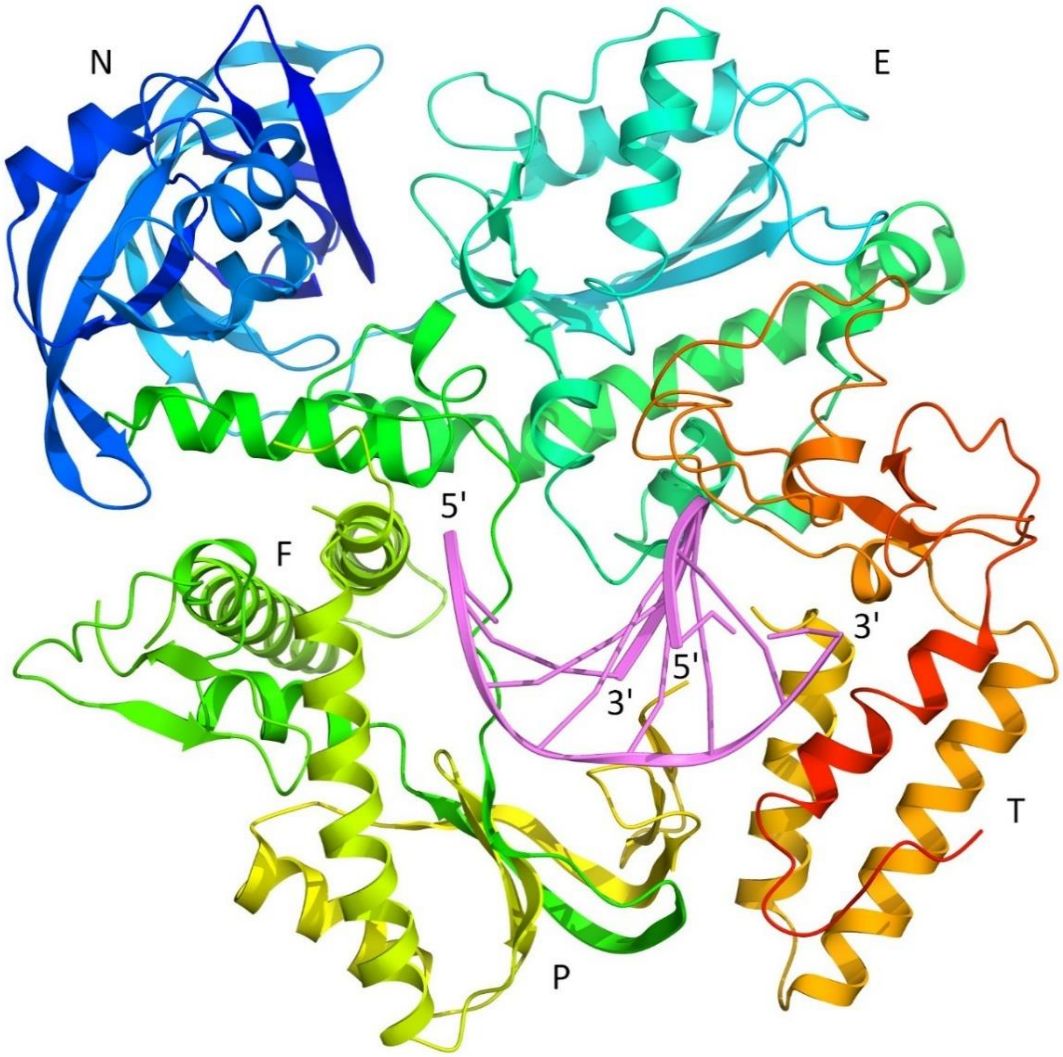


Figure 5.6 A modelled complex of Pc-polymerase with DNA. It was generated by fitting the enzyme domains separately with their corresponding parts in the *pfu*-DNA complex. Again, the N-terminal, exonuclease, palm, fingers and thumb domains are labelled as N, E, P, F and T, respectively. The primer-template duplex of DNA is coloured in purple. [Figure from (Guo *et al*., 2017c), reproduced with permission of the International Union of Crystallography (http://journals.iucr.org)].

5.4.4 Electrostatic surface

The presence of the clefts T and D can be demonstrated by the solvent accessible surface of Pc-polymerase, as shown in Figure 5.7. There are many acidic groups which dominate the electrostatic surface potential at the active sites

of the polymerase and the exonuclease in the centre region. These acidic residues tend to bind the catalytically essential magnesium ions. While the outer regions are characterised by a more positive potential which is likely to interact with the sugar-phosphate backbone of DNA.

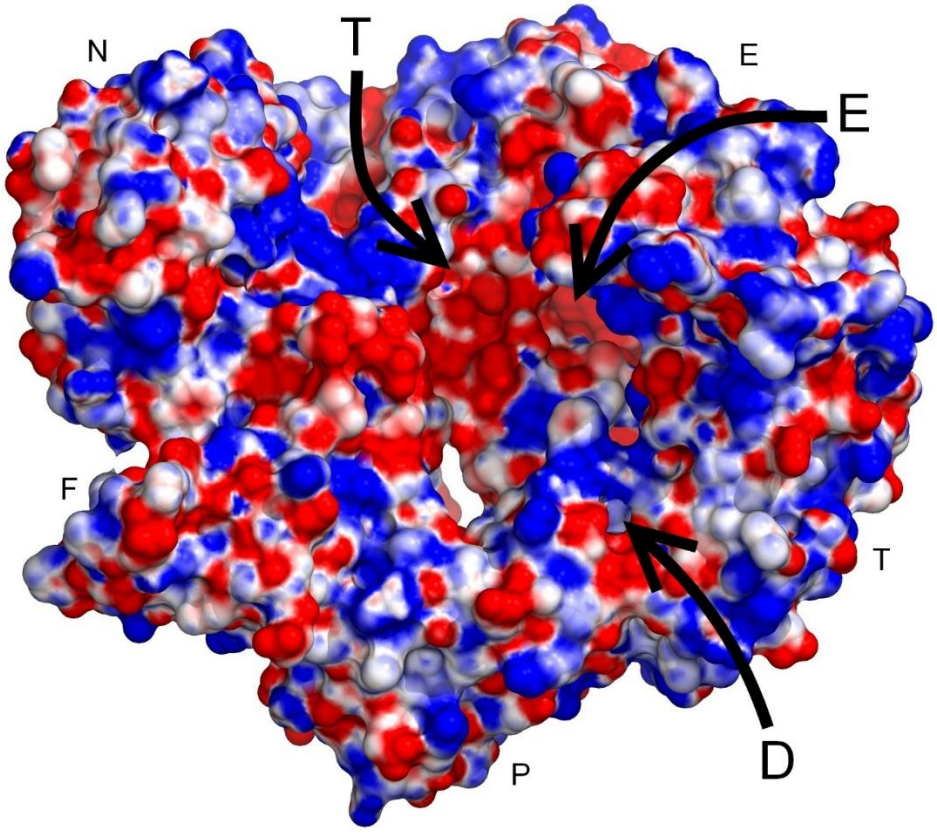


Figure 5.7 The solvent accessible surface of Pc-polymerase. The surface is coloured by electrostatic potential and the duplex cleft, the template cleft and the editing channel are labelled as D, T and E, respectively, with their general directions indicated by the arrows. The central hole is considered as the entry channel for incoming dNTPs that are added to the template. [Figure from (Guo *et al*., 2017c), reproduced with permission of the International Union of Crystallography (http://journals.iucr.org)].

5.4.5 Thermostability

The great thermostability of thermophilic proteins can be attributed to several factors. These proteins tend to have greater hydrophobicity (Haney *et al.*, 1997),

more hydrogen bonds (Vogt and Argos, 1997; Vogt *et al.*, 1997) and salt bridges (Haney *et al.*, 1997; Kumar *et al.*, 2000b; Yip *et al.*, 1995; Yip *et al.*, 1998), increased helical content, low occurrence of thermolabile residues such as Cys and Ser (Russell *et al.*, 1997), high occurrence of Arg, Tyr and Pro (Bogin *et al.*, 1998; Haney *et al.*, 1997; Watanabe *et al.*, 1997), amino acid substitutions within and outside the secondary structures (Haney *et al.*, 1997; Russell *et al.*, 1997; Zuber, 1988), better packing, smaller and less numerous cavities, deletion or shortening of loops (Russell *et al.*, 1997), increased surface area buried upon oligomerization (Salminen *et al.*, 1996) and increased polar surface area (Haney *et al.*, 1997; Vogt and Argos, 1997; Vogt *et al.*, 1997). However, it should be note that no single factor proposed to contribute toward protein thermostability is 100% consistent in all the thermophilic proteins. Kumar *et al*. (2000a) observed that the most consistent trend is shown by side chain-side chain hydrogen bonds and salt bridges. They may rigidify a thermophilic protein in the room-temperature range and the protein may still be flexible enough at high temperature in order to function (Jaenicke and Böhm, 1998).

By comparison of the thermophilic Pc-polymerase with the mesophilic family B DNA polymerase II from *E. coli* (PDB ID: 1Q8I), it was found that no trend could be observed for most of these factors, for example, the hydrogen bond content is similar in the two enzymes. About 75% of residues in Pc-polymerase form hydrogen bonds while this number is 77% for the *E. coli* enzyme. However, analysis showed that there is significant difference in the salt-bridge content. Initial analysis using a cut-off distance of 4 Å between donor and acceptor atoms revealed 60 ionic side-chain interactions in Pc-polymerase and 40 in the *E. coli* enzyme. These numbers increased to 212 and 124 (a ratio of almost 2:1) when

the cut-off distance was set to 6 Å considering that buried electrostatic interactions tend to have a longer range than 4 Å because of the lower dielectric constant of the protein interior. Further investigation revealed that this finding is consistent in pfu, KOD1 and other archaeal DNA polymerases. Thus when the conditions for electrostatic interactions are more relaxed, the difference between the thermophilic and mesophilic enzymes is more significant, which suggests that the extreme thermostability of Pc-polymerase may, at least partially, attribute to the longer-range electrostatic effects.

5.5 Summary

The family B DNA polymerase from *Pyrobaculum calidifontis* (Pc-polymerase) is magnesium-ion dependent and highly thermostable, which is indicated by an optimal temperature of 75 °C and a half-life of 4.5 h at 95 °C. It is of high-fidelity due to its error-correcting 3'-5' exonuclease activity. Unlike other broadly used high-fidelity DNA polymerases which have very high sequence identity to each other, such as those from *Pyrococcus furiosus*, *Thermococcus kodakarensis* and *Thermococcus gorgonarius*, Pc-polymerase has a very low sequence identity of approximately 37%. The crystal structure of Pc-polymerase has been determined at 2.8 Å. The domains are arranged in a circular disc-like shape with a narrow central channel, which is similar to other DNA polymerases. There are a few connected crevices in one face of the 'disc' which are involved in binding of single-strand and duplex DNA. The central channel is thought to allow incoming nucleoside triphosphates to access the active site. Pc-polymerase has several unique structural features that distinguish it from other archaeal DNA polymerases. The complex of the enzyme with the primer-template duplex of DNA was modelled which suggests a large movement of the thumb domain upon

DNA binding. The surface potential of the central region of the molecule, where catalytic magnesium ions bind, is dominated by acidic groups. Although there are many factors which may contribute to the great thermostability of thermophilic proteins, the high thermal stability of PC-polymerase may be mainly attributed to the large number of salt bridges.

The structural model and reflection files have been deposited in the PDB with accession code: 5MDN.

Chapter 6

Structure and function of the type III pullulan hydrolase from *Thermococcus kodakarensis*

6.1 Introduction

6.1.1 Pullulan and starch

Pullulan, also known as *α*-1,4- or *α*-1,6-glucan, is a polysaccharide synthesised by the fungus *Aureobasidium pullulans* from starch (Kim *et al.*, 1990), and is composed of repeating units of maltotriose linked by *α*-1,6-glycosidic bonds or repeating units of isopanose joined by *α*-1,4-glycosidic bonds (Leathers, 2003). The ratio of *α*-1,4- to *α*-1,6-glycosidic bonds in pullulan is 2:1. It is mainly used by the cells to resist predation and desiccation and is involved in diffusion of molecules both into and out of cells. Pullulan has been used as a model substrate for studying starch-debranching enzymes (Plant *et al.*, 1986) as well as in the food and pharmaceutical industries (Shingel, 2004; Singh *et al.*, 2008).

Starch is one of the most abundant polysaccharides and acts as a storage form of energy produced by green plants. Unlike pullulan which is soluble in water, starch is insoluble in water. Starches from different origins, plant organs and growth conditions have significantly different physical properties while most of them are a mixture of two high molecular weight polymers known as amylose and amylopectin (Swinkels, 1985). Amylose is a linear molecule formed by 100-10,000 D-glucose units connected by *α*-1,4-linkages while amylopectin, which constitutes approximately 73-80% of starch, consists of 24-30 *α*-1,4-linked D-glucose units joined by *α*-1,6-bonds, resulting in molecules with 9,600-15,900 glucose units (Takeda *et al.*, 2003). Amylose and amylopectin possess a latent aldehyde group at the end of the polymeric chain that is known as the reducing end. Depending on their origin, starches have various industrial applications such as manufacture of glucose, maltose syrups and production of other oligosaccharides (Rendleman, 1997; Van der Maarel *et al.*, 2002).

6.1.2 Enzymes involved in starch catabolism

Due to the complexity of its structure, depolymerisation of starch into oligosaccharides and smaller sugars requires a range of enzymes including endoamylases, such as *α*-amylases (EC 3.2.1.1) which cleave the chain internally, and exoamylases, such as glucoamylases (EC 3.2.1.3) which remove the terminal monosaccharides sequentially (Hii *et al.*, 2012). Transferases hydrolyse *α*-1,4-glycosidic bonds and transfer part of the donor to form a new *α*-1,4- [e.g. amylomaltase (EC 2.4.1.25)] or *α*-1,6- [e.g. branching enzyme (EC 2.4.1.18)] glycosidic bond with the acceptor. While the converse reaction, namely the hydrolysis of *α*-1,6-glycosidic bonds, is catalysed by debranching enzymes that are classified into the indirect and direct groups (Fogarty and Kelly, 1990). The indirect enzyme, amylo-1,6-glucosidase, requires the prior modification of the substrate by a transferase to leave a single *α*-1,6-linked glucose moiety at the branch point. The direct enzymes, known as isoamylases and pullulanases, can hydrolyse *α*-1,6-glycosidic bonds directly from unmodified substrate (Hii *et al.*, 2012). Hydrolysis of starch by amylase enzymes produces low molecular weight dextrins and greatly reduces glucose yield, which can be improved by the addition of pullulanases. In addition, by using pullulanase together with *β*-amylase in the starch saccharification process, maltose yield could be increased by about 20-25% (Poliakoff and Licence, 2007).

Pullulanases, or more precisely pullulan-hydrolysing enzymes, are grouped into the glycosyl hydrolase family 13 (GH13), 49 (GH49) or 57 (GH57) (Janeček *et al.*, 2014; MacGregor *et al.*, 2001) and, based on their substrate specificities and reaction products, are classified into five groups: pullulanase I and II, pullulan hydrolase I, II and III (Hii *et al.*, 2012). Pullulanase I (EC 3.2.1.41), which was

previously called R-enzyme, hydrolyses *α*-1,6-glycosidic bonds in starch, pullulan, glycogen and limit dextrins but does not degrade *α*-1,4-glycosidic bonds. Pullulanase II or amylopullulanase (EC 3.2.1.1/41) acts on both *α*-1,4- and *α*-1,6-linkages in polysaccharides such as starch and limit dextrins, while it generally hydrolyses at *α*-1,6-glycosidic bonds in pullulan, producing maltotriose (Nisha and Satyanarayana, 2013). Type I pullulan hydrolase (neopullulanases, EC 3.2.1.135) hydrolyses *α*-1,4-bonds of pullulan to generate panose and also hydrolyses both linkages of starch and related polysaccharides with a low efficiency (Kuriki *et al.*, 1988). Pullulan hydrolase II (isopullulanase EC 3.2.1.57) cleaves *α*-1,4-linkages of pullulan and panose, however, it does not act on starch or dextran (Aoki and Sakano, 1997). Type III pullulan hydrolase cleaves both *α*-1,4- and *α*-1,6-glycosidic bonds of pullulan, resulting in the formation of maltotriose, panose, maltose and glucose. Until now, only two enzymes in the last class have been reported, those from *Thermococcus aggregans* (TA-PUL) (Niehaus *et al.*, 2000) and *Thermococcus kodakarensis* (Ahmad *et al.*, 2014). The characteristics of these pullulan-hydrolysing enzymes are summarised in Figure 6.1.

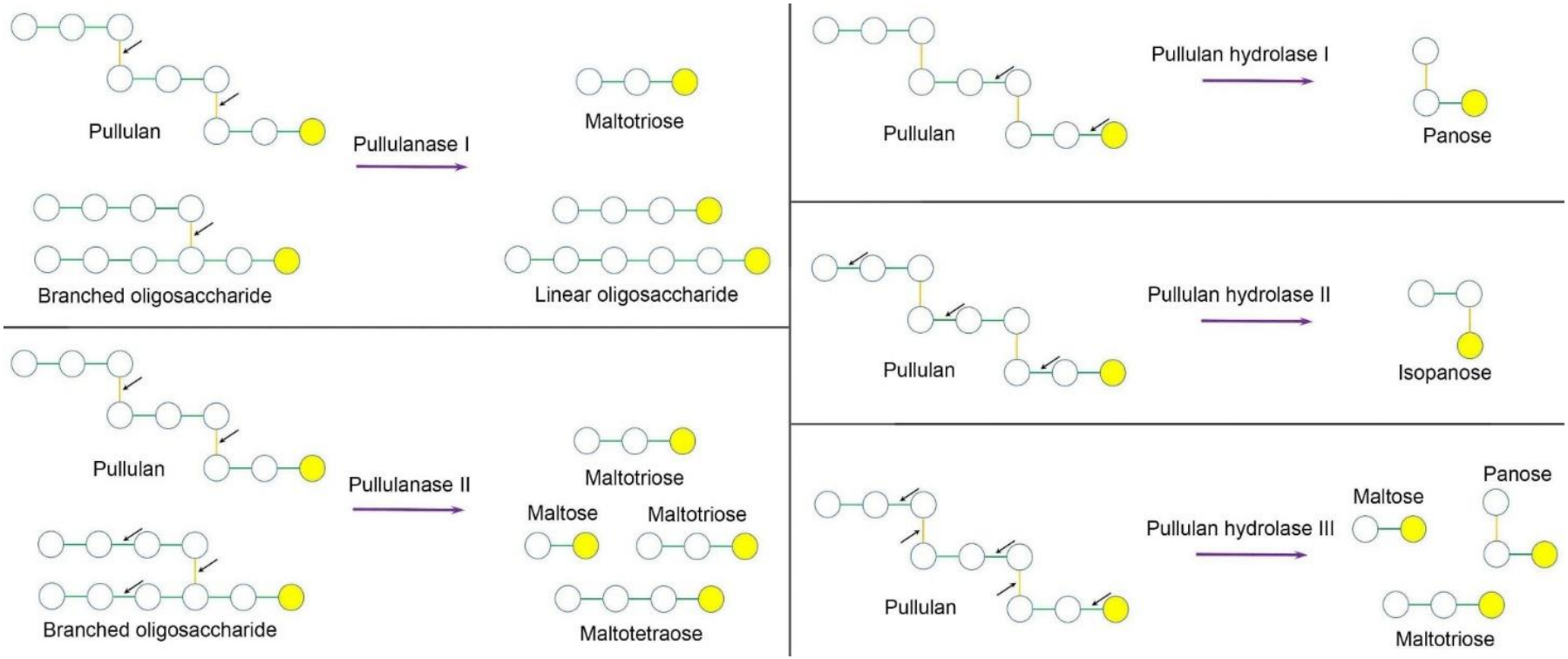


Figure 6.1 Reactions catalysed by different pullulan hydrolysing enzymes. The cleavage sites of pullulanases I and II as well as pullulan hydrolase I, II and III are indicated by the arrows and the products are shown. [Figure generated based on (Nisha and Satyanarayana, 2016)].

6.1.3 Catalytic mechanism

The catalytic mechanism of pullulan-hydrolysing enzymes is similar to that of the *α*-amylase family. The mechanism is characterised by *α*-retaining double replacement involving two catalytic residues in the active site: a Glu as the catalyst and an Asp as the nucleophile (Figure 6.2). The catalytic process can be divided into five steps: 1) Glu534 donates a proton to the glycosidic bond O and Asp601 nucleophilically attacks the C1 of glucose G1 (A); 2) an oxocarbonium ion-like transition state is formed followed by the formation of a covalent intermediate (B); 3) the covalent bond is then attacked by a water or a glucose molecule (B) which replaces the protonated glucose G2 ; 4) another transition state is formed again; 5) a H is transferred from the water or the glucose to Glu534 followed by the formation of a hydroxyl group or a new glycosidic bond (C) (Koshland, 1953; Van Der Maarel *et al.*, 2002). Asp503 is not directly involved in the catalytic process, instead, it binds to two OH groups of the substrate and plays an important role in the distortion of the substrate (Uitdehaag *et al.*, 1999).

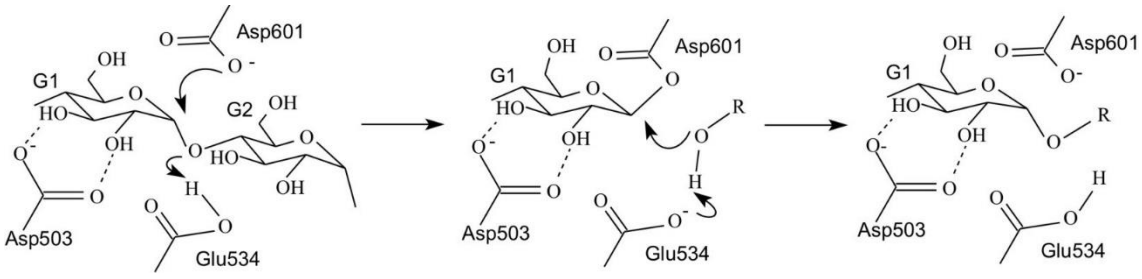


Figure 6.2 The catalytic mechanism of pullulan-hydrolysing enzymes. The amino acids are numbered according to the type III pullulan hydrolase from *T. kodakarensis* (TK-PUL). [Figure generated based on Van Der Maarel *et al*. (2002).].

6.1.4 Characteristics of the type III pullulan hydrolase from *T. kodakarensis*

The type III pullulan hydrolase from *T. kodakarensis* (TK-PUL) possesses both pullulanase and *α*-amylase activities. The enzyme has a molecular mass of 84.4 kDa with 748 amino acids. TK-PUL has the highest activity at 95-100 °C and the pH optima of 3.5 and 4.2 in acetate and citrate buffers, respectively, although it is active in a broad pH range from 3.0 to 8.5. It does not require any metal ions for the activity and has shown a broad range of substrate specificity including the ability to act on pullulan, *β*,*γ*-cyclodextrin, starch, amylose, amylopectin, dextrin and glycogen (Ahmad *et al.*, 2014). Interestingly, cyclodextrins are well-known competitive inhibitors of pullulanases (Duffner *et al.*, 2000) and none of the other enzymes have been reported to hydrolyse pullulan so efficiently. In addition, TK-PUL has a unique ability to hydrolyse maltotriose into maltose and glucose, which has not been reported in other homologous enzymes.

6.1.5 Applications of pullulan hydrolysing enzymes

Pullulan hydrolysing enzymes have high market value in starch saccharification industries to produce glucose, maltose, maltotriose and maltotetraose syrups. They increase the glucose or maltose production by 2% and 20-25%, respectively, and reduce the total reaction time and cost (Jensen and Norman, 1984). These enzymes are also employed as antistaling agents in the bread industry. Staling causes undesirable changes such as loss of bread flavour, decrease in moisture content and crispness of the crust during storage. It is due to the retrogradation of the amylopectin in starch and can be retarded by shortening the amylopectin chain length (Champenois *et al.*, 1999). These enzymes hydrolyse the branched maltodextrins produced by *α*-amylase (Carroll *et al.*, 1987), which is required for

making bread, and eliminate the gumminess of the bread. They are also involved in the preparation of resistant starch, panose and isopanose containing syrups (Machida *et al.*, 1986; Zhang and Jin, 2011).

6.2 Project aim

The aim of the project was to determine the crystal structure of TK-PUL, which would be the first structure of a type III pullulan hydrolyse, and to identify the differences between it and those from other classes of pullulan hydrolysing enzymes.

6.3 Methods

6.3.1 Crystallisation

TK-PUL was expressed and purified by Ahmad *et al*. (2014). Crystal screening was carried out by use of the sitting drop method using the same screening kits and robot as described in section 2.3.1. Protein samples at 5 mg/ml, 10 mg/ml and 16 mg/ml were screened at 21 °C with and without 1 mM Ca^{2+}, D-glucose, maltose, maltotriose, panose, n-dodecyl *α*-D-maltoside, *α*-cyclodextrin, *β*-cyclodextrin and *γ*-cyclodextrin (all in 10× molar excess except for Ca^{2+}). Many crystal clusters for ligand-free TK-PUL were obtained in the Morpheus conditions D1, E1, F1, G1 and H1 and many single crystals for the enzyme with n-dodecyl *α*-D-maltoside were obtained in the Morpheus B9 condition. However, all of them were of poor diffraction quality which did not allow structure determination. Further optimisation revealed that the best ligand-free crystal (Figure 6.3a, obtained at 10 mg/ml in 0.1 M carboxylic acids, 0.1 M buffer system 1 pH 6.2, 21% P500MME_P20K) diffracted to 2.8 Å whilst the best crystal with n-dodecyl *α*-D-maltoside added (Figure 6.3b, obtained at 12 mg/ml in 0.09 M halogen, 0.1 M

buffer system 3 pH 8.5, 40% P500MME_P20K) [For details of the conditions, see (Gorrec, 2009)] only diffracted to 4.2 Å, although these crystals looked a lot better. Selected crystals were mounted in loops before flash-cooling in liquid nitrogen.

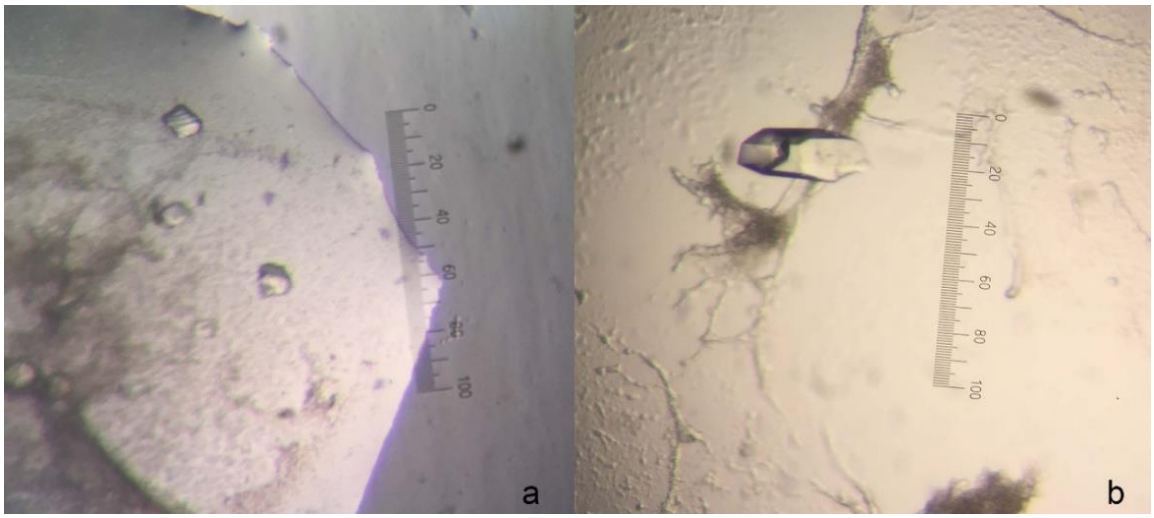


Figure 6.3 Crystals of the ligand-free (a) and ligand-bound (b) TK-PULs. One small unit on the ruler is 10 microns.

6.3.2 Data collection and data processing

X-ray diffraction data were collected at station I03, DLS and the raw data were processed using *DIALS* (Gildea *et al.*, 2014; Waterman *et al.*, 2013) to separate multiple lattices for the ligand-free crystal and to integrate diffraction spots for the ligand-free and ligand-bound crystals in the space groups *C*2 and $P3_221$, respectively. Both were later confirmed by *Pointless* (Evans, 2006; Evans, 2011) and molecular replacement. Scaling and data reduction were carried out using *Aimless* (Evans and Murshudov, 2013b) and data quality was checked by *Phenix.xtriage* (Zwart *et al.*, 2005). *Matthews_coef* (Kantardjieff and Rupp, 2003; Matthews, 1968) suggested a solvent content of 53.7% and 71.0% for the ligand-free and ligand-bound structures, respectively, with one molecule per ASU for both.

6.3.3 Structure determination

A partial structure solution for the ligand-free TK-PUL was identified by molecular replacement and refinement using the *MrBUMP* website (Keegan and Winn, 2008). The *Thermus thermophilus* HB8 pullulanase structure (Ttha1563, PDB ID: 2Z1K, 37% sequence identity with TK-PUL, Niwa *et al*., to be published) was used as the search model and residues from 286 to 764 were modelled into the electron density which gave a *Phaser* log-likelihood gain (LLG) value of 235.0, a translation function z-score (TFZ) of 12.7 and an R_{free} value of 44.7%. Since there was a large positive density at the N-terminal end of the partial structure, molecular replacement was repeated by use of *Molrep* (Vagin and Teplyakov, 2010) with the option SAPTF + local phased RF + phased TF using the partial solution as a fixed model and the residues 103-220 of *Staphylothermus marinus* maltogenic amylase (SMMA, PDB ID: 4AEE) as the search model for the partial structure (This was performed with the help from Dr Ronan Keegan, STFC, England). This enabled placement of an extra 101 residues (185-285) into the electron density which filled up all the map and brought the R_{free} value down to 39.5%. It was not possible to build the first 184 N-terminal residues due to the lack of the electron density. SDS-PAGE of the TK-PUL sample showed two strong bands of approximately 68 kDa and 15 kDa, which suggested that these residues might have been cleaved during sample preparation or storage. It was spotted that a lot of residues were shifted forward or backward at this stage, thus many rounds of manual rebuilding and correction were carried out including the relocation of residues guided by the electron density for the large aromatic side chains using *Coot* (Emsley and Cowtan, 2004; Emsley *et al.*, 2010). This was followed by further restrained refinement with *Phenix.refine* (Adams *et al.*, 2010;

Afonine *et al.*, 2012; Echols *et al.*, 2012). The ligand-bound structure was then determined by molecular replacement using the ligand-free structure as the search model followed by a few rounds of refinement. Manual rebuilding or correction was not possible due to the low resolution. All the statistics for data collection, data processing, structure determination and refinement of the ligand-free structure are shown in table 6.1. The *VADAR* (Willard *et al.*, 2003) and *ESBRI* (Costantini *et al.*, 2008) online services were used to analyse hydrogen bonds, salt-bridges and other factors related to thermostability of the enzymes.

Table 6.1 X-ray statistics for the ligand-free TK-PUL structure. Values in parentheses are for the outer resolution shell.

Beamline	I03 (DLS)
Wavelength (Å)	0.9762
Space group	*C*2
Unit-cell parameters	
a (Å)	192.6
b (Å)	63.9
c (Å)	56.1
β (°)	93.8
Resolution (Å)	96.11-2.80 (2.95-2.80)
R_{merge} (%)	16.7 (90.0)
R_{meas} (%)	20.1 (106.2)
$CC_{½}$ (%)	98.5 (53.6)
Completeness (%)	99.8 (98.9)
Average $I/\sigma(I)$	5.4 (1.2)

Multiplicity	3.3 (3.3)
No. of observed reflections	56,033 (7,968)
No. of unique reflections	17,039 (2,444)
Wilson plot *B*-factor (Å^2)	47.8
Solvent content (%)	52.8
R_{factor} (%)	24.2
R_{free} (%)	27.9
RMSD bond lengths (Å)	0.004
RMSD bond angles (°)	0.621
No. of reflections in working set	16,960
No. of reflections in test set	909
Mean protein *B*-factor (Å^2)	47.3

6.4 Results and discussion

6.4.1 Tertiary structure of TK-PUL

Figure 6.4 shows the tertiary structure of TK-PUL, which consists of an N-terminal, a central catalytic and a C-terminal domain. The N-terminal domain contains residues 185-280 forming an anti-parallel *β*-barrel structure. The central catalytic domain shows a TIM-barrel structure which is composed of residues from 281 to 694. The C-terminal anti-parallel *β*-barrel domain is formed by residues 694 onwards. The three domains form a triangle which are spatially close to each other.

Following the first *β*-strand in the central domain is a region (300-350) inserted in the TIM-barrel which contains two *α*-helices, a *β*-hairpin and a few loops that are involved in binding a calcium ion. The third *β*-strand is followed by another

insertion of 50 residues forming a very compact, kernel-like subdomain consisting of helical and strand features. There are a few helical segments following the 5th and the 6th strands and the last *β*-strand (8th) is followed by a flap-like structure which partially covers the active site and interacts with the kernel-like subdomain.

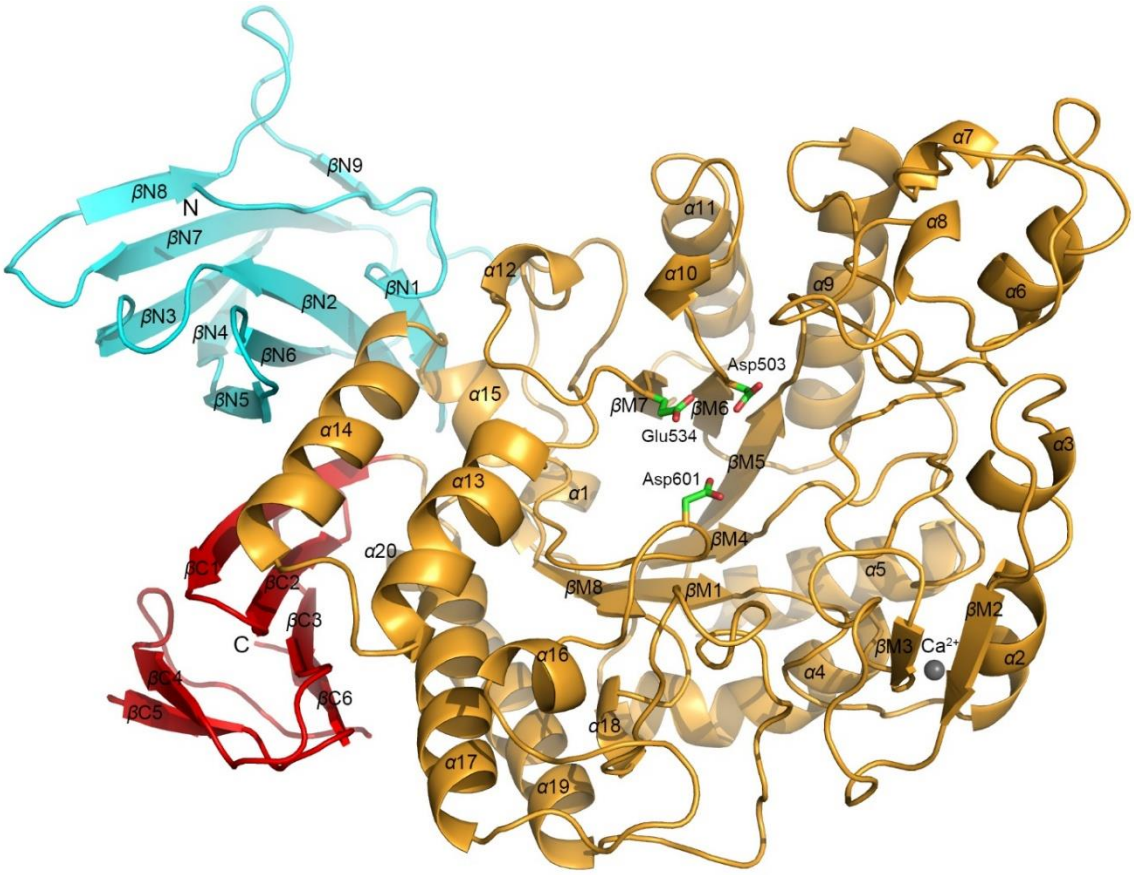


Figure 6.4 Crystal structure of TK-PUL. The N-terminal, central and C-terminal domains are coloured as cyan, orange and red, respectively. The three residues forming the catalytic triad are shown in ball-and-stick and the calcium ion is shown as grey sphere.

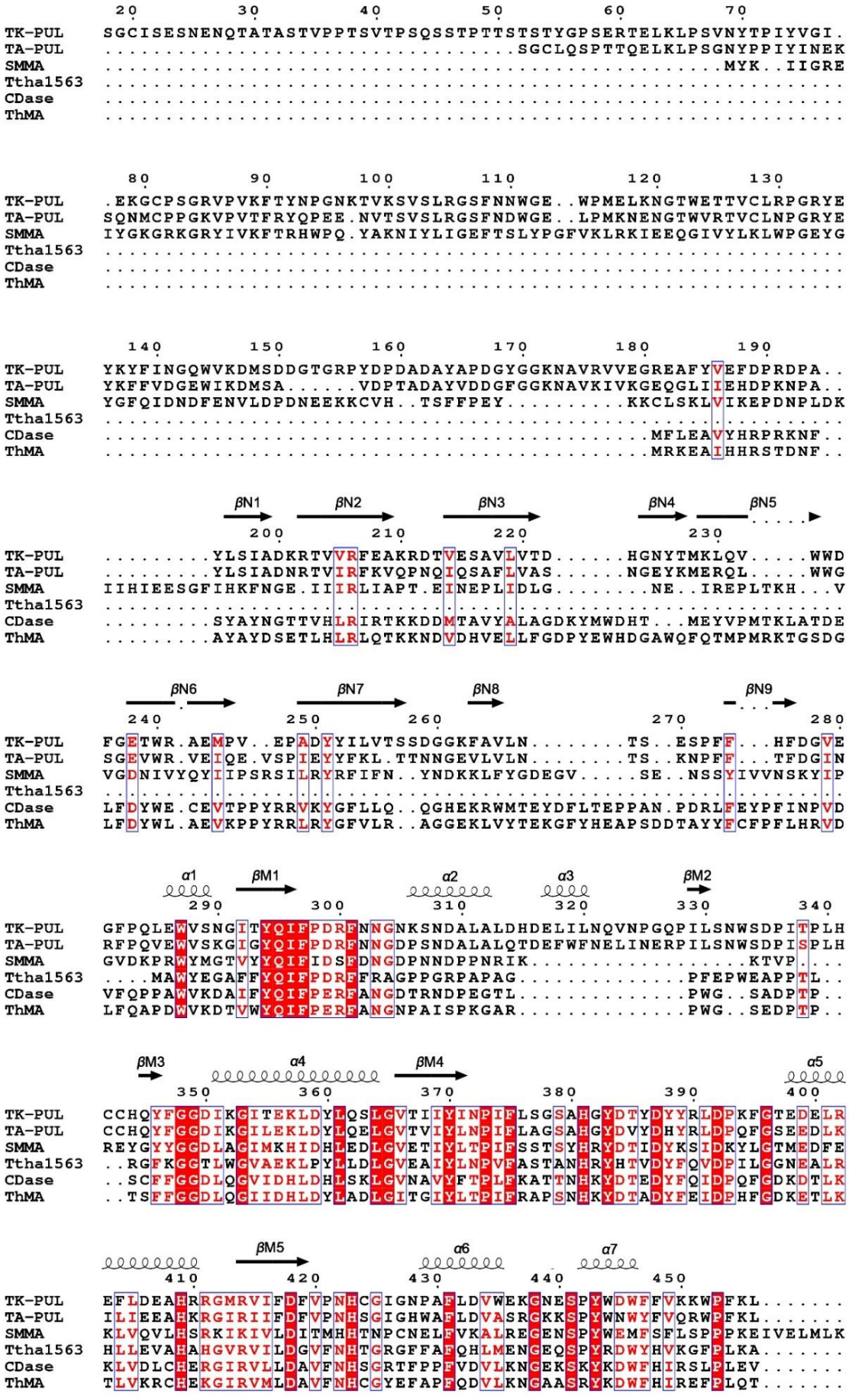
6.4.2 Structural difference with homologues

The sequence alignment of TK-PUL with several homologous structures is shown in figure 6.5. There is a signal peptide containing 17 amino acids before the N-terminus of TK-PUL which was included in the numbering scheme as previously reported by Ahmad *et al*. (2014), thus it is also included here to make it clear and

consistent with the report. TK-PUL has a 65.9% sequence identity with TA-PUL over 704 residues, whilst it shares only 27%-37% sequence identity with other aligned homologues over approximately 480 residues which cover the central and the C-terminal domains of TK-PUL. The N-terminal residues 185-280 of TK-PUL do not share high sequence similarity with the corresponding residues in SMMA, however, their tertiary structures are similar. Structure prediction using the *HHpred* website (Soding *et al.*, 2005) suggested that residues 78-180 are folded in a similar way to the corresponding region in the AMP-activated protein kinase from *Rattus norvegicus* (PDB ID: 4YEF) although this region could not be modelled due to the lack of electron density.

Figure 6.6 illustrates the structural superimposition of TK-PUL with the maltogenic amylases from *Thermus sp*. (ThMA, PDB ID: 1SMA) (Kim *et al.*, 1999) and *Bacillus sp*. I-6 (PDB ID: 1EA9) (Lee *et al.*, 2002). The TIM-barrel domain of TK-PUL shares a high structural similarity with that of the homologous proteins, while the N- and C-terminal domains are slightly different. TK-PUL is also homologous with the type I pullulan hydrolases from *T. thermophilus* (PDB ID: 2Z1K) and *B. stearothermophilus* (PDB ID: 1J0H) in both sequence and structure, ignoring that the former lacks the N-terminal domain. However, the loop regions at the active site end of the catalytic domain are quite different in these structures. For examples, the TK-PUL enzyme lacks the large insertion in the *α*-helical region following the 6^{th} strand, which forms a flap over the active site cavity in the *T. thermophilus* structure. The spatially-adjacent loop following the 7^{th} strand is substantially larger in the TK-PUL enzyme. The calcium-binding loop following the first *β*-strand of the TIM-barrel is considerably larger in TK-PUL and is

oriented toward the active site and is likely to be involved in peripheral interactions with large oligosaccharide substrates.


TK-PUL
TA-PUL
SMMA
Ttha1563
CDase
ThMA
βN1
βN2
βN3
βN4
βN5
βN6
βN7
βN8
βN9
α1
βM1
α2
α3
βM2
βM3
α4
βM4
α5
βM5
α6
α7

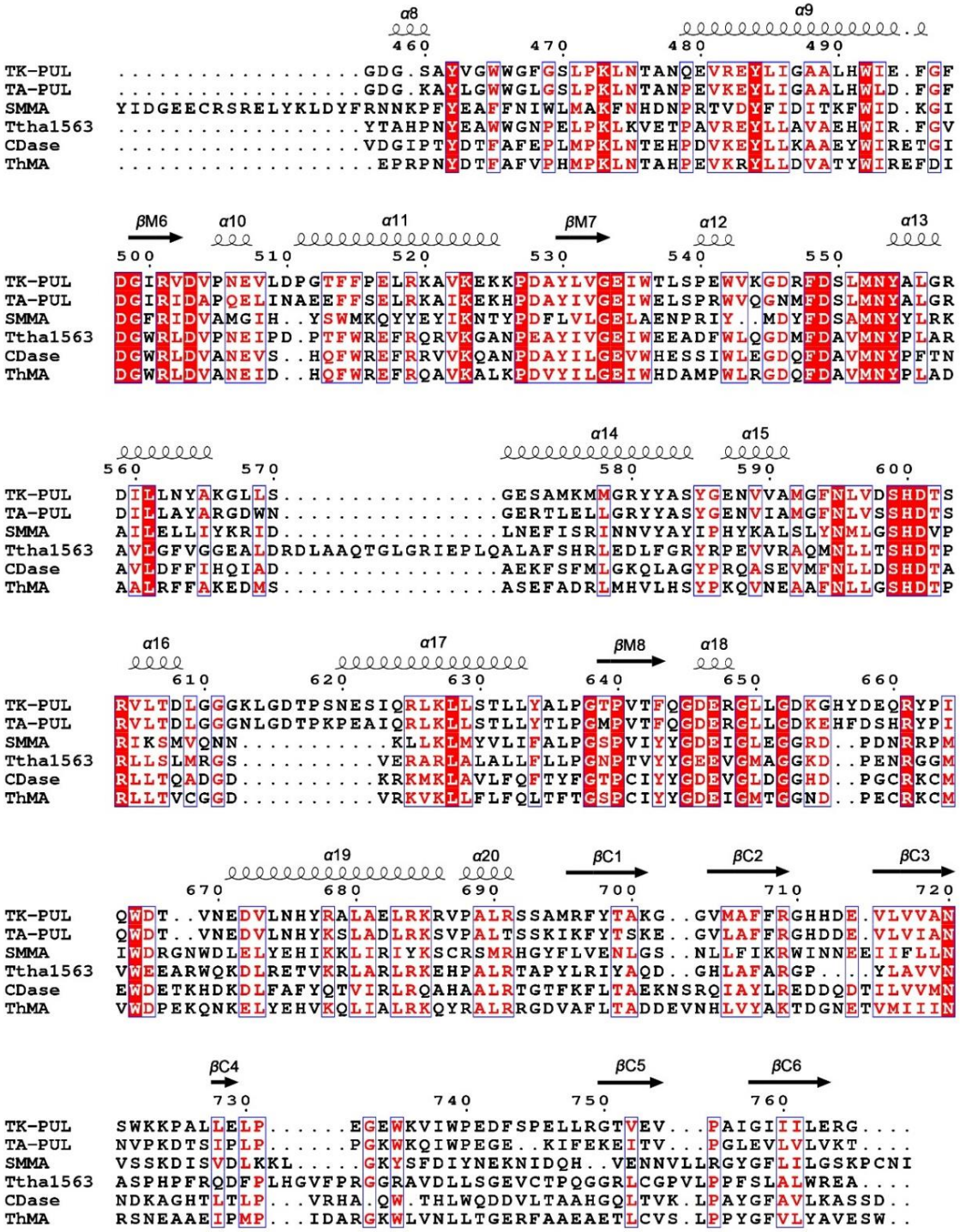


Figure 6.5 Sequence alignment and the secondary structure characteristics of TK-PUL with other homologues. Alpha-helices and *β*-strands are labelled according to the TK-PUL structure except for the N-terminal 184 residues which are missing in the structure. All the conserved residues are boxed and the fully conserved residues coloured in white on a red background, while the not fully conserved residues are coloured in red. Alignment was performed using *ESPript* 3.0 website (Gouet *et al.*, 2003; Robert and Gouet, 2014).

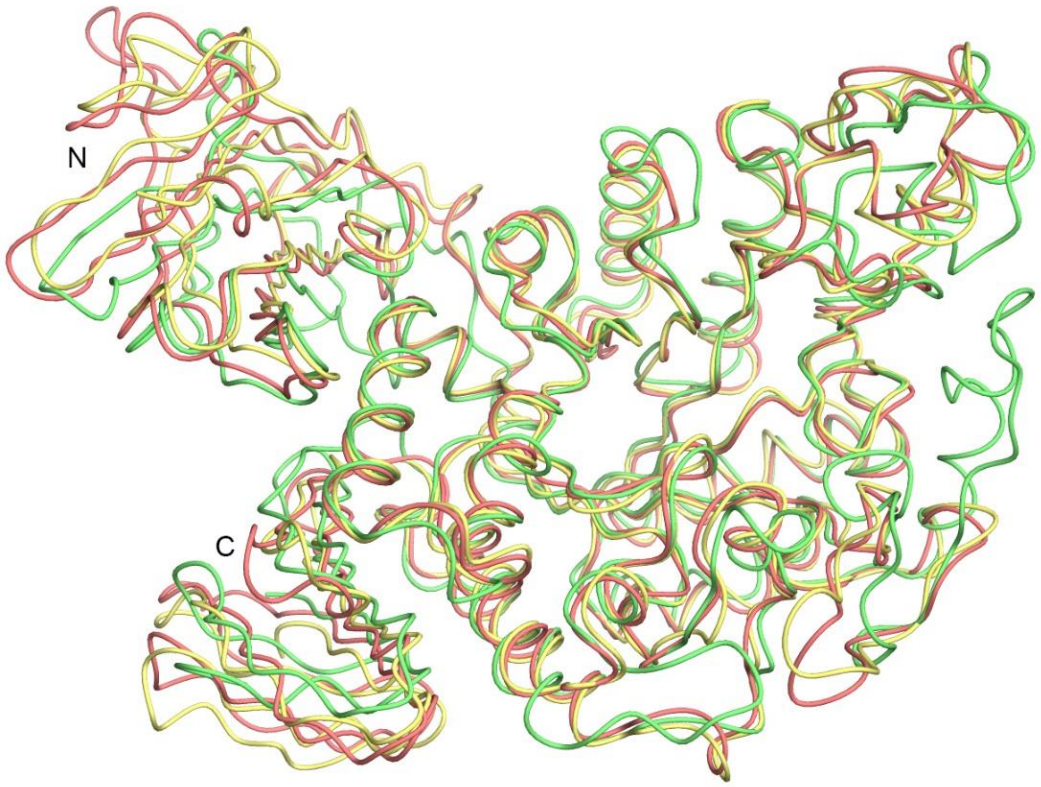


Figure 6.6 Superimposition of TK-PUL with two other homologues. TK-PUL, *Bacillus* cyclomaltodextrinase (CDase, PDB ID: 1EA9) and ThMA (PDB ID: 1SMA) are coloured in green, yellow and pink, respectively.

6.4.3 Active site

The catalytic triad composed of residues Asp503, Glu534 and Asp601 is located at the base of the cavity that contains many exposed aromatic residues as shown in Figure 6.7. The very exposed aromatic side chains of Trp465 and Phe468 suggest a role of binding the hydrophobic faces of substrates which have also been identified in related enzymes (Hondoh *et al.*, 2003; Ohtaki *et al.*, 2006).

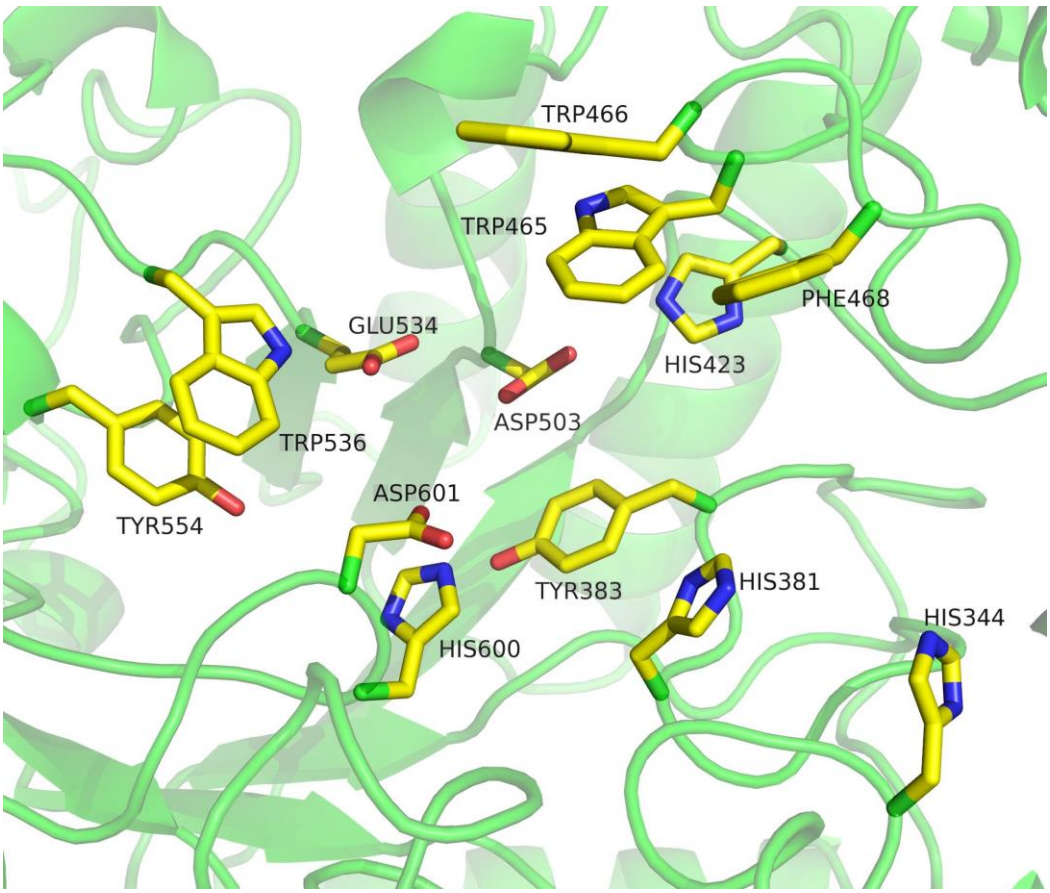


Figure 6.7 The active site of TK-PUL. The catalytic triad and other aromatic residues that are involved in carbohydrate binding are shown as ball-and-stick.

TK-PUL was also co-crystallised with n-dodecyl *α*-D-maltoside and the best crystal only diffracted to 4.2 Å, which did not allow for a precisely determined structure due to the poor electron density. However, molecular replacement using the ligand-free structure as a search model and the following refinement indicated a correct solution. Inspection of the active site identified a large positive feature of electron density where a maltose molecule could be fitted in (Fig 6.8). Superposition of this complex structure with the structure of *Bacillus subtilis* str. 168 type I pullulanase complexed with a maltose (PDB ID: 2E9B) indicated that these two sugar molecules superposed quite well. This suggests that it is very likely for the TK-PUL to possess a maltose molecule in the active site as a hydrolysis product of the substrate n-dodecyl *α*-D-maltoside, which was cleaved

at the glycosidic bond between the alkyl and the maltosidic parts. Despite the fact that the precise position of the maltose molecule cannot be located, it is in hydrogen-bonding distance of the catalytic residues and forms many hydrophobic interactions with the aromatic side chains of the residues around.

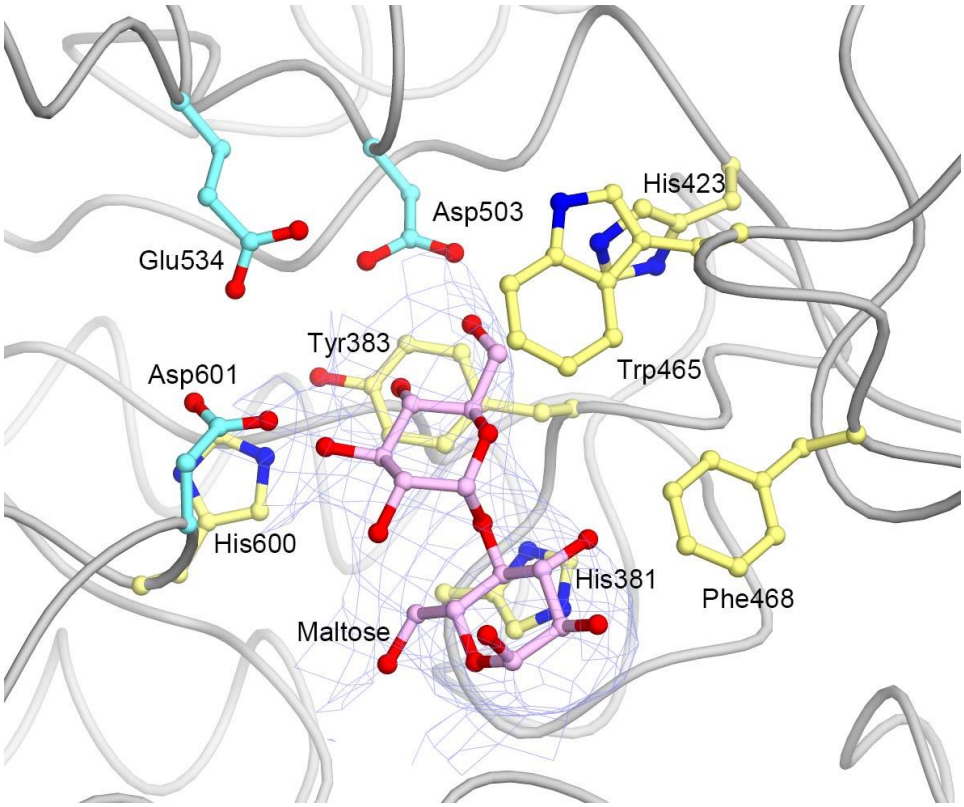


Figure 6.8 $2F_o$-F_c map of maltose in the complex structure. The maltose product is shown as ball-and-stick in pink, the catalytic residues and the aromatic residues in the active site are shown as ball-and-stick in cyan and yellow, respectively.

6.4.4 Calcium binding loop and vicinal disulphide

As mentioned in section 6.4.1, the region formed by residues 300-350 is involved in calcium binding by octahedral coordination as shown in Figure 6.8. The calcium binding site has been reported in many other homologous proteins such as the R-47 *α*-amylase II from *Thermoactinomyces vulgaris* (PDB ID: 1WZK, to be published) and the neopullulanases from *Bacillus stearothermophilus* (PDB ID:

1J0H) (Hondoh *et al.*, 2003). The residues that participate in calcium binding, through their main chain or side chain, include Asn303, Gly305, Asn308, Asp309, Gly348 and Asp350. The calcium ion is buried, completely dehydrated and is likely to play an important role in the stabilisation of the protein. There is also a rare vicinal disulphide bridge in this region which is formed by residues Cys342 and Cys343. This may have an influence on the substrate specificity of TK-PUL in a redox-dependent manner because the covalently linked cysteine side chains are oriented towards the active site.

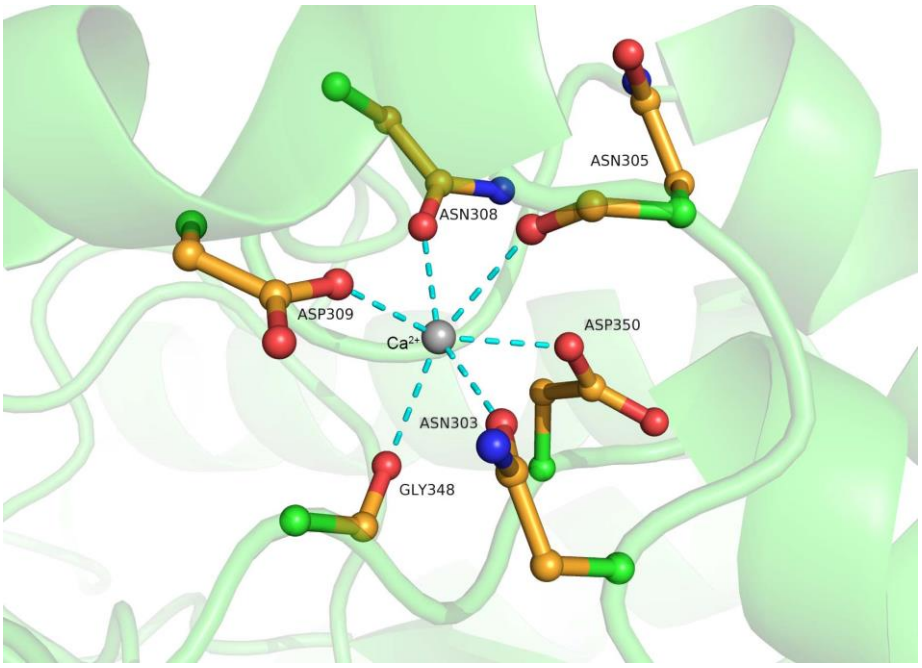


Figure 6.9 The calcium binding site of TK-PUL. Several main chain- and side chain-carbonyl groups are involved.

6.4.5 Thermostability

As mentioned in section 5.3.4.5, many factors may contribute to the thermostability of thermophilic proteins including increased hydrophobicity, more hydrogen bonds and salt bridges, increased helical contents, low occurrence of thermolabile residues such as Cys and Ser, high occurrence of Arg, Tyr and Pro, amino acid substitutions within and outside the secondary structures, better packing, smaller and less numerous cavities, deletion or shortening of loops, increased surface area buried upon oligomerization and increased polar surface area. The comparison of some of these factors for several thermophilic and mesophilic pullulan hydrolysing enzymes is shown in table 6.2 and those which may contribute to the thermostability are coloured as blue. The thermophilic enzymes show higher contents of salt bridges, helical segments, Pro, Arg and Tyr than the mesophilic ones, whilst the Ser content is much lower. Hydrogen bond content is similar in all the enzymes whist Cys content is slightly higher in the thermophilic enzymes which does not agree with the trend mentioned above.

Table 6.2 Thermostability-related factors for several thermophilic and mesophilic pullulan hydrolysing enzymes.

	Thermophilic				Mesophilic		
Enzyme*	TK-PUL	1SMA	1J0H	2Z1K	2YOC	2FH6	2WAN
Salt bridges (%)	25.7	29.9	30.6	30.1	20.4	17.6	16.5
H-bonds (%)	74	73	76	77	75	76	74
Helix content (%)	25	23	23	31	22	25	15
Pro content (%)	6.3	6.0	6.0	8.4	3.9	4.5	4.8
Arg content (%)	4.8	6.0	5.8	8.4	3.9	4.1	2.1
Tyr content (%)	4.8	5.1	5.4	9.4	3.6	3.6	4.3
Cys content (%)	0.8	1.4	1.4	0.4	0.6	0.6	0.1
Ser content (%)	6.3	2.9	3.6	1.9	9.2	9.3	6.5

* All the enzymes are represented by their PDB ID (except for TK-PUL) as follow: 1SMA, the maltogenic amylases from *Thermus sp.*; 1J0H, the type I pullulan hydrolases from *B. stearothermophilus*; 2Z1K, the type I pullulan hydrolases from *T. thermophilus*; 2YOC, the pullulanase from *Klebsiella oxytoca*; 2FH6, the pullulanase from *Klebsiella aerogenes*; 2WAN, the pullulanase from *Bacillus acidopullulyticus*.

Hydrogen bonds and salt bridges may rigidify a thermophilic protein in the room-temperature range and the protein may still be flexible at high temperature in order to function (Jaenicke and Böhm, 1998). Kumar *et al*. (2000a) suggested that the difference in the number of salt bridges between the thermophilic and mesophilic homologues appear to correlate with the melting temperature T_m of the thermophilic proteins . Alpha-helices enhance the rigidity and stability of proteins more than *β*-strands and loops and thermophilic helices favour Arg and avoid His and Cys as compared with mesophilic helices (Kumar *et al.*, 2000a; Warren and Petsko, 1995). Pro can only adopt a limited conformation due to the rigidity of the pyrrolidine ring. It was identified that thermophilic proteins tend to have Pro residues at the second sites of *β*-turns or the first turns of *α*-helices (Bogin *et al.*, 1998; Watanabe *et al.*, 1997), thus Pro exerts crucial effects on protein thermostability by controlling their folding. Arg and Tyr may be useful in both short and long range interactions due to their large side chains. In addition, the guanidium side chain of Arg can form salt bridges which stabilises proteins. In contrast, the short side chains of Cys and Ser mostly only form local interactions and they tend to undergo oxidation at high temperatures (Russell *et al.*, 1997).

6.5 Summary

The crystal structure of a thermoacidophilic type III pullulan hydrolase, TK-PUL, has been determined to a resolution of 2.8 Å which would be the first structure of a type III pullulan hydrolase. The unique properties of TK-PUL, e.g. great thermostability and extraordinary stability over a broad pH range, make it an ideal enzyme for the starch industry. The first 184 residues are missing in the structure which might have already been cleaved during protein preparation. The structure

of the last part of the N-terminal (185-280) and the C-terminal domains are different from the homologous structures and the loop regions at the active site end of the catalytic domain are quite different. The structure of the first part (residues 78-180) of the N-terminal domain is predicted to be similar to the corresponding region in the AMP-activated protein kinase from *Rattus norvegicus*. The complex structure suggested that n-dodecyl *α*-D-maltoside may be a substrate for TK-PUL, however, no further information can be obtained due to the low resolution. The region formed by residues 300-350 is involved in calcium binding, which has been reported in other homologous proteins. There is also a rare vicinal disulphide bridge formed by residues Cys342 and Cys343 which may have an influence on the substrate specificity. The thermostability of TK-PUL and a few homologues may be attribute to several factors including the increased content of salt bridges, helical segments, Pro, Arg and Tyr and the decreased content of Ser.

Chapter 7

Structure and function of the Thermostable L-asparaginase from *Thermococcus kodakarensis*

7.1 Introduction

L-asparaginase (EC 3.5.1.1) catalyses the hydrolysis of asparagine to aspartic acid and ammonia. Plants transport nitrogen, in the form of L-asparagine, from their roots to growing tissues thus they have a high demand on this enzyme (Atkins et al., 1975; Sieciechowicz et al., 1988). When amino acids become the primary carbon source in bacteria in anaerobic conditions, the expression level of asparaginase can be increased by 100-fold (Cedar and Schwartz, 1967; Cedar and Schwartz, 1968). This is an important up-regulation under anaerobic conditions as the metabolites of asparagine (as well as glutamine) can feed into the citric acid cycle. In contrast, the preferred carbon source glucose is a catabolite repressor of asparaginase expression. Enzymes in this family vary in their activity on glutamine to produce glutamic acid. Thus asparaginases and glutaminases are necessary for cell growth in ammonia-deficient media and their expression is activated by the presence of these amino acids in the medium.

7.1.1 Applications

7.1.1.1 L-asparaginase as an anticancer agent

L-asparaginase has been broadly used as a chemotherapeutic agent for treatment of acute lymphoblastic leukaemia (ALL) and other hematopoietic malignancies. ALL is the most common childhood acute leukaemia and contributes to approximately 80% of childhood leukaemias and 20% of adult leukaemias (Fullmer *et al.*, 2010). The history can be traced back to the 1950s, when Kidd spotted that the progression of murine lymphoma was limited by guinea pig serum (Kidd, 1953). This discovery attracted broad interest and it was found that only guinea pig serum had the anti-lymphoma activity compared with

other animals such as horse and rabbit, and it was only effective against certain cancer types. In the 1960s, Broome identified that it was the L-asparaginase in guinea pig serum which contributed mainly to the anti-lymphoma activity (Broome, 1961; Broome, 1963). Treatment with L-asparaginase has been shown to improve event-free survival for ALL from <10% to >80% in the last few years (Möricke et al., 2008; Pui et al., 2009; Silverman et al., 2001).

Cancer cells, such as lymphatic cells, have high demand on asparagine for their survival and proliferation (Kiriyama *et al.*, 1989; Stams *et al.*, 2003). However, due to the lack of L-asparagine synthetase required for L-asparagine synthesis, leukemic lymphoblasts and some other tumour cells can only obtain this amino acid from blood serum. L-asparaginase hydrolyses asparagine from blood serum, leading tumours to a state of cell death (apoptosis), while healthy cells are not affected because they possess enough L-asparagine synthetase. In addition, studies have shown that L-asparaginase inhibits the mTOR pathway and induces an autophagic process which contributes to its anti-leukaemic activity and greatly affects leukaemia cells (Russell *et al.*, 2014; Song *et al.*, 2015). Unlike conventional cancer therapy, L-asparaginase treatment is highly discriminatory.

L-asparaginases have been found in various sources including mammals, birds, plants, bacteria, fungi and archaea but only those from *E. coli* (EcAII) and *Erwinia chrysanthemi* (ErAII) have been approved for the treatment of ALL. The *E. coli* enzyme shows a higher activity whilst the *E. chrysanthemi* enzyme has been used to treat patients that are allergic to the former (Albertsen *et al.*, 2001). In the USA, the most commonly used form of L-asparaginase is a covalent conjugation with PEG, or PEGylation, which improves the bioavailability, biostability and reduces the immunological response. The elimination half-life of PEG-

asparaginase (e.g. 6 days) is five times longer than the native EcAII preparations and nine times longer than the ErAII ones. This is an important improvement since the enzyme shows a peak activity in the fifth day after an intramuscular injection (Shrivastava *et al.*, 2016). In Europe, it is currently a second-line treatment only for patients who are allergic to native asparaginases. Side effects of L-asparaginase therapy such as immune responses, allergies and anaphylactic shock (Soares *et al.*, 2002) may be attributed to several reasons including its L-glutaminase activity which reduces the plasma L-glutamine level (Avramis *et al.*, 2002; Villa *et al.*, 1986). Thus looking for alternative sources of L-asparaginase with less or no side effects is of great importance.

7.1.1.2 Other applications

L-asparaginase has been used to develop biosensors for the analysis of asparagine levels in leukaemia and in the food industry. In the designed experiments, hydrolysis of asparagine by this enzyme produces ammonium ions which induce a pH change that further changes the colour and absorption (Kumar *et al.*, 2013). This is a reliable, cheap and user-friendly approach compared with other conventionally used methods.

The enzyme is also widely applied in the food industry. Acrylamide, also known as 2-propenamide, is a colourless and odourless crystalline solid that has potent neurotoxicity. It is largely produced from heat-induced reactions, e.g. frying or baking starchy foods over 120 °C, and is formed between the *α*-amino group of asparagine and carbonyl group of reducing sugars such as glucose (Friedman, 2003). Therefore treatment of these foods with L-asparaginase prior to cooking significantly reduces acrylamide formation.

In addition, L-asparaginase is also involved in the biosynthesis of amino acids such as lysine, methionine and threonine.

7.1.2 Mechanism of L-asparaginase hydrolysis

The mechanism of L-asparaginase hydrolysis has not been fully understood. However, the process can be divided into two steps through an intermediate known as *β*-acyl-enzyme (Figure 7.1). Firstly, the activated catalytic residue of L-asparaginase nucleophilically attacks the amide carbon of the substrate L-asparagine and produces the *β*-acyl-enzyme intermediate. This is followed by the nucleophilic attack of the ester carbon by a water molecule which produces the product and frees up the enzyme (Verma *et al.*, 2007).

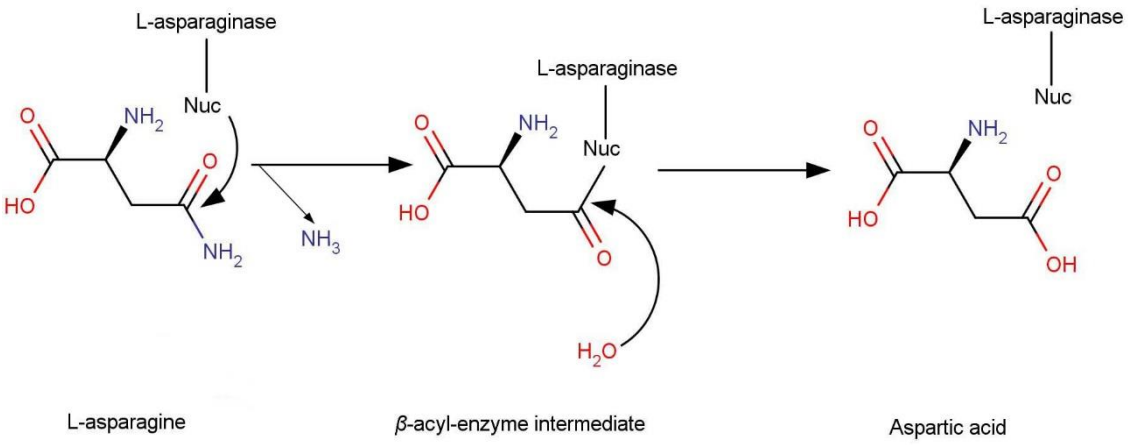


Figure 7.1 The mechanism of L-asparaginase hydrolysis. Nucleophilic attacks are indicated by the arrows. [Figure from (Guo *et al*., 2017b) originally generated based on Verma *et al*. (2007)].

Besides their activity on L-asparagine, these enzymes from different organisms showed different activities on L-glutamine and are put into two classes. One class has an activity of 2-10% to the L-asparagine hydrolysis activity, while enzymes in

the other class have comparable L-asparaginase and L-glutaminase activities (Boyd and Phillips, 1971; Chohan and Rashid, 2013; Davidson *et al.*, 1977).

7.1.3 The L-asparaginase from *Thermococcus kodakarensis*

Bacteria produce two forms of L-asparaginases, a cytoplasmic form (type I) and a periplasmic form (type II) which is more active and used in chemotherapy. Generally, the enzymes form dimers or tetramers (dimers of dimers) with the molecular mass of each subunit being around 35 kDa. Type I and type II L-asparaginases share low amino acid sequence identities e.g. those from *E. coli* have a sequence identity of only 24%. Although the two enzymes catalyse the same reaction, type I has a lower affinity for the substrate, for example, the *E. coli* type I L-asparaginase (EcA) has a K_M for asparagine of 3.5 mM whilst the type II enzyme has that value in the µM range (Schwartz *et al.*, 1966). The two isozymes can be distinguished by their sensitivity to thermal activation and by their solubility in ammonium sulphate solution (Yao *et al.*, 2005).

The thermostable L-asparaginase from *Thermococcus kodakarensis,* TkA, is a type I L-asparaginase that is composed of 328 amino acids with a molecular mass of 35.5 kDa (Chohan and Rashid, 2013). The enzyme is active as a homodimer in solution with the highest activity observed at pH 9.5 and 85 °C. It showed a K_M value of 5.5 mM against L-asparagine while no glutaminase activity was observed. In addition, TkA also exhibited D-asparagine activity which was about 50% to that of L-asparagine.

7.2 Project aim

The aim of the project was to determine the crystal structure of the thermostable L-asparaginase, TkA, as well as to identify the differences between it and the homologous enzymes from other organisms, which may improve or expand their applications.

7.3 Methods

7.3.1 Protein preparation and crystallisation

Recombinant TkA was expressed and purified according to the method described by Chohan & Rashid (2013) and the enzyme was stored in 20 mM Tris-HCl pH 8.0. Screening for crystallisation conditions was conducted by use of the sitting-drop method with the same Mosquito robot as mentioned in the previous chapters. Three different protein concentrations, 5mg/ml, 10 mg/ml and 16 mg/ml were applied to the Morpheus screening kit from Molecular Dimensions (Suffolk, UK) and the plates were stored at both 21 °C and 4 °C for crystallisation. Crystals started to appear in many conditions after 2 days and the best crystals (Figure 7.2) were obtained at a protein concentration of 10 mg/ml in the optimised C6 condition [0.09 M NPS, 0.1 M buffer system 2, pH 6.5-7.0, 28.5% (v/v) EDO_P8K] [For details of the condition, see (Gorrec, 2009)] at 21 °C. Selected crystals were transferred to a 10 µl drop containing 50% of paratone N MD2-08 and 50% of paraffin oil to remove the mother liquor surrounding the crystals before flash-cooling and data collection.

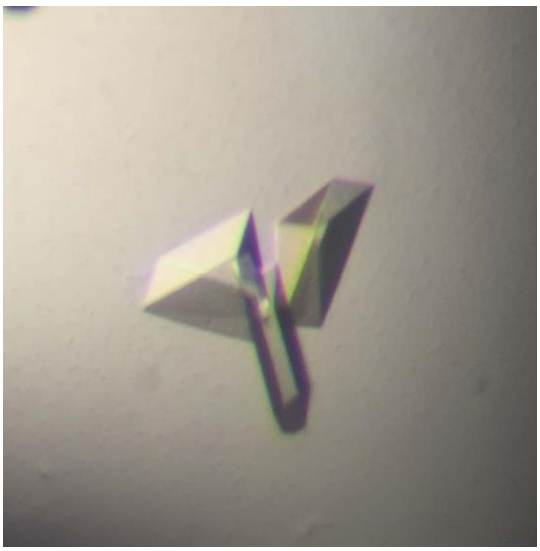


Figure 7.2 TkA crystals. These crystals were pentahedrons composed of two equilateral triangular- and three rectangular-faces. Each side of the equilateral triangle is approximately 250 microns and the width of the rectangle is approximately 60 microns.

7.3.2 Data collection, data processing and structure determination

X-ray data collection was carried out remotely at station I03, DLS at 100 K using a Pilatus3 6M detector. Automatic data processing using *xia2* (Winter, 2010) indicated that all the crystals were triclinic and belonged to the space group *P*1. Since this is an uncommon space group, other possibilities were checked by the integration of the raw diffraction images with *DIALS* (Waterman *et al.*, 2013) and scaling with *Aimless* (Evans and Murshudov, 2013b), which suggested that *P*1 should be the best option. The best crystal diffracted to 2.2 Å and the data were of good quality, as suggested by *Phenix.xtriage* (Zwart *et al.*, 2005). Analysis using *Matthews_coef* (Kantardjieff and Rupp, 2003; Matthews, 1968) suggested that the TkA crystal possesses 6 monomers per ASU with a solvent content of 47.32%.

The structure was determined by molecular replacement using the program *Phaser MR* (McCoy *et al.*, 2007) with the structure of the type I L-asparaginase from *Pyrococcus horikoshii* (PhA, PDB ID: 1WLS, 59% sequence identity to TkA) (Yao *et al.*, 2005) as the search model. Manual rebuilding and correction were accomplished using *Coot* (Emsley *et al.*, 2010) followed by TLS, local NCS and restrained refinement by use of *Refmac*5 (Murshudov *et al.*, 1997; Murshudov *et al.*, 2011). Model validation was performed using *MolProbity* (Chen *et al.*, 2010). All the statistics for data collection, data processing, structure determination and refinement are shown in table 7.1. The *VADAR* (Willard *et al.*, 2003) and *ESBRI* (Costantini *et al.*, 2008) online services were used to analyse hydrogen bonds, salt-bridges and other factors related to thermostability of the enzymes.

Table 7.1 X-ray statistics for the TkA structure. Values in parentheses are for the outer resolution shell [Table from (Guo *et al*., 2017b)].

Beamline	I03 (DLS)
Wavelength (Å)	0.9763
Space group	*P*1
Unit-cell parameters	
a (Å)	70.7
b (Å)	71.0
c (Å)	107.7
α (°)	72.1
β (°)	76.2
γ (°)	87.8
Resolution (Å)	68.65-2.18 (2.26-2.18)

R_{merge} (%)	6.3 (121.9)
R_{meas} (%)	7.5 (143.4)
$CC_{½}$ (%)	99.8 (58.8)
Completeness (%)	97.4 (97.2)
Average $I/\sigma(I)$	10.1 (1.3)
Multiplicity	3.5 (3.6)
No. of observed reflections	342,457 (35,236)
No. of unique reflections	98,478 (9,841)
Wilson plot *B*-factor (Å^2)	50.0
Solvent content (%)	47.3
R_{factor} (%)	19.8
R_{free} (%)	22.6
RMSD bond lengths (Å)	0.017
RMSD bond angles (°)	2.010
No. of reflections in working set	98,475
No. of reflections in test set	4,894
Mean protein *B*-factor (Å^2)	47.7

7.4 Results and discussion

7.4.1 Quality of the model

The structure of TkA L-asparaginase was determined by molecular replacement and was refined to a resolution of 2.2 Å. Data reprocessing with *DIALS* suggested only two possible Bravais lattice types: *mI* or *aP*. The *mI* lattice had an *RMSD* values of 1.44 and an R_{merge} value of 70.3% when the data were integrated into the corresponding *C*2 space group. However, the *aP* lattice had these two values

of 0.14 and 6.3%, respectively, when the data were integrated into the *P*1 space group. Thus the crystal belongs to the *P*1 space group with 6 monomers in the ASU, which gives a solvent content of 47.3%. The electron density of the first four chains (A, B, C and D) is of good quality whilst that of chain E is relatively poor. Many regions in chain F have very poor electron density. Indeed, the first four chains are characterised with *B*-factors of around 43.6 $Å^2$, while chains E and F have higher *B*-factors of 53.3 and 59.3 $Å^2$, respectively. Data analysis suggested that the data were of good quality and no anisotropy or translational NCS was spotted. Analysis with *Find_ncs* from the *PHENIX* suite (Terwilliger, 2013) suggested six NCS related chains (Figure 7.3) and deleting of the bad region in chain F increased the R-values by approximately 3%. Density-modification did not make any improvement. Analysis with *MolProbity* indicated that 95.5% of the residues are in the Ramachandran favoured region with 1.0% outliers.

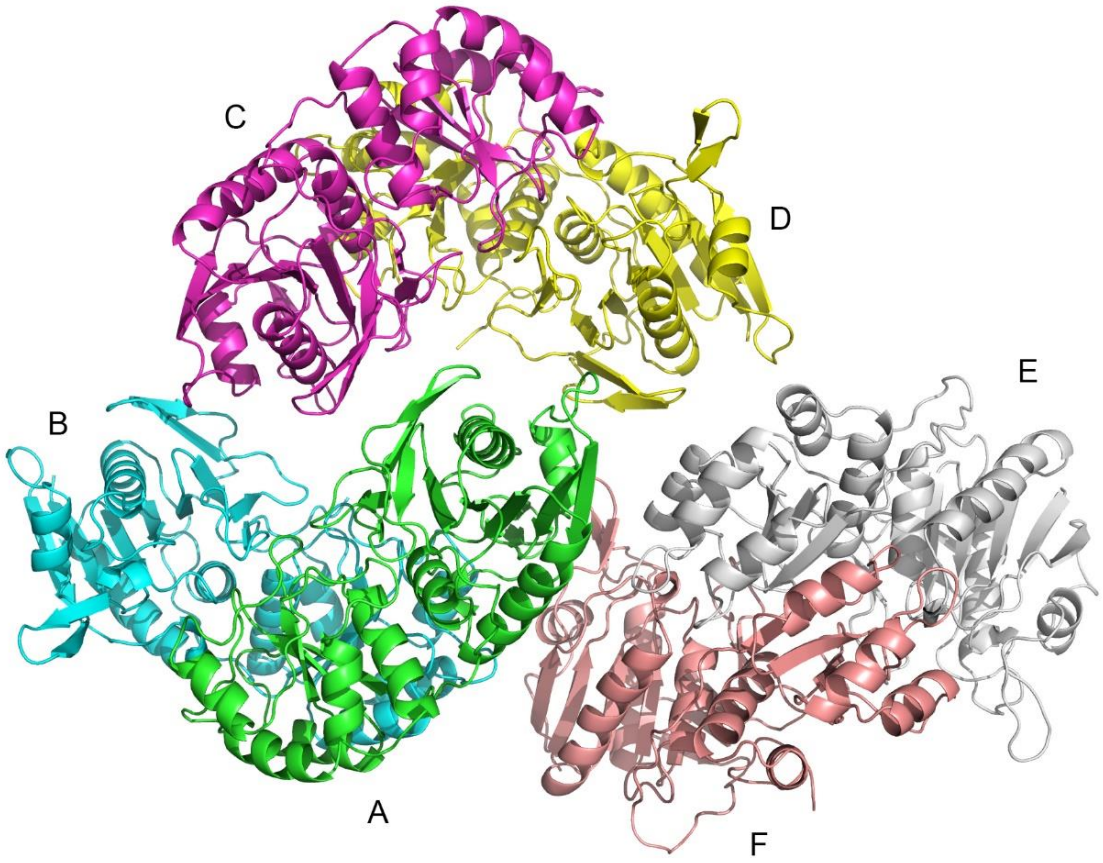


Figure 7.3 The six NCS related monomers of TkA in the ASU. Chain idenfiers are shown.

7.4.2 Overall structure

The six chains in the ASU of TkA share a very similar structure with *RMSD* values ranging from 0.08 to 0.15 Å for C^{α} atoms. When considering chain A only, the *RMSD* value between TkA and *E. coli* type I & type II L-asparaginases are 1.11 Å and 1.67 Å, respectively. Thus TkA has a similar overall fold to that of the type I and II L-asparaginases.

Each subunit of TkA consists of an N-terminal and a C-terminal *α*/*β* domain connected by a linker loop formed by residues 185-203 (Figure 7.4). The N-terminal domain contains an 8-stranded mixed *β*-sheet (*β*1, *β*4-8, *β*11 and *β*12)

flanked by 4 *α*-helices (*α*1-4). It has been shown that the *β*-hairpin composed of strands *β*2 and *β*3 is highly flexible and is often characterised by poor or no electron density. The *β*-hairpin is involved in substrate binding and catalysis and adopts ‘open’ and ‘closed’ conformations (Nguyen *et al.*, 2016). However, this area is characterised with good electron density in all the subunits of TkA. Chain B and D show a more ‘open’ conformation compared with EcA (PDB ID: 2P2D) (Yun *et al.*, 2007) whilst the other chains adopt a more ‘closed’ conformation, leaving that of the *E. coli* enzyme somewhere between them (Figure 7.5). Of greater significance than the fact that these subunits adopt either ‘open’ or ‘closed’ conformations in TkA is that the *α*1-helix has moved to a great extent toward the active site in all the subunits. The *α*4-helix has also moved slightly toward the active site. Thus it is very likely that these two helical segments also participate in substrate recognition in addition to the flexible *β*-hairpin. Leaving the main sheet between the two domains is another *β*-hairpin formed by *β*9 and *β*10, which may participate in subunit adhesion. The same as other type I L-asparaginases, TkA is active as a homodimer and the active site (indicated by the red star), which is located in the pocket around the two catalytically important residues Thr11 and Thr85, is composed of residues from two neighbouring monomers of the dimer.

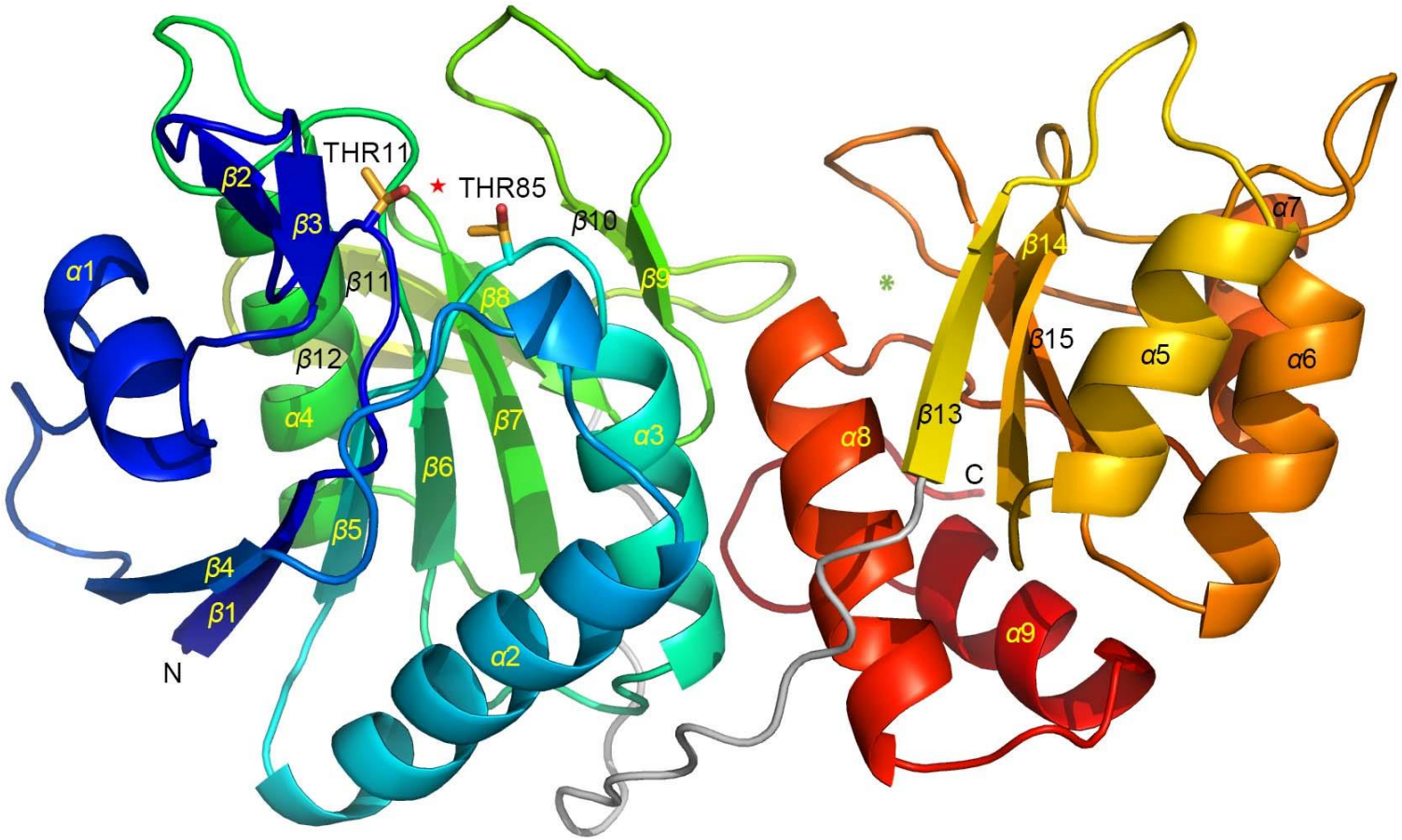


Figure 7.4 Overall structure of TkA. The monomeric structure is composed of an N-terminal and a C-terminal *α/β* domain connected by a loop (coloured as grey). The two key threonine residues involved in catalysis are shown in ball-and-stick. The active site and the putative allosteric site are indicated by the red star and the green asterisk, respectively [Figure from (Guo *et al*., 2017b)].

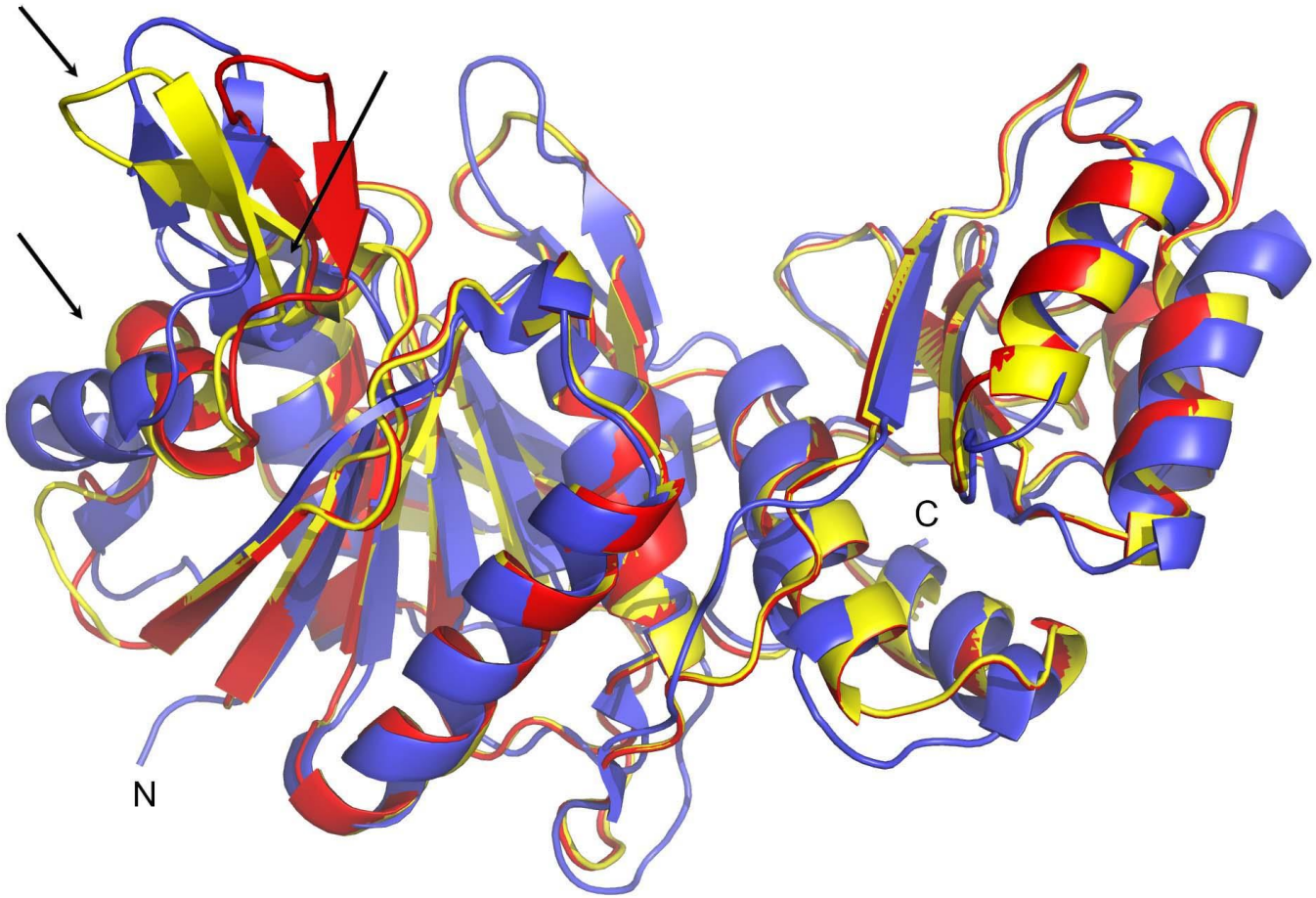


Figure 7.5 Structural superposition of TkA with the *E. coli* type I L-asparaginase. The structures are in approximately the same position as that shown in Figure 7.3. The chains A, B of TkA and the *E. coli* enzyme (PDB ID: 2P2D) are coloured as red, yellow and blue, respectively. The arrows indicate the segments which adopt different conformations in response to substrate binding [Figure from (Guo *et al*., 2017b)].

The relatively smaller C-terminal domain is formed mainly by a 3-stranded parallel *β*-sheet (*β*13-15) and 5 *α*-helices (*α*5-9). There is also a putative allosteric site which is located between the *β*15-strand and the *α*8-helix (indicated by the green asterisk in Figure 7.4) and is involved in asparagine binding (Yun *et al.*, 2007) which activates a cooperative conformational switch from an inactive to an active form.

The sequence alignment of TkA with several homologues is shown in Figure 7.6. TkA shares a 58.3% and 60.6% sequence identity over all residues with the L-asparaginase from *Pyrococcus furiosus* (PfA) and the L-asparaginase I homologue protein from *Pyrococcus horikoshii* (PhA), respectively. It only has sequence identities below 30% with EcA, EcAII and ErAII. However, there are many highly conserved regions in all of these proteins including the active site, which suggests that they are likely to share a similar tertiary structure.

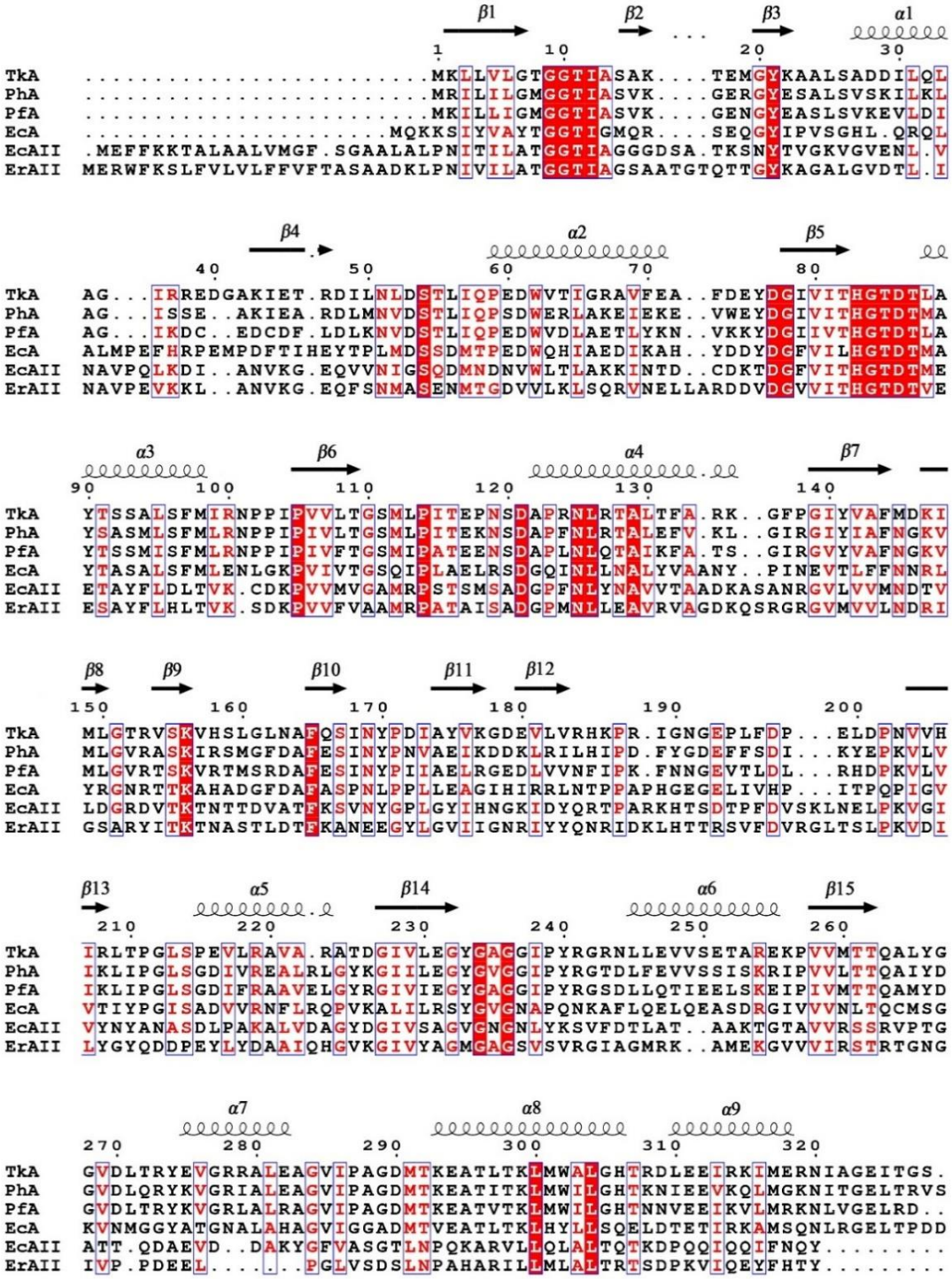


Figure 7.6 Sequence alignment and the secondary structure characteristics of TkA with other homologues. Alpha-helices and *β*-strands are labelled according to the TkA structure. All the conserved residues are boxed and the fully conserved residues coloured as white with a red background, while the not fully conserved residues are coloured as red. Alignment was performed using *ESPript* 3.0 website (Gouet *et al.*, 2003; Robert and Gouet, 2014) [Figure from (Guo *et al*., 2017b)].

7.4.3 Active site

As mentioned before, TkA is active as a homodimer (Figure 7.7a) and many residues from two neighbouring subunits are involved in substrate recognition as well as catalysis. Structural comparison identified that these residues in TkA include Thr11, Tyr21, Ser54, Thr55, Thr85, Asp86 and Lys156 from one subunit as well as Tyr233' and Glu275' from the neighbouring subunit, most of which are conserved in L-asparaginases (Figure 7.7b). One of the two key residues participating in catalysis, Thr11, resides in the *β*-hairpin whose flexibility is considered to be deeply involved in the activity of the enzyme. The other key residue, Thr85, is located in the loop between the *α*3-helix and *β*5-strand and mediates sequential 'ping-pong' nucleophilic attacks together with Thr11 during amidohydrolysis (Harms *et al.*, 1991). Since some unexpected electron density was identified in the active site of each monomer, many molecules have been fitted in an effort to identify what it is including the substrate asparagine and the product aspartic acid. However, this was finally interpreted and successfully refined as a phosphate ion which was one of the components in the crystallisation buffer and has been reported in other asparaginase structures (Tomar *et al.*, 2014; Wehner *et al.*, 1992). The phosphate ion occupies the binding site for the substrate and forms many interactions with the side chains of the neighbouring residues including Thr11 and Thr85. Tomar *et al*. also suggested that binding of a ligand (asparagine, citrate or phosphate) stabilises the flexible *β*-hairpin which acts as a gatekeeper and prevents further substrate entry. However, in the publication reported by Yun *et al*. (2007), this hairpin is still not visible in the Asp- or Asn-bound structures. In addition, the B and D chains in the TkA enzyme adopt

a more ‘open’ conformation than that of the ligand-free *E. coli* type I L-asparaginase (PDB ID: 2P2D).

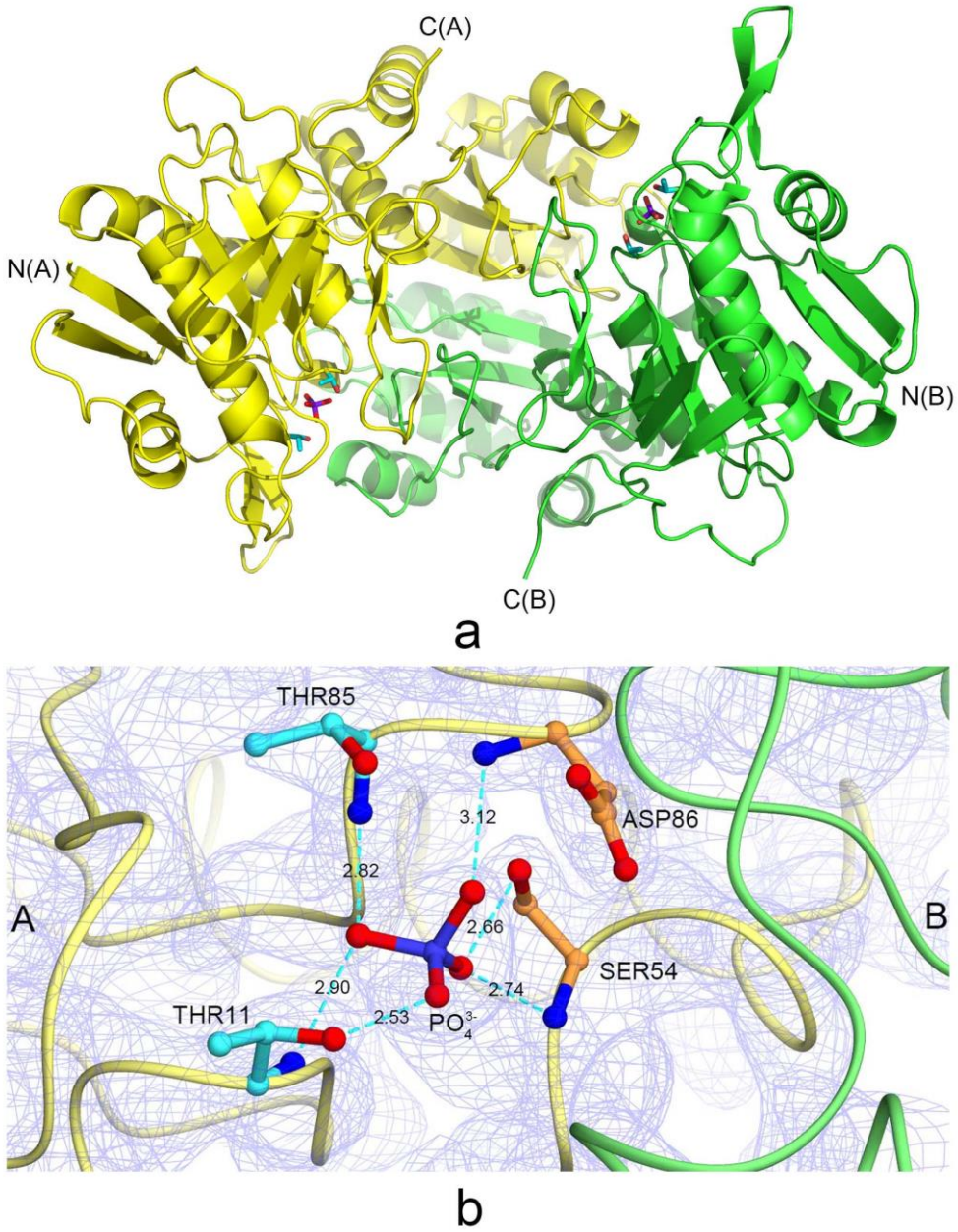


Figure 7.7 Dimer assembly and the active site of TkA. a) The dimer assembly formed between chain A (yellow) and B (green). b) The active site formed by residues from both chain A (yellow) and B (green). The two key threonine residues involved in catalysis are coloured as cyan and other residues participate in substrate recognition are coloured as orange. A phosphate (purple) ion has been identified to bind tightly in the active site in each subunit, with the hydrogen bonds showing in dashed lines [Figure from (Guo *et al*., 2017b)].

Yun *et al*. (2007) indicated that L-asparaginases which have high L-glutaminase activity possess smaller residues at the equivalent position of residue Gly237' in TkA. They predicted that those having a glycine at this position should have substantial glutaminase activity. However, whilst TkA does possess a glycine at this position, no glutaminase activity has been observed.

7.4.4 Thermostability

As mentioned in section 5.3.4.5, many factors may contribute to the thermostability of thermophilic proteins including increased hydrophobicity, more hydrogen bonds and salt bridges, increased helical contents, low occurrence of thermolabile residues such as Cys and Ser, high occurrence of Arg, Tyr and Pro, amino acid substitutions within and outside the secondary structures, better packing, smaller and less numerous cavities, deletion or shortening of loops, increased surface area buried upon oligomerization and increased polar surface area. The comparison of some of these factors for several thermophilic and mesophilic L-asparaginases is shown in table 7.2 and those which may contribute to the thermostability are coloured as blue. The dimeric structure was used in the calculation. The factors that may contribute to the thermostability of the investigated thermophilic L-asparaginases include the increased content of salt bridges and Arg and the decreased content of Cys and Ser. The salt bridge content seems to be the most consistent and significant factor that can be attributed to the great thermostability of all the investigated thermophilic enzymes in this thesis including the DNA polymerases, the pullulan hydrolysing enzymes and the L-asparaginases.

Table 7.2 Thermostability-related factors for several thermophilic and mesophilic L-asparaginases [Table from (Guo *et al*., 2017b)].

	Thermophilic			Mesophilic		
Enzyme*	TkA	1WLS	4Q0M	2P2D	3NTX	2OCD
Salt bridges (%)	20.0	17.5	19.0	13.9	9.2	9.9
H-bonds (%)	73	72	73	71	73	73
Helix content (%)	29	30	28	29	29	28
Pro content (%)	5.5	4.0	4.0	6.5	5.3	6.2
Arg content (%)	7.0	4.6	4.6	4.4	4.1	3.0
Tyr content (%)	3.0	3.7	3.4	4.1	3.3	3.9
Cys content (%)	0	0	0.6	0.3	0.3	0.6
Ser content (%)	4.3	6.7	4.9	4.7	7.4	5.9

* All the enzymes are represented by their PDB ID (except for TkA) as follow: 1WLS, the L-asparaginase I from *Pyrococcus horikoshii*; 4Q0M, the L-asparaginase I from *Pyrococcus furiosus*; 2P2D, the L-asparaginase I from *E. coli*; 3NTX, the L-asparaginase from *Yersinia pestis*; 2OCD, the L-asparaginase I from *Vibrio cholerae*.

7.5 Summary

The crystal structure of TkA has been determined at 2.2 Å in the *P*1 space group with 6 monomers in the ASU forming three dimeric pairs. Each subunit of TkA consists of an N-terminal and a C-terminal *α*/*β* domain connected by a linker loop. TkA is active as a homodimer and many residues from the neighbouring molecules in a dimer are involved in substrate recognition as well as catalysis. The N-terminal domain contains a highly flexible *β*-hairpin which adopts 'open' and 'closed' conformations. The *β*-hairpin is observed to adopt different

conformations in different subunits of the TkA structure. It was usually only observed in L-asparaginase structures that adopt a ‘closed’ conformation whilst it is characterised with good electron density in all the subunits in TkA structure. One phosphate ion has been built in the active site. The great thermostability of TkA may be attributed to the higher Arg content, lower numbers of Cys and Ser, but mainly to the increased content of salt bridges.

Chapter 8

Expression, purification and crystallisation of the juvenile hormone diol kinase from the silk worm, *Bombyx mori*

8.1 Introduction

Juvenile hormones (JHs) are a family of insect acyclic sesquiterpenoids which have multiple roles in regulation of many aspects of insect physiology including metamorphosis, sexual maturation, development and reproduction (Riddiford *et al.*, 2003; Wyatt and Davey, 1996). For example, these compounds maintain growth of the larva and are also involved in the production of eggs in female insects. JHs are produced in the corpora allata, a pair of endocrine glands behind the brain, released into the haemolymph and are transported to various tissues.

JH titres are precisely regulated by biosynthesis and degradation at different developmental stages. There are at least three enzymes participating in the JH degradation pathways known as JH esterase (JHE) (Hammock and Sparks, 1977), JH epoxide hydrolase (JHEH) (Share and Roe, 1988) and JH diol kinase (JHDK) (Maxwell *et al.*, 2002a). JHE is involved in the first pathway in which the methyl ester moiety of JH (Figure 8.1a) is hydrolysed to produce JH acid (JHa, Figure 8.1b). JHEH takes part in the second pathway where the epoxide moiety of JH is hydrolysed to form JH diol (JHd, Figure 8.1c). JHDK converts JHd to JH diol phosphate (JHdp, Figure 8.1d) which is the principal end product of JH degradation (Halarnkar *et al.*, 1993).

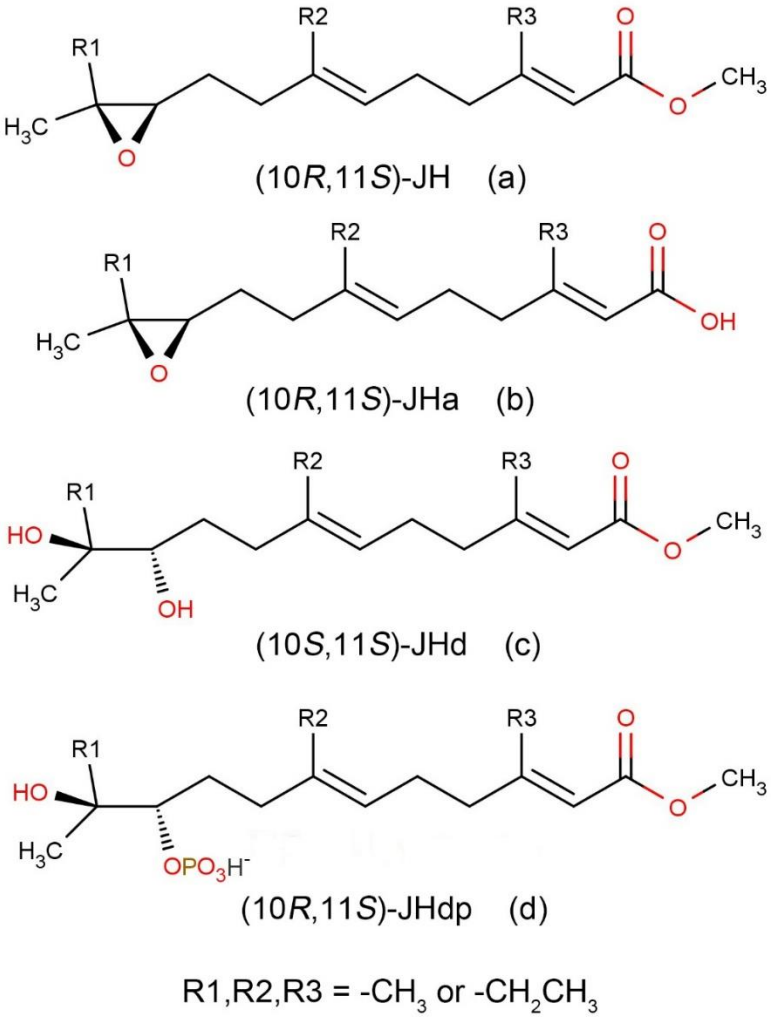


Figure 8.1 Structure of JHs and JH metabolites.

To date only three JHDKs have been reported including those from *Manduca sexta* (*Manse*-JHDK) (Maxwell *et al.*, 2002a; Maxwell *et al.*, 2002b), *Bombyx mori* (*Bommo*-JHDK) (Li *et al.*, 2005) and *Spodoptera litura* (SlJHDK) (Zeng *et al.*, 2015). *Bommo*-JHDK has a 54% sequence identity with both *Manse*-JHDK and SlJHDK. It is also considered to be homologous to *Drosophila melanogaster* sarcoplasmic calcium-binding protein-2 (dSCP2) and shares a 46.4% sequence identity with dSCP2 (Maxwell *et al.*, 2002b). Proteins in this family usually contain EF-hands that bind calcium ions (Nakayama and Kretsinger, 1994) and Li *et al*. (2005) suggested that JHDKs belong to a novel class of kinases with high

architectural similarity to calcium-binding proteins. *Bommo*-JHDK shares 36% sequence identity with the calexcitin from *Doryteuthis pealeii* which is a calcium binding protein composed of 4 EF-hands (PDB ID: 2CCM). The sequence alignment for *Bommo*-JHDK with several homologous proteins is shown in Figure 8.2.

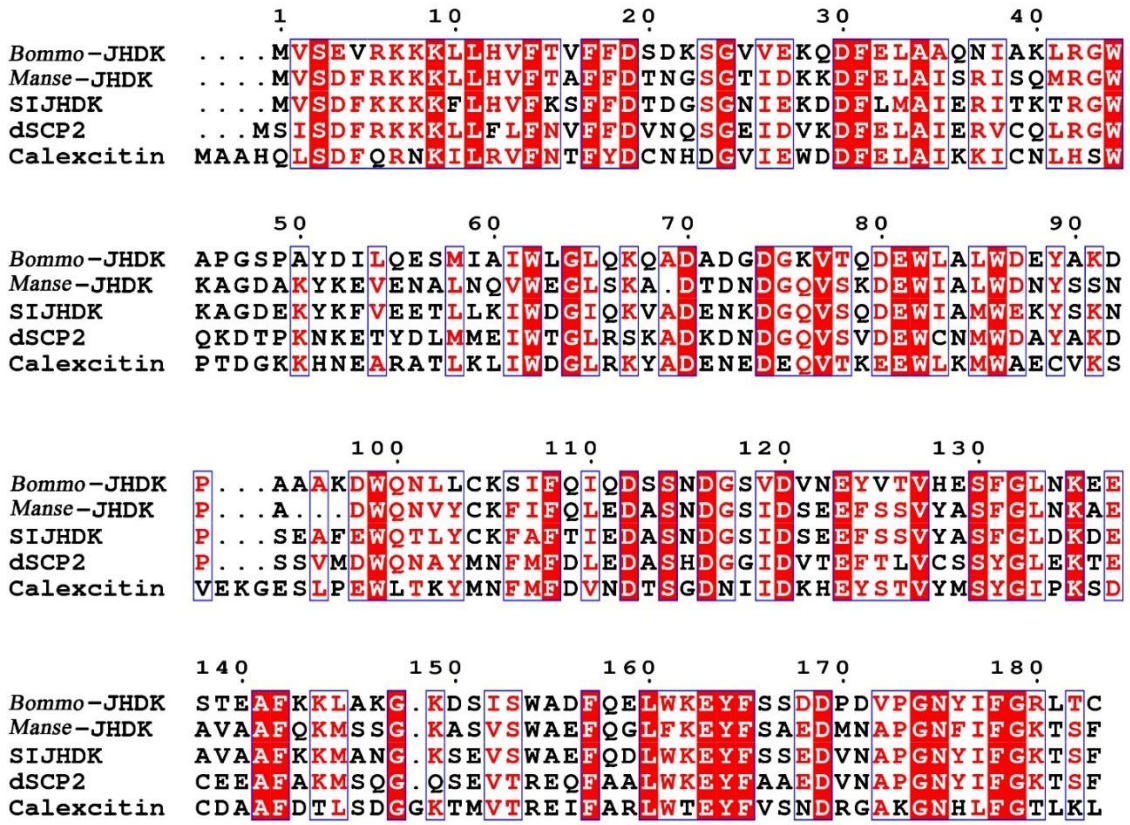


Figure 8.2 Sequence alignment and secondary structure characteristics of *Bommo*-JHDK with other homologues. All the conserved residues are boxed and the fully conserved residues coloured in white with a red background, while the not fully conserved residues are coloured in red. Alignment was performed using *ESPript* 3.0 website (Gouet *et al.*, 2003; Robert and Gouet, 2014).

JHDKs are active as homodimers with a molecular mass of 20-21 kDa for each subunit. They are characterised by pH optima of 8.0-8.5 and an optimal temperature of around 22 °C. These enzymes require magnesium ions for

catalysis and are very prone to metal poisoning. Calcium ions inhibit JHDK activity at micromolar levels. They have a preference for ATP to convert JH to JHdp, but can also catalyse the reaction using GTP as the phosphate donor. Due to their role in degradation of JH in insects, JHDKs may be a good drug target for development of pesticides.

8.2 Project aim

The aim of the project was to determine the crystal structure of *Bommo*-JHDK, and to identify the structural characteristics of this 'novel' class of kinases.

8.3 Methods

8.3.1 Plasmid re-construction

The *Bommo*-JHDK gene was a gift from Sheng Li, Chinese Academy of Sciences and Weihua Xu, University of Science and Technology of China. The gene was previously inserted into a pET-11a plasmid (Novagen, Millipore, Hertfordshire, UK) by a former group member, which did not allow a highly purified protein to be produced due to the lack of an affinity tag. Attempts to crystallise the crude protein failed to produce any crystal.

The pET-11a plasmid containing the JHDK gene was amplified in DH5*α* cells (Invitrogen, Thermo Fisher Scientific, Dartford, UK), purified by a GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific, Dartford, UK) and concentrated to 200 ng/µl followed by a double-digest with BamHI and NdeI restriction enzymes at 37 °C for 3 h to extract the *Bommo*-JHDK gene. The gene was separated from the digested and non-digested plasmid by electrophoresis on an agarose gel and was then extracted from the gel by electrophoresis again followed by a

purification using ethanol precipitation. It was then ligated on a pET-16b plasmid (pre-digested by BamHI and NdeI) (Novagen, Darmstadt, Germany), which has an N-terminal 10×His-tag, using a T4 DNA ligase following the protocol described in the manual (Thermo Fisher Scientific, Dartford, UK). The *JHDK*-pET-16b complex was amplified, purified and the presence of the *JHDK* gene was confirmed by gene sequencing.

8.3.2 Protein expression and purification

DNA transformation was undertaken according to Method i of the appendices, protein expression was performed following Method ii of the appendices using the standard method without heat shock. Protein purification was achieved firstly by using a HisTrap HP column (GE Healthcare, Buckinghamshire, UK) (binding buffer: 20 mM imidazole, 50 mM NaH_2PO_4, 300 mM NaCl, pH 8.0, elution buffer: 500 mM imidazole in binding buffer), followed by a Superdex 75 (GE Healthcare, Buckinghamshire, UK) gel-filtration column (buffer: 50 mM Tris, 100 mM NaCl, pH 7.5).

8.3.3 Crystallisation

Screening for crystallisation conditions was accomplished using the sitting-drop method at 21 °C with the same screening kits as mentioned in section 3.2.1. The same Mosquito crystal screening robot was used to dispense 400 nl of each protein, at 10 mg/ml and 20 mg/ml, plus 400 nl of the corresponding well solution into each drop. Only one crystal was obtained in the JCSG-*plus* H3 condition (0.1 M Bis-Tris pH 5.5, 25% PEG 3,350) after a few months. Further optimisation using 24 well hanging drop plates with 2 µl drop size showed that crystals with better diffraction quality (Figure 8.3) could be obtained reproducibly in exactly the

same condition, however, it took approximately 5 months for the protein to crystallise. Efforts were made in order to identify conditions that require a much shorter time for crystallisation, such as different crystallisation temperatures and additives as well as methods like proteolysis (Keegan *et al.*, 2014) and lysine methylation to reduce surface entropy (Sledz *et al.*, 2010) were tried, but no crystal-like object was obtained.

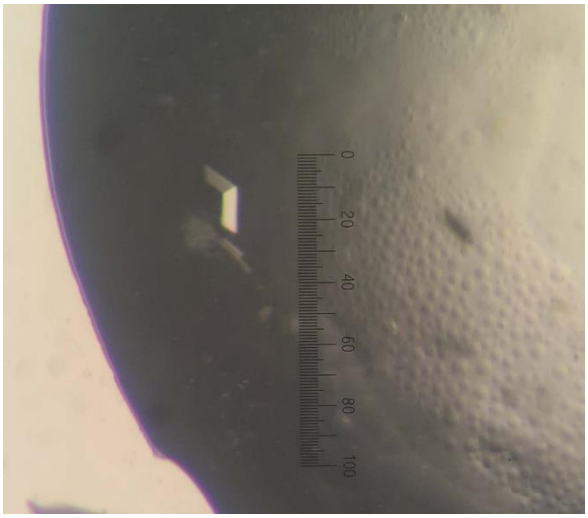


Figure 8.3 A crystal of *Bommo*-JHDK. One small unit on the ruler is 10 microns.

8.3.4 Making a heavy metal derivative

Since attempts to determine the crystal structure by molecular replacement failed to give any correct solution, a heavy metal derivative was made by soaking the selected crystals overnight in a drop of the corresponding well solution containing 10 mM Na_2PtCl_4. These crystals were then transferred into a drop of the well solution but no Na_2PtCl_4 to 'soak out' any unbound heavy metal ions. All the crystals including the native and the derivative ones were cryo-protected with 30% glycerol and flash-cooled before data collection.

8.3.5 Data collection, data processing and attempts to determine the structure

Data collections for both the native and the derivative crystals were performed at station I03, DLS. As guided by a fluorescence scan (Figure 8.4), three *MAD* datasets were collected for each derivative crystal at different wavelengths known as the peak (pk), the inflection (if) and the high remote (hrm) points. Automatic data processing using *xia2* (Winter, 2010) indicated that both crystals were monoclinic and belonged to the space group $P2_1$, which was confirmed by *Aimless* (Evans and Murshudov, 2013b). The native and the derivative crystals diffracted to 2.0 Å and 3.9 Å, respectively. All the statistics for data collection and data processing are shown in table 8.1. Analysis using *Matthews_coef* (Kantardjieff and Rupp, 2003) suggested 4, 5 and 6 molecules per ASU with a solvent content of 61.4%, 51.8% and 42.1%.

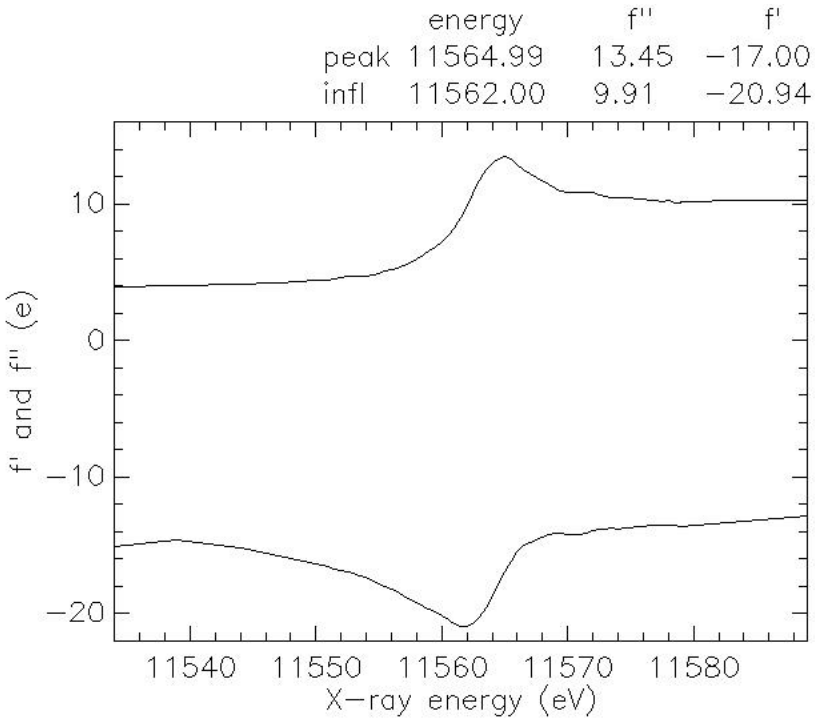


Figure 8.4 Fluorescence spectrum of a JHDK-Pt derivative crystal.

Table 8.1 Data collection and data processing statistics for the native and derivative *Bommo*-JHDK crystals. Values in parentheses are for the high resolution shell.

	Native	Derivative		
		pk	if	hrm
Beamline			I03	
Wavelength (Å)	0.9763	1.0721	1.0723	1.0633
Space group			$P2_1$	
Unit-cell parameters				
a (Å)	65.6	66.4	66.3	66.4
b (Å)	79.4	76.8	76.6	76.8
c (Å)	100.9	105.8	105.5	105.8
α (°)	90.0	90.0	90.0	90.0
β (°)	91.6	91.3	91.3	91.3
γ (°)	90.0	90.0	90.0	90.0
Resolution (Å)	50.59-1.99 (2.02-1.99)	55.68-4.08 (4.15-4.08)	66.24-3.91 (3.97-3.91)	56.77-3.72 (3.78-3.72)
R_{merge} (%)	10.7 (153.0)	10.6 (50.3)	10.0 (51.5)	9.0 (77.3)
R_{meas} (%)	11.6 (165.4)	12.9 (60.5)	12.2 (63.4)	11.0 (93.2)
$CC_{½}$ (%)	99.9 (62.0)	98.4 (70.8)	98.5 (68.4)	98.7 (65.7)
Completeness (%)	100.0 (100.0)	98.0 (97.4)	98.0 (94.8)	92.0 (16.1)
Anomalous completeness (%)		87.1 (4.3)	86.9 (4.2)	81.2 (8.0)
Average $I/\sigma(I)$	10.8 (1.2)	5.7 (2.3)	6.0 (2.0)	6.3 (1.2)
Multiplicity	6.8 (6.9)	3.0 (3.0)	3.0 (2.8)	3.0 (3.2)

Anomalous multiplicity		1.6 (1.6)	1.6 (1.5)	1.6 (1.7)
Anomalous correlation		-0.02 (-0.03)	-0.07 (-0.11)	-0.11 (-0.11)
No. of observed reflections	482,495	25,382	28,541	31,210
No. of unique reflections	71,164	8,478	9,556	10,490
Wilson plot *B*-factor ($Å^2$)	32.4	105.0	109.6	106.5

Structure determination using computer programs and online services based on molecular replacement did not give any correct solution including the programs *Phaser MR* (McCoy *et al.*, 2007), *Molrep* (Vagin and Teplyakov, 2010) and the online services *MrBUMP* (Keegan and Winn, 2008) and *BALBES* (Long *et al.*, 2008). Attempts on different search models and different parts of the search models also failed. Molecular replacement with all the structures in the whole PDB database was carried out using the website *SIMBAD* (Keegan *et al.*, 2016), which did not give any correct result, either. Experimental phasing was carried out using the *SHELX* (Sheldrick, 2010) and *CRANK2* (Skubák and Pannu, 2013b) web services, unfortunately, *MAD* phasing was not successful which was likely to be because the anomalous signal was only significant at resolution lower than 10 Å.

8.4 Results and discussion

8.4.1 Plasmid re-construction, protein expression and purification

The original *Bommo*-JHDK gene in a pET-11a plasmid did allow for successful protein expression, however, due to the lack of an affinity tag, attempts to purify the protein with ammonium sulphate precipitation and ion exchange gave a mixture of JHDK with a few contaminants including a nucleic acid. Screening for crystallisation conditions with the mixture did not produce any crystals. The *JHDK*-pET-11a construct was amplified and was incubated with BamHI and NdeI to extract the JHDK gene. Figure 8.5 shows the result of the double-digest, the bands for the insert at the bottom clearly indicate that the extraction was successful (pET-11a is 5677 bp and the JHDK gene is 558 bp). Since it was a double-digest and the NdeI restriction enzyme was already expired for some time, the efficiency of the digest was very low. The insert was then extracted from the gel by electrophoresis and purified by ethanol precipitation, which was followed by the ligation of the gene into a pET-16b plasmid. The successful ligation was confirmed by gene sequencing.

Bommo-JHDK was expressed in BL21 (DE3) cells and was purified by a Ni-column followed by a Superdex75 gel-filtration column, as indicated by the bands at around 25 kDa in Figure 8.6 which stand for the protein plus a few extra residues at the N-terminal including the 10×His-tag. Approximately 10 mg of highly purified protein was obtained per litre culture.

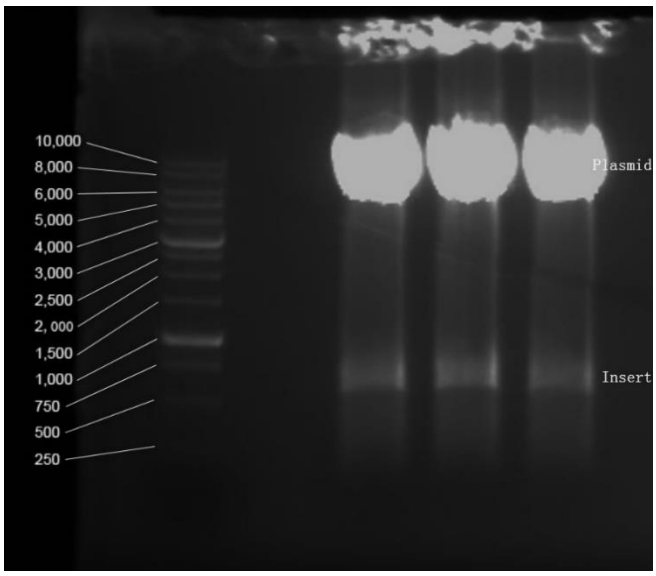


Figure 8.5 Gel electrophoresis result for the *JHDK*-pET-11a double-digest.

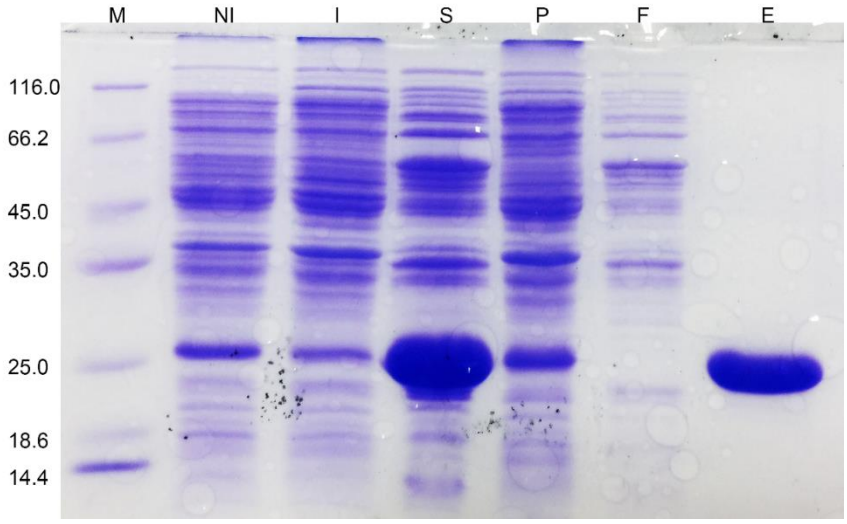


Figure 8.6 SDS-PAGE for *Bommo*-JHDK expression and purification. M, NI, I, S, P, F and E indicate the marker, non-induced, induced, supernatant, pellet, flow-through and eluted samples, respectively.

8.4.2 Crystallisation

As described in section 8.3.3, *Bommo*-JHDK crystals can be obtained reproducibly in 0.1 M Bis-Tris pH 5.5, 25% (w/v) PEG 3,350, however, it takes approximately 5 months to grow these crystals to a suitable size for diffraction.

Running an SDS-PAGE with the dissolved crystals showed a band with a molecular mass of around 22 kDa which is about 2-3 kDa smaller than the sample after purification. Thus the protein might suffer some proteolysis during long time crystallisation process in which some of the N-terminal residues were cleaved off which was essential for the crystallisation. The same phenomenon was observed in a homologous protein, calexicitin, which was expressed with a pET-16b plasmid as well (Beaven *et al.*, 2005; Erskine *et al.*, 2015).

Due to the long time required for crystallisation, many efforts were made in order to get crystals in a shorter time. Successful methylation for lysine residues, chymotrypsin and trypsin proteolysis are indicated by lanes MT, C and T in Figure 8.7. For those treated with chymotrypsin and trypsin, different fragments were separated by gel-filtration and were subjected to crystal screen separately. Different metal ions were also added during crystallisation. However, no crystal was obtained.

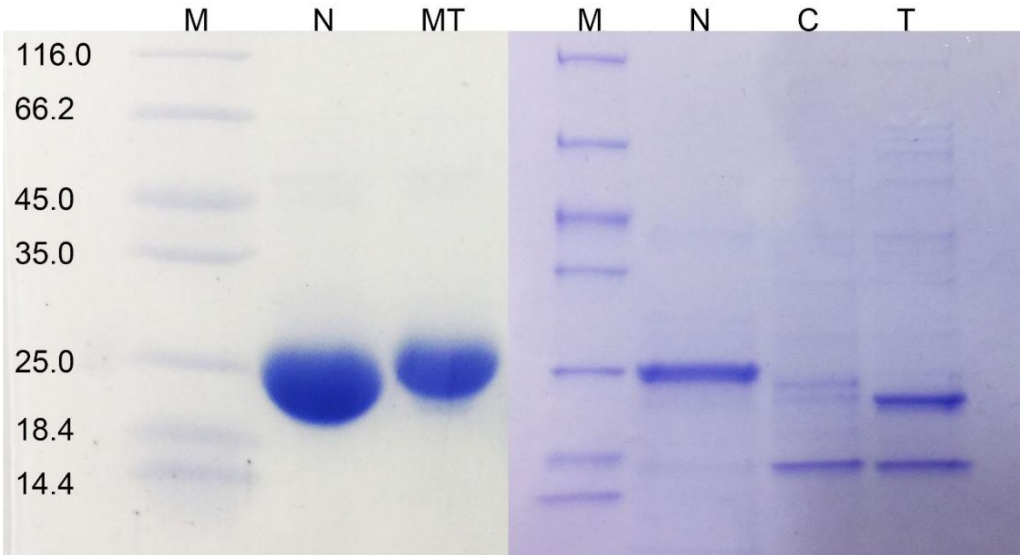


Figure 8.7 Lysine methylation and proteolysis of *Bommo*-JHDK by chymotrypsin and trypsin. The marker, native, methylated, chymotrypsinolysed and trypsinolysed samples are indicated by M, N, MT, C and T, respectively.

8.4.3 Attempts to determine the structure

Molecular replacement did not give any correct solution. Possible reasons could be 1) the protein may have a completely different structure from that of the 'homologues' identified by sequence comparison, ignoring that the closest structure (calexcitin, PDB ID: 2CCM) has only 36% sequence identity; 2) the protein may have many local/domain movements which frustrate the programs.

The auto-processed anomalous data are of poor quality which is indicated by statistics such as the low anomalous completeness and the negative anomalous correlation. As shown in Figure 8.8, the anomalous signal for the data collected at the 'if' and 'hrm' points is only significant at resolution lower than 10 Å, whilst that of the pk data is not significant at all. The data collected at the 'pk' point suffered some radiation damage due to a not appropriate data collection strategy in which they were collected after those collected at the 'if' and 'hrm' points, perhaps a more gentle strategy needs to be used next time. Data reprocessing can improve the anomalous completeness to over 86% for both the inner and outer shells at 7 Å. However, the anomalous correlation is still lower than 0.25 for the inner shell and negative for the outer shell. *MAD* phasing and auto building using the auto-processed and reprocessed data failed to give a correction solution as indicated by a *FOM* values below 30.

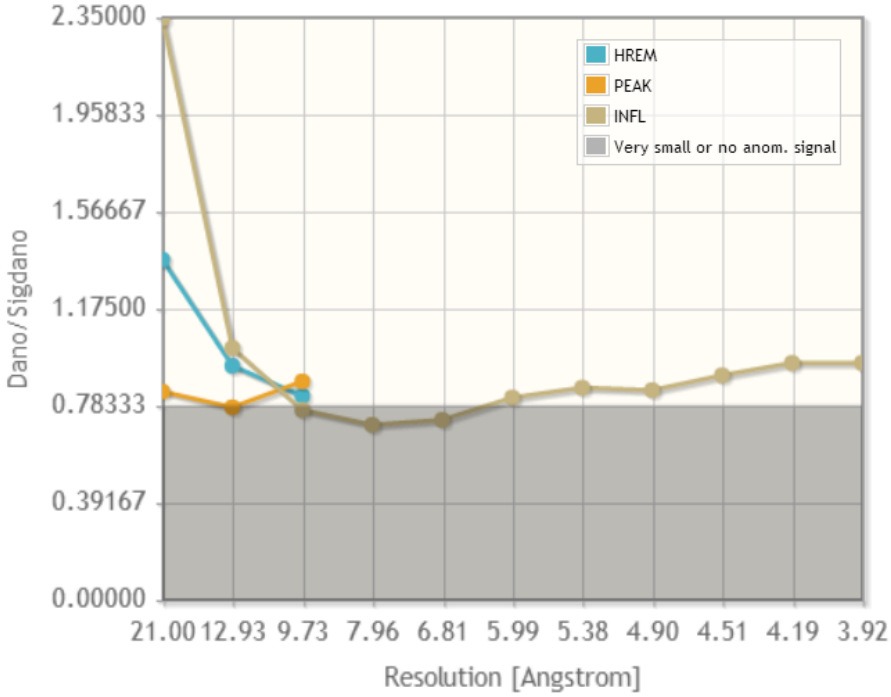


Figure 8.8 Anomalous signal analysis against resolution.

8.5 Future work

1. Better data collection strategies can be carried out in future in order to collect as much data as possible with a minimum radiation damage.

2. It is worth making a selenomethionine derivative protein in which the possible heavy atom sites are already known.

3. With the development of technologies, other phasing methods such as phasing with calcium or sulphur atoms become more feasible and may help to determine the *Bommo*-JHDK structure. The recently available long-wavelength MX beamline I23 at DLS is designed for solving crystallographic phase problem using the small anomalous signals from sulphur or phosphorous, etc.

Summary

X-ray crystallography is one of the dominant methods in studying protein structures which are essential in understanding their functions. Seven different proteins have been studied by X-ray crystallography combined with other methods and the information obtained can be used as guidance for applications in many areas such as drug discovery and biological industry.

The outbreaks of human epidemic nonbacterial gastroenteritis are mainly caused by noroviruses. The ligand-free crystal structure of SV3CP has been determined to 1.30 Å and a system for growing high-quality crystals has been established. The non-covalent compounds that have been identified by the *in-crystallo* screening with SV3CP give useful information for engineering novel compounds as therapeutic agents. Screening with covalent fragments is on-going.

The enzyme PBGD is one of the key enzymes in tetrapyrrole biosynthesis. It catalyses the formation of a linear tetrapyrrole from four molecules of the substrate PBG. Three mutations have been made affecting Asp82 in BPBGD which are D82A, D82E and D82N and their crystal structures have been determined at resolutions of 2.7, 1.8 and 1.9 Å, respectively. These structures reveal that whilst the D82E mutant possesses the DPM cofactor, in the D82N and D82A mutants, the cofactor is likely to be missing or disordered. Comparison of the mutant BPBGD structures with that of the WT enzyme shows that there are significant domain movements and suggests that the enzyme adopts “open” and “closed” conformations, potentially in response to substrate binding.

PDI is a glycoprotein which inhibits both the aspartic protease cathepsin D and the serine protease trypsin. The first crystal structure of PDI has been determined to a resolution of 2.1 Å, revealing that PDI adopts a typical *β*-trefoil fold with a

core and several protruding inhibitory loops, which is typical of the Kunitz-family protease inhibitors. The NAG moiety was confirmed by clear electron density. Possible reactive-site loops for cathepsin D and trypsin have been studied which indicate that PDI is an unusual bi-functional inhibitor.

Investigations of the other three thermostable enzymes gave insight into their catalytic mechanism and factors contributing to the high thermostability. Pc-polymerase is a family B DNA polymerase from *Pyrobaculum calidifontis*. The crystal structure of Pc-polymerase has been refined to a resolution of 2.8 Å and several unique features have been identified which may account for its high processivity. A complex model with the primer-template duplex of DNA suggests the large movement of the thumb domain upon DNA binding. The high thermostability may be attributed to the large number of salt bridges.

TK-PUL is in the class of pullulan-hydrolysing enzymes that have been widely used in starch saccharification industries. The crystal structure of TK-PUL, which would be the first structure of a type III pullulan hydrolyse, has been determined which revealed that the structure of the last part of the N-terminal and the C-terminal domains are different from the homologous structures and the loop regions at the active site end of the catalytic domain are quite different. The complex structure suggested that n-Dodecyl *α*-D-maltoside may be a substrate for TK-PUL, however, no further information can be obtained due to the low resolution. The calcium binding site and the rare vicinal disulphide bridge have also been studied. The thermostability of TK-PUL and a few homologs may be attribute to several factors including the increased content of salt bridges, helical segments, Pro, Arg and Tyr and the decreased content of Ser.

TkA catalyzes the hydrolysis of asparagine to aspartic acid and ammonia. The crystal structure of TkA has been determined at 2.2 Å. Many residues from the neighboring molecules in a dimer are involved in substrate recognition as well as catalysis. The N-terminal domain contains a highly flexible *β*-hairpin which adopts 'open' and 'closed' conformations. The *β*-hairpin are observed to adopt different conformations in different subunits of TkA structure. The *β*-hairpin was usually only observed in L-asparaginase structures that adopt a 'closed' conformation whilst it is characterised with good electron density in all the subunits in TkA structure. One phosphate ion has been built in the active site formed by the neighbouring subunits. The great thermostability of TkA may be attributed to the higher Arg content, lower numbers of Cys and Ser, but mainly to the increased content of salt bridges.

Juvenile hormone diol kinases (JHDKs) are a class of enzymes which are involved in the JH degradation pathway. Proteins in this family usually contain GTP-binding motifs and EF-hands that bind calcium ions. The expression, purification and crystallisation systems for *Bommo*-JHDK have been established, however, more efforts are required to determine the crystal structure.

The phasing method molecular replacement was predominantly involved in the determination of the structures summarized above. Since the introduction of the original concept of molecular replacement in the early 1960s (Huber, 1965; Rossmann and Blow, 1962), it has developed impressively, partly due to the dramatic evolution of computer hardware and software, which has resulted in a faster, more flexible and in many cases fully automated methodology. In addition to the relatively fast programs *Molrep* and *Phaser MR*, pipelines such as *MrBUMP* and *BALBES* are available which automate the process of structure solution

including model selection, preparation, molecular replacement, refinement, etc. *Mr_Rosetta* (DiMaio *et al*., 2011) takes advantage of the structure-modelling field and combines that with crystallographic molecular replacement, model-building, density modification and refinement. *ARCIMBOLDO* (Rodriguez *et al*., 2009) and *AMPLE* (Bibby *et al*., 2012) attempt *ab initio* modelling by performing molecular replacement solution with small fragments and then build up full structures from these fragments. *SIMBAD* carries out sequence independent molecular replacement based on the available database and is good to identify contaminant proteins or to help find homologous structures in difficult cases.

The structure determination (or attempts) of the proteins involved in the projects were mainly based on molecular replacement but each was slightly different. In the most straightforward cases where there was high sequence identity between the search model and the target structure (e.g. BPBGD mutants, PDI and TkA), the structure was determined by simply running *Phaser MR* or *Molrep*. For the SV3CP structure, previous attempts (by many) to determine the structure in the $P6_1$ space group (2.8 Å) by molecular replacement using the 2IPH structure (protein-inhibitor complex) as the search model failed to give a correct solution, which might be due to a combination of packing conflicts and structural difference (e.g. the moved *β*-hairpin and the loop which adopts different conformations in different chains). Truncating the last 8 residues (which are not observed in the inhibitor-free structure) did give a correct solution in *Phaser MR* and clashes between these residues from neighbouring molecules were observed in the *C*2 and $P6_1$ structures by superposing them with the 2IPH structure separately. Attempts to determine the structure of Pc-polymerase by *Phaser MR* and *Molrep* did not give a correct solution due to the large domain movements in the target

structure compared with the search model. The problem was solved by using the *BALBES* online service which searched for the domains separately with models generated from different DNA polymerases. The structure determination of the last 500 or so residues for TK-PUL was performed by using the *MrBUMP* web service. The N-terminal residues 185-280 were identified by a combination of the phase information from the partial structure with the use of *Molrep* (SAPTF + local phased RF + phased TF) to position a search model generated from a remote homolog identified by *HHpred*. Attempts to determine the JHDK structure by both molecular replacement and *MAD* have not yet succeeded. In addition to the programs and online services such as *Phaser MR*, *MrBUMP* and *BALBES*, *SIMBAD* has been used but this did not give any plausible solutions. Efforts to determine the structure by *ARCIMBOLDO* were hampered due to the lack of a grid or supercomputer. While other molecular replacement services, such as *Mr_Rosetta*, *AMPLE* and experimental phasing methods are worth trying as future work. The structure determination and challenges encountered are summarized in table S.

Table S Summary for structure determination and challenges.

Target	Resolution (Å)	Space group	Phasing program	Search model	Sequence identity (%)	Challanges	Solved?
SV3CP	1.30	*C*2	*Phaser MR*	2IPH	100	Packing conflicts	Yes
BPBGD	1.81, 1.87, 2.76	$P2_12_12_1$	*Molrep*	4MLV	99	No	Yes
PDI	2.11, 2.83	*C*2 $C222_1$	*Molrep*	3TC2	73	No	Yes
Pc-polymerase	2.80	$P2_1$	*BALBES*	2XHB 3A2F 1WN7	35 36 35	Large domain movements, various structural similarity for different domains.	Yes
TK-PUL	2.80	*C*2	*MrBUMP, Molrep*	2Z1K 4AEE	37 10	Large domain movements, low sequence identity for the second searh model.	Yes
TkA	2.18	*P*1	*Phaser MR*	1WLS	59	No	Yes
JHDK	1.99	$P2_1$	*Phaser MR*, *MrBUMP*, *BALBES*, *SIMBAD*, *SHELX*, *CRANK2*	Many	Up to 35	Bad search model? Weak anomalous signal.	No

References

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C., Zwart, P.H., 2010. *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica. Section D, Structural Biology* 66, 213-221.

Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T.C., Urzhumtsev, A., Zwart, P.H., Adams, P.D., 2012. Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallographica. Section D, Structural Biology* 68, 352-367.

Ahmad, N., Rashid, N., Haider, M.S., Akram, M., Akhtar, M., 2014. Novel maltotriose-hydrolyzing thermoacidophilic type III pullulan hydrolase from *Thermococcus kodakarensis*. *Applied and Environmental Microbiology* 80, 1108-1115.

Albertsen, B., Jakobsen, P., Schrøder, H., Schmiegelow, K., Carlsen, N.T., 2001. Pharmacokinetics of *Erwinia* asparaginase after intravenous and intramuscular administration. *Cancer Chemotherapy and Pharmacology* 48, 77-82.

Ali, S.F., Rashid, N., Imanaka, T., Akhtar, M., 2011. Family B DNA polymerase from a hyperthermophilic archaeon *Pyrobaculum calidifontis*: Cloning, characterization and PCR application. *Journal of Bioscience and Bioengineering* 112, 118-123.

Amo, T., Paje, M.L.F., Inagaki, A., Ezaki, S., Atomi, H., Imanaka, T., 2002. *Pyrobaculum calidifontis* sp. nov., a novel hyperthermophilic archaeon that grows in atmospheric air. *Archaea* 1, 113-121.

Anand, K., Palm, G.J., Mesters, J.R., Siddell, S.G., Ziebuhr, J., Hilgenfeld, R., 2002. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra *α*‐helical domain. *The EMBO Journal* 21, 3213-3224.

Anderson, P.M., Desnick, R.J., 1980. Purification and properties of uroporphyrinogen I synthase from human erythrocytes. Identification of stable enzyme-substrate intermediates. *The Journal of Biological Chemistry* 255, 1993-1999.

Andino, R., Rieckhof, G.E., Baltimore, D., 1990. A functional ribonucleoprotein complex forms around the 5' end of poliovirus RNA. *Cell* 63, 369-380.

Aoki, H., Sakano, Y., 1997. A classification of dextran-hydrolysing enzymes based on amino-acid-sequence similarities. *Biochemical Journal* 323, 859.

Argyris, E.G., Dornadula, G., Nunnari, G., Acheampong, E., Zhang, C., Mehlman, K., Pomerantz, R.J., Zhang, H., 2006. Inhibition of endogenous reverse transcription of human and nonhuman primate lentiviruses: potential for development of lentivirucides. *Virology* 353, 482-490.

Arndt, U., Champness, J., Phizackerley, R., Wonacott, A., 1973. A single-crystal oscillation camera for large unit cells. *Journal of Applied crystallography* 6, 457-463.

Atha, D.H., Ingham, K.C., 1981. Mechanism of precipitation of proteins by polyethylene glycols. Analysis in terms of excluded volume. *Journal of Biological Chemistry* 256, 12108-12117.

Atkins, C.A., Pate, J.S., Sharkey, P.J., 1975. Asparagine metabolism - key to the nitrogen nutrition of developing legume seeds. *Plant Physiology* 56, 807-812.

Atmar, R.L., Opekun, A.R., Gilger, M.A., Estes, M.K., Crawford, S.E., Neill, F.H., Graham, D.Y., 2008. Norwalk virus shedding after experimental human infection. *Emerging Infectious Diseases* 14, 1553-1557.

Avramis, V.I., Sencer, S., Periclou, A.P., Sather, H., Bostrom, B.C., Cohen, L.J., Ettinger, A.G., Ettinger, L.J., Franklin, J., Gaynon, P.S., 2002. A randomized comparison of native *Escherichia coli* asparaginase and polyethylene glycol conjugated asparaginase for treatment of children with newly diagnosed standard-risk acute lymphoblastic leukemia: a Children's Cancer Group study. *Blood* 99, 1986-1994.

Awan, S.J., Siligardi, G., Shoolingin-Jordan, P.M., Warren, M.J., 1997. Reconstitution of the holoenzyme form of *Escherichia coli* porphobilinogen deaminase from apoenzyme with porphobilinogen and preuroporphyrinogen: a study using circular dichroism spectroscopy. *Biochemistry* 36, 9273-9282.

Axford, D., Owen, R.L., Aishima, J., Foadi, J., Morgan, A.W., Robinson, J.I., Nettleship, J.E., Owens, R.J., Moraes, I., Fry, E.E., 2012. *In situ* macromolecular crystallography using microbeams. *Acta Crystallographica. Section D, Structural Biology* 68, 592-600.

Azarkan, M., Dibiani, R., Goormaghtigh, E., Raussens, V., Baeyens-Volant, D., 2006. The papaya Kunitz-type trypsin inhibitor is a highly stable *β*-sheet glycoprotein. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1764, 1063-1072.

Azim, N., Deery, E., Warren, M.J., Erskine, P., Cooper, J.B., Wood, S.P., Akhtar, M., 2013. Crystallization and preliminary X-ray characterization of the tetrapyrrole-biosynthetic enzyme porphobilinogen deaminase from *Bacillus megaterium*. *Acta Crystallographica. Section F, Structural Biology Communications* 69, 906-908.

Azim, N., Deery, E., Warren, M.J., Wolfenden, B.A., Erskine, P., Cooper, J.B., Coker, A., Wood, S.P., Akhtar, M., 2014. Structural evidence for the partially oxidized dipyrromethene and dipyrromethanone forms of the cofactor of porphobilinogen deaminase: structures of the *Bacillus megaterium* enzyme at near-atomic resolution. *Acta Crystallographica. Section D, Biological Crystallography* 70, 744-751.

Baert, L., Uyttendaele, M., Stals, A., Van Coillie, E., Dierick, K., Debevere, J., Botteldoorn, N., 2009. Reported foodborne outbreaks due to noroviruses in

Belgium: the link between food and patient investigations in an international context. *Epidemiology and Infection* 137, 316-325.

Balzar, D., 1992. Profile fitting of X-ray diffraction lines and Fourier analysis of broadening. *Journal of Applied Crystallography* 25, 559-570.

Bao, R., Zhou, C.Z., Jiang, C., Lin, S.X., Chi, C.W., Chen, Y., 2009. The ternary structure of the double-headed arrowhead protease inhibitor API-A complexed with two trypsins reveals a novel reactive site conformation. *The Journal of Biological Chemistry* 284, 26676-26684.

Bartenschlager, R., Ahlborn-Laake, L., Mous, J., Jacobsen, H., 1994. Kinetic and structural analyses of hepatitis C virus polyprotein processing. *Journal of Virology* 68, 5045-5055.

Barton, G.J., 1993. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Engineering* 6, 37-40.

Bartsch, S.M., Lopman, B.A., Ozawa, S., Hall, A.J., Lee, B.Y., 2016. Global Economic Burden of Norovirus Gastroenteritis. *PloS One* 11, e0151219.

Battye, T.G., Kontogiannis, L., Johnson, O., Powell, H.R., Leslie, A.G., 2011. *iMOSFLM*: a new graphical interface for diffraction-image processing with *MOSFLM*. *Acta Crystallographica. Section D, Biological Crystallography* 67, 271-281.

Bazan, J.F., Fletterick, R.J., 1988. Viral cysteine proteases are homologous to the trypsin-like family of serine proteases: structural and functional implications. *Proceedings of the National Academy of Sciences of the United States of America* 85, 7872-7876.

Beaven, G.D., Erskine, P.T., Wright, J.N., Mohammed, F., Gill, R., Wood, S.P., Vernon, J., Giese, K.P., Cooper, J.B., 2005. Crystallization and preliminary X-ray diffraction analysis of calexcitin from *Loligo pealei*: a neuronal protein implicated in learning and memory. *Acta Crystallographica. Section F, Structural Biology Communications* 61, 879-881.

Bell-McGuinn, K.M., Garfall, A.L., Bogyo, M., Hanahan, D., Joyce, J.A., 2007. Inhibition of cysteine cathepsin protease activity enhances chemotherapy regimens by decreasing tumor growth and invasiveness in a mouse model of multistage cancer. *Cancer Research* 67, 7378-7385.

Benes, P., Vetvicka, V., Fusek, M., 2008. Cathepsin D - many functions of one aspartic protease. *Crit Rev Oncol Hematol* 68, 12-28.

Bergmann, E.M., Mosimann, S.C., Chernaia, M.M., Malcolm, B.A., James, M., 1997. The refined crystal structure of the 3C gene product from hepatitis A virus: specific proteinase activity and RNA recognition. *Journal of Virology* 71, 2436-2448.

Bernal, J.D., Crowfoot, D., 1934. X-ray photographs of crystalline pepsin. *Nature* 133, 794-795.

Bernstein, D.I., Atmar, R.L., Lyon, G.M., Treanor, J.J., Chen, W.H., Jiang, X., Vinjé, J., Gregoricus, N., Frenck, R.W., Moe, C.L., 2015. Norovirus vaccine against experimental human GII. 4 virus illness: a challenge study in healthy adults. *Journal of Infectious Diseases* 211, 870-878.

Bertolotti-Ciarlet, A., Crawford, S.E., Hutson, A.M., Estes, M.K., 2003. The 3' end of Norwalk virus mRNA contains determinants that regulate the expression and stability of the viral capsid protein VP1: a novel function for the VP2 protein. *Journal of Virology* 77, 11603-11615.

Bhat, A., Roberts, L.R., Dwyer, J.J., 2015. Lead discovery and optimization strategies for peptide macrocycles. *European Journal of Medicinal Chemistry* 94, 471-479.

Bibby, J., Keegan, R.M., Mayans, O., Winn, M.D., Rigden, D.J., 2012. AMPLE: a cluster-and-truncate approach to solve the crystal structures of small proteins using rapidly computed ab initio models. *Acta Crystallographica. Section D, Biological Crystallography* 68, 1622-1631.

Birtley, J.R., Knox, S.R., Jaulent, A.M., Brick, P., Leatherbarrow, R.J., Curry, S., 2005. Crystal structure of foot-and-mouth disease virus 3C protease new insights into catalytic mechanism and cleavage specificity. *Journal of Biological Chemistry* 280, 11520-11527.

Blakeley, M.P., Ruiz, F., Cachau, R., Hazemann, I., Meilleur, F., Mitschler, A., Ginell, S., Afonine, P., Ventura, O.N., Cousido-Siah, A., Haertlein, M., Joachimiak, A., Myles, D., Podjarny, A., 2008. Quantum model of catalysis based on a mobile

proton revealed by subatomic X-ray and neutron diffraction studies of h-aldose reductase. *Proceedings of the National Academy of Sciences of the United States of America* 105, 1844-1848.

Blakeney, S.J., Cahill, A., Reilly, P.A., 2003. Processing of Norwalk virus nonstructural proteins by a 3C-like cysteine proteinase. *Virology* 308, 216-224.

Blow, D., Crick, F., 1959. The treatment of errors in the isomorphous replacement method. *Acta Crystallographica* 12, 794-802.

Blow, D.M., 2003. How Bijvoet made the difference: the growing power of anomalous scattering. *Methods Enzymol* 374, 3-22.

Bogin, O., Peretz, M., Hacham, Y., Burstein, Y., Korkhin, Y., Frolow, F., 1998. Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brockii* alcohol dehydrogenase. *Protein Science* 7, 1156-1163.

Boniotti, B., Wirblich, C., Sibilia, M., Meyers, G., Thiel, H.-J., Rossi, C., 1994. Identification and characterization of a 3C-like protease from rabbit hemorrhagic disease virus, a calicivirus. *Journal of Virology* 68, 6487-6495.

Boyd, J.W., Phillips, A.W., 1971. Purification and properties of L-asparaginase from *Serratia marcescens*. *Journal of Bacteriology* 106, 578-587.

Brautigam, C.A., Steitz, T.A., 1998. Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes. *Current Opinion in Structural Biology* 8, 54-63.

Broome, J., 1961. Evidence that the L-asparaginase activity of guinea pig serum is responsible for its antilymphoma effects.

Broome, J., 1963. Evidence that the L-asparaginase of guinea pig serum is responsible for its antilymphoma effects: II. Lymphoma 6C3HED cells cultured in a medium devoid of L-asparagine lose their susceptibility to the effects of guinea pig serum *in vivo*. *The Journal of Experimental Medicine* 118, 121.

Bugg, T.D., Braddick, D., Dowson, C.G., Roper, D.I., 2011. Bacterial cell wall assembly: still an attractive antibacterial target. *Trends in Biotechnology* 29, 167-173.

Carroll, J.O., Boyce, C.O., Wong, T.M., Starace, C.A. 1987. Bread antistaling method. *Google Patents*.

Cater, S.A., Lees, W.E., Hill, J., Brzin, J., Kay, J., Phylip, L.H., 2002. Aspartic proteinase inhibitors from tomato and potato are more potent against yeast proteinase A than cathepsin D. *Biochim Biophys Acta* 1596, 76-82.

Cedar, H., Schwartz, J.H., 1967. Localization of the two L-asparaginases in anaerobically grown *Escherichia coli*. *Journal of Biological Chemistry* 242, 3753-3755.

Cedar, H., Schwartz, J.H., 1968. Production of L-asparaginase II by *Escherichia coli*. *Journal of Bacteriology* 96, 2043-2048.

Chadwick, D.J., Ackrill, K., 1994. The biosynthesis of the tetrapyrrole pigments *John Wiley & Sons Ltd*.

Champenois, Y., Della Valle, G., Planchot, V., Buleon, A., Colonna, P., 1999. Influence of *α*-amylases on bread staling and on retrogradation of wheat starch models. *Sciences des Aliments* 19, 471-486.

ChemAxon, L. 2013. *Marvin Sketch* 6.0.0.

Chen, J.J., London, I.M., 1995. Regulation of protein synthesis by heme-regulated eIF-2 alpha kinase. *Trends in Biochemical Sciences* 20, 105-108.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C., 2010. *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica. Section D, Structural Biology* 66, 12-21.

Chien, A., Edgar, D.B., Trela, J.M., 1976. Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus*. *Journal of Bacteriology* 127, 1550-1557.

Chohan, S.M., Rashid, N., 2013. TK1656, a thermostable l-asparaginase from *Thermococcus kodakaraensis*, exhibiting highest ever reported enzyme activity. *Journal of Bioscience and Bioengineering* 116, 438-443.

Clarke, I., Estes, M., Green, K., Hansman, G., Knowles, N., Koopmans, M., Matson, D., Meyers, G., Neill, J., Radford, A., 2012. Caliciviridae. *Virus Taxonomy*, 977-986.

Clarke, I.N., Lambden, P.R., 1997. The molecular biology of caliciviruses. *The Journal of General Virology* 78 (Pt 2), 291-301.

Coates, L., Erskine, P.T., Mall, S., Gill, R., Wood, S.P., Myles, D.A., Cooper, J.B., 2006. X-ray, neutron and NMR studies of the catalytic mechanism of aspartic proteinases. *European Biophysics Journal : EBJ* 35, 559-566.

Collins, P.M., Ng, J.T., Talon, R., Nekrosiute, K., Krojer, T., Douangamath, A., Brandao-Neto, J., Wright, N., Pearce, N.M., von Delft, F., 2017. Gentle, fast and effective crystal soaking by acoustic dispensing. *Acta Crystallographica. Section D, Structural Biology* 73, 246-255.

Comeau, S.R., Gatchell, D.W., Vajda, S., Camacho, C.J., 2004a. *ClusPro*: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* 20, 45-50.

Comeau, S.R., Gatchell, D.W., Vajda, S., Camacho, C.J., 2004b. *ClusPro*: a fully automated algorithm for protein - protein docking. *Nucleic Acids Research* 32, W96-W99.

Consortium, P.G.S., 2011. Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189-195.

Cooper, J.B., Lilliston, M., Brooks, D., Swords, B., 2014. Experience with a pharmacy technician medication history program. *American Journal of Health-system Pharmacy : AJHP : Official Journal of the American Society of Health-System Pharmacists* 71, 1567-1574.

Copeland, W.C., Wang, T.S., 1993. Mutational analysis of the human DNA polymerase alpha. The most conserved region in alpha-like DNA polymerases is involved in metal-specific catalysis. *The Journal of Biological Chemistry* 268, 11028-11040.

Correa, P., 1981. Epidemiological correlations between diet and cancer frequency. *Cancer Research* 41, 3685-3689.

Costantini, S., Colonna, G., Facchiano, A.M., 2008. *ESBRI*: a web server for evaluating salt bridges in proteins. *Bioinformation* 3, 137-138.

Courter, J.D., Girotto, J.E., Salazar, J.C., 2008. Tipranavir: a new protease inhibitor for the pediatric population. *Expert Review of Anti-infective Therapy* 6, 797-803.

Cowtan, K., 2006. The *Buccaneer* software for automated model building. 1. Tracing protein chains. *Acta Crystallographica. Section D, Structural Biology* 62, 1002-1011.

Cox, O.B., Krojer, T., Collins, P., Monteiro, O., Talon, R., Bradley, A., Fedorov, O., Amin, J., Marsden, B.D., Spencer, J., 2016. A poised fragment library enables

rapid synthetic expansion yielding the first reported inhibitors of PHIP (2), an atypical bromodomain. *Chemical Science* 7, 2322-2330.

Cristofoli, W.A., Wiebe, L.I., De Clercq, E., Andrei, G., Snoeck, R., Balzarini, J., Knaus, E.E., 2007. 5-Alkynyl Analogs of Arabinouridine and 2'-Deoxyuridine: Cytostatic Activity against Herpes Simplex Virus and Varicella-Zoster Thymidine Kinase Gene-Transfected Cells. *Journal of Medicinal Chemistry* 50, 2851-2857.

Crowther, R., 1972. The fast rotation function. *International Science Reviews Series* 13, 173-178.

Cullis, A.F., Muirhead, H., Perutz, M., Rossmann, M., North, A. 1961. The Structure of Haemoglobin. VIII. A Three-Dimensional Fourier Synthesis at 5.5 Å Resolution: Determination of the Phase Angles, pp. 15-38 Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, Vol. 265. *The Royal Society*.

Damalanka, V.C., Kim, Y., Kankanamalage, A.C.G., Lushington, G.H., Mehzabeen, N., Battaile, K.P., Lovell, S., Chang, K.-O., Groutas, W.C., 2017. Design, synthesis, and evaluation of a novel series of macrocyclic inhibitors of norovirus 3CL protease. *European Journal of Medicinal Chemistry* 127, 41-61.

Dauter, Z., 2006. Estimation of anomalous signal in diffraction data. *Acta Crystallographica. Section D, Biological Crystallography* 62, 867-876.

Dauter, Z., Adamiak, D.A., 2001. Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy. *Acta Crystallographica. Section D, Structural Biology* 57, 990-995.

Davidson, L., Brear, D.R., Wingard, P., Hawkins, J., Kitto, G.B., 1977. Purification and properties of L-glutaminase-L-asparaginase from *Pseudomonas acidovorans*. *Journal of Bacteriology* 129, 1379-1386.

De Clercq, E., 2013. Dancing with chemical formulae of antivirals: A panoramic view (Part 2). *Biochemical Pharmacology* 86, 1397-1410.

De Clercq, E., 2014. Current race in the development of DAAs (direct-acting antivirals) against HCV. *Biochemical Pharmacology* 89, 441-452.

De Clercq, E., 2015. Development of antiviral drugs for the treatment of hepatitis C at an accelerating pace. *Reviews in Medical Virology* 25, 254-267.

De Strooper, B., 2010. Proteases and proteolysis in Alzheimer disease: a multifactorial view on the disease process. *Physiological Reviews* 90, 465-494.

Delarue, M., Poch, O., Tordo, N., Moras, D., Argos, P., 1990. An attempt to unify the structure of polymerases. *Protein Engineering* 3, 461-467.

Delfau, M.H., Picat, C., de Rooij, F.W., Hamer, K., Bogard, M., Wilson, J.H., Deybach, J.C., Nordmann, Y., Grandchamp, B., 1990. Two different point G to A mutations in exon 10 of the porphobilinogen deaminase gene are responsible for acute intermittent porphyria. *The Journal of Clinical Investigation* 86, 1511-1516.

Desai, R., Hembree, C.D., Handel, A., Matthews, J.E., Dickey, B.W., McDonald, S., Hall, A.J., Parashar, U.D., Leon, J.S., Lopman, B., 2012. Severe outcomes are associated with genogroup 2 genotype 4 norovirus outbreaks: a systematic literature review. *Clinical Infectious Diseases* 55, 189-193.

Diederichs, K., Karplus, P.A., 1997. Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nature Structural Biology* 4, 269-275.

DiMaio, F., Terwilliger, T.C., Read, R.J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H.L., 2011. Improved molecular replacement by density-and energy-guided protein structure optimization. *Nature* 473, 540-543.

Division of Viral Diseases, N.C.f.I., Respiratory Diseases, C.f.D.C., Prevention, 2011. Updated norovirus outbreak management and disease prevention guidelines. *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and Reports* 60, 1-18.

Doerr, A., 2006. Widening the protein crystallization bottleneck. *Nature Methods* 3, 961.

Döbelin, N., 2014. X-rays & Diffraction. [pdf slides]. Retrieved from

http://profex.doebelin.org/wp-content/uploads/2014/02/Lesson-1-X-rays-and-Diffraction.pdf

Drenth, J., 2007. Principles of protein X-ray crystallography *Springer Science & Business Media*.

Duffner, F., Bertoldo, C., Andersen, J.T., Wagner, K., Antranikian, G., 2000. A new thermoactive pullulanase from *Desulfurococcus mucosus*: cloning, sequencing, purification, and characterization of the recombinant enzyme after expression in *Bacillus subtilis*. *Journal of Bacteriology* 182, 6331-6338.

Echols, N., Grosse-Kunstleve, R.W., Afonine, P.V., Bunkoczi, G., Chen, V.B., Headd, J.J., McCoy, A.J., Moriarty, N.W., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C., Adams, P.D., 2012. Graphical tools for macromolecular crystallography in *PHENIX*. *Journal of Applied Crystallography* 45, 581-586.

Emsley, P., Cowtan, K., 2004. *Coot*: model-building tools for molecular graphics. *Acta Crystallographica. Section D, Structural Biology* 60, 2126-2132.

Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, K., 2010. Features and development of *Coot*. *Acta Crystallographica. Section D, Structural Biology* 66, 486-501.

Enzo, S., Fagherazzi, G., Benedetti, A., Polizzi, S., 1988. A profile-fitting procedure for analysis of broadened X-ray diffraction peaks. I. Methodology. *Journal of Applied Crystallography* 21, 536-542.

Erskine, P., Fokas, A., Muriithi, C., Rehman, H., Yates, L., Bowyer, A., Findlow, I., Hagan, R., Werner, J., Miles, A.J., 2015. X-ray, spectroscopic and normal-mode dynamics of calexcitin: structure–function studies of a neuronal calcium-signalling protein. *Acta Crystallographica Section D, Structural Biology* 71, 615-631.

Erskine, P.T., Beaven, G.D., Hagan, R., Findlow, I.S., Werner, J.M., Wood, S.P., Vernon, J., Giese, K.P., Fox, G., Cooper, J.B., 2006. Structure of the neuronal protein calexcitin suggests a mode of interaction in signalling pathways of learning and memory. *Journal of Molecular Biology* 357, 1536-1547.

Evans, P., 2006. Scaling and assessment of data quality. *Acta Crystallographica. Section D, Structural Biology* 62, 72-82.

Evans, P.R., 2011. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta crystallographica. Section D, Structural Biology* 67, 282-292.

Evans, P.R., Murshudov, G.N., 2013a. How good are my data and what is the resolution? *Acta Crystallographica. Section D, Structural Biology* 69, 1204-1214.

Evans, P.R., Murshudov, G.N., 2013b. How good are my data and what is the resolution? *Acta Crystallographica. Section D: Structural Biology* 69, 1204-1214.

Florescu, D.F., Hill, L.A., McCartan, M.A., Grant, W., 2008. Two cases of Norwalk virus enteritis following small bowel transplantation treated with oral human serum immunoglobulin. *Pediatric Transplantation* 12, 372-375.

Fogarty, W.M., Kelly, C.T., 1990. Recent advances in microbial amylases, Microbial enzymes and biotechnology, *Springer* 71-132,

Forterre, P., Brochier, C., Philippe, H., 2002. Evolution of the Archaea. *Theoretical Population Biology* 61, 409-422.

Frange, P., Touzot, F., Debre, M., Heritier, S., Leruez-Ville, M., Cros, G., Rouzioux, C., Blanche, S., Fischer, A., Avettand-Fenoel, V., 2012. Prevalence and clinical impact of norovirus fecal shedding in children with inherited immune deficiencies. *The Journal of Infectious Diseases* 206, 1269-1274.

French, S., Wilson, K., 1978. On the treatment of negative intensity observations. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 34, 517-525.

Frenck, R., Bernstein, D.I., Xia, M., Huang, P., Zhong, W., Parker, S., Dickey, M., McNeal, M., Jiang, X., 2012. Predicting susceptibility to norovirus GII.4 by use of a challenge model involving humans. *The Journal of Infectious Diseases* 206, 1386-1393.

Friedman, M., 2003. Chemistry, biochemistry, and safety of acrylamide. A review. *Journal of Agricultural and Food Chemistry* 51, 4504-4526.

Friesema, I., Vennema, H., Heijne, J., de Jager, C., Teunis, P., van der Linde, R., Duizer, E., van Duynhoven, Y., 2009. Differences in clinical presentation between norovirus genotypes in nursing homes. *Journal of Clinical Virology* 46, 341-344.

Fujinaga, M., Read, R., 1987. Experiences with a new translation-function program. *Journal of Applied Crystallography* 20, 517-521.

Fullmer, A., O'Brien, S., Kantarjian, H., Jabbour, E., 2010. Emerging therapy for the treatment of acute lymphoblastic leukemia. *Expert Opinion on Emerging Drugs* 15, 1-11.

Galleschi, L., Friggeri, M., Repiccioli, R., Come, D., Corbineau, F. 1993. Aspartic proteinase inhibitor from wheat: some properties, pp. 207-211 *Proceedings of Fourth Int. Workshop Seads*: *Basic and Applied Aspects of Seed Biology*.

Geerlof, A., Brown, J., Coutard, B., Egloff, M.P., Enguita, F.J., Fogg, M.J., Gilbert, R.J., Groves, M.R., Haouz, A., Nettleship, J.E., Nordlund, P., Owens, R.J., Ruff, M., Sainsbury, S., Svergun, D.I., Wilmanns, M., 2006. The impact of protein characterization in structural proteomics. *Acta Crystallographica. Section D, Structural Biology* 62, 1125-1136.

Gerstein, M., Lesk, A.M., Chothia, C., 1994. Structural mechanisms for domain movements in proteins. *Biochemistry* 33, 6739-6749.

Ghosh, E., Kumari, P., Jaiman, D., Shukla, A.K., 2015. Methodological advances: the unsung heroes of the GPCR structural revolution. *Nature Reviews. Molecular Cell Biology* 16, 69-81.

Gildea, R.J., Waterman, D.G., Parkhurst, J.M., Axford, D., Sutton, G., Stuart, D.I., Sauter, N.K., Evans, G., Winter, G., 2014. New methods for indexing multi-lattice diffraction data. *Acta Crystallographica. Section D, Structural Biology* 70, 2652-2666.

Gill, R., Kolstoe, S.E., Mohammed, F., Al, D.B.A., Mosely, J.E., Sarwar, M., Cooper, J.B., Wood, S.P., Shoolingin-Jordan, P.M., 2009. Structure of human porphobilinogen deaminase at 2.8 A: the molecular basis of acute intermittent porphyria. *The Biochemical Journal* 420, 17-25.

Gorrec, F., 2009. The MORPHEUS protein crystallization screen. *Journal of Applied Crystallography* 42, 1035-1042.

Gouet, P., Robert, X., Courcelle, E., 2003. *ESPript/ENDscript*: Extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Research* 31, 3320-3323.

Grakoui, A., McCourt, D., Wychowski, C., Feinstone, S., Rice, C., 1993. Characterization of the hepatitis C virus-encoded serine proteinase: determination of proteinase-dependent polyprotein cleavage sites. *Journal of Virology* 67, 2832-2843.

Green, K.Y., 2014. Norovirus infection in immunocompromised hosts. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases* 20, 717-723.

Greene, C.M., McElvaney, N.G., 2009. Proteases and antiproteases in chronic neutrophilic lung disease - relevance to drug discovery. *British Journal of Pharmacology* 158, 1048-1058.

Guo, J., Erskine, P., Coker, A.R., Wood, S.P., Cooper, J.B., 2017a. Structural studies of domain movement in active-site mutants of porphobilinogen deaminase from Bacillus megaterium. *Acta Crystallographica Section F, Structural Biology Communications* 73, 612-620.

Guo, J., Coker, A.R., Wood, S.P., Cooper, J.B., Chohan, S.M., Rashid, N., Akhtar, M., 2017b. Structure and function of the thermostable L-asparaginase from

Thermococcus kodakarensis. *Acta Crystallographica. Section D, Structural Biology* 73, 889-895.

Guo, J., Erskine, P.T., Coker, A.R., Wood, S.P., Cooper, J.B., Mares, M., Baudys, M., 2015. Structure of a Kunitz-type potato cathepsin D inhibitor. *Journal of Structural Biology* 192, 554-560.

Guo, J., Zhang, W., Coker, A.R., Wood, S.P., Cooper, J.B., Ahmad, S., Ali, S., Rashid, N., Akhtar, M., 2017c. Structure of the family B DNA polymerase from the hyperthermophilic archaeon *Pyrobaculum calidifontis*. *Acta Crystallographica. Section D, Structural Biology* 73.

Habib, H., Fazili, K.M., 2007. Plant protease inhibitors: a defense strategy in plants. *Biotechnology and Molecular Biology Reviews* 2, 68-85.

Hahn, T., 2002. Editor. International Tables for Crystallography, Vol. A. Space-Group Symmetry: Brief Teaching Edition.

Halarnkar, P., Jackson, G., Straub, K., Schooley, D., 1993. Juvenile hormone catabolism in *Manduca sexta*: Homologue selectivity of catabolism and identification of a diol-phosphate conjugate as a major end product. *Cellular and Molecular Life Sciences* 49, 988-994.

Hammock, B.D., Sparks, T.C., 1977. A rapid assay for insect juvenile hormone esterase activity. *Analytical Biochemistry* 82, 573-579.

Hamre, D., Bernstein, J., Donovick, R., 1950. Activity of *p*-aminobenzaldehyde, 3-thiosemicarbazone on vaccinia virus in the chick embryo and in the mouse. *Proceedings of the Society for Experimental Biology and Medicine* 73, 275-278.

Haney, P., Konisky, J., Koretke, K., Luthey-Schulten, Z., Wolynes, P., 1997. Structural basis for thermostability and identification of potential active site residues for adenylate kinases from the archaeal genus *Methanococcus*. *Proteins Structure Function and Genetics* 28, 117-130.

Hannapel, D.J., 1993. Nucleotide and deduced amino acid sequence of the 22-kilodalton cathepsin D inhibitor protein of potato (*Solanum tuberosum* L.). *Plant Physiology* 101, 703-704.

Harada, Y., Lifchitz, A., Berthou, J., Jolles, P., 1981. A translation function combining packing and diffraction information: an application to lysozyme (high-temperature form). *Acta Crystallographica Section A, Crystal Physics, Diffraction, Theoretical and General Crystallography* 37, 398-406.

Harms, E., Wehner, A., Aung, H.-P., Röhm, K., 1991. A catalytic role for threonine-12 of *E. coli* asparaginase II as established by site-directed mutagenesis. *FEBS Letters* 285, 55-58.

Hayward, S., Berendsen, H.J., 1998. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins* 30, 144-154.

Hayward, S., Lee, R.A., 2002. Improvements in the analysis of domain motions in proteins from conformational change: *DynDom* version 1.50. *Journal of Molecular Graphics & Modelling* 21, 181-183.

Herbers, K., Prat, S., Willmitzer, L., 1994. Cloning and characterization of a cathepsin D inhibitor gene from *Solanum tuberosum* L. *Plant Molecular Biology* 26, 73-83.

Hii, S.L., Tan, J.S., Ling, T.C., Ariff, A.B., 2012. Pullulanase: role in starch hydrolysis and potential industrial applications. *Enzyme Research* 2012, 921362.

Hirsch, C.D., Hamilton, J.P., Childs, K.L., Cepela, J., Crisovan, E., Vaillancourt, B., Hirsch, C.N., Habermann, M., Neal, B., Buell, C.R., 2014. Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. *The Plant Genome* 7.

Hondoh, H., Kuriki, T., Matsuura, Y., 2003. Three-dimensional structure and substrate binding of *Bacillus stearothermophilus* neopullulanase. *Journal of Molecular Biology* 326, 177-188.

Huber, R., 1965. Die automatisierte faltmolekülmethode. *Acta Crystallographica* 19, 353-356.

Huhti, L., Szakal, E.D., Puustinen, L., Salminen, M., Huhtala, H., Valve, O., Blazevic, V., Vesikari, T., 2011. Norovirus GII-4 causes a more severe gastroenteritis than other noroviruses in young children. *Journal of Infectious Diseases* 203, 1442-1444.

Hussey, R.J., Coates, L., Gill, R.S., Erskine, P.T., Coker, S.F., Mitchell, E., Cooper, J.B., Wood, S., Broadbridge, R., Clarke, I.N., Lambden, P.R., Shoolingin-Jordan, P.M., 2011. A structural study of norovirus 3C protease specificity: binding of a designed active site-directed peptide inhibitor. *Biochemistry* 50, 240-249.

Hüssy, P., Langen, H., Mous, J., Jacobsen, H., 1996. Hepatitis C virus core protein: carboxy-terminal boundaries of two processed species suggest cleavage by a signal peptide peptidase. *Virology* 224, 93-104.

Hutson, A.M., Atmar, R.L., Estes, M.K., 2004. Norovirus disease: changing epidemiology and host susceptibility factors. *Trends in Microbiology* 12, 279-287.

Hutson, A.M., Atmar, R.L., Graham, D.Y., Estes, M.K., 2002. Norwalk virus infection and disease is associated with ABO *histo*-blood group type. *Journal of Infectious Diseases* 185, 1335-1337.

Incardona, M.F., Bourenkov, G.P., Levik, K., Pieritz, R.A., Popov, A.N., Svensson, O., 2009. *EDNA*: a framework for plugin-based applications applied to X-ray experiment online data analysis. *Journal of Synchrotron Radiation* 16, 872-879.

Jaeger, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K., 2012. Global landscape of HIV-human protein complexes. *Nature* 481, 365-370.

Jaenicke, R., Böhm, G., 1998. The stability of proteins in extreme environments. *Current Opinion in Structural Biology* 8, 738-748.

Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Science of the USA* 96, 3801-3806.

Jancarik, J., Kim, S.-H., 1991. Sparse matrix sampling: a screening method for crystallization of proteins. *Journal of Applied Crystallography* 24, 409-411.

Janeček, Š., Svensson, B., MacGregor, E.A., 2014. *α*-Amylase: an enzyme specificity found in various families of glycoside hydrolases. *Cellular and Molecular Life Sciences* 71, 1149-1170.

Jashni, M.K., Mehrabi, R., Collemare, J., Mesarich, C.H., de Wit, P.J., 2015. The battle in the apoplast: further insights into the roles of proteases and their inhibitors in plant-pathogen interactions. *Front Plant Science* 6, 584.

Jensen, B., Norman, B., 1984. Bacillus acidopullulyticus Pullulanase: application and regulatory aspects for use in the food industry [Promozyme]. *Process Biochemistry*.

Jiang, X., Wang, M., Wang, K., Estes, M.K., 1993. Sequence and genomic organization of Norwalk virus. *Virology* 195, 51-61.

Johnson, P.C., Mathewson, J.J., DuPont, H.L., Greenberg, H.B., 1990. Multiple-challenge study of host susceptibility to Norwalk gastroenteritis in US adults. *Journal of Infectious Diseases* 161, 18-21.

Joosten, R.P., Long, F., Murshudov, G.N., Perrakis, A., 2014. The *PDB_REDO* server for macromolecular structure model optimization. *IUCrJ* 1, 213-220.

Jordan, P.M., Warren, M.J., 1987. Evidence for a dipyrromethane cofactor at the catalytic site of *E. coli* porphobilinogen deaminase. *FEBS Letters* 225, 87-92.

Jordan, P.M., Woodcock, S.C., 1991. Mutagenesis of arginine residues in the catalytic cleft of *Escherichia coli* porphobilinogen deaminase that affects dipyrromethane cofactor assembly and tetrapyrrole chain initiation and elongation. *The Biochemical Journal* 280 (Pt 2), 445-449.

Jørgensen, M., Stensballe, A., Welinder, K.G., 2011. Extensive post-translational processing of potato tuber storage proteins and vacuolar targeting. *FEBS J* 278, 4070-4087.

Jung, K., Wang, Q., Kim, Y., Scheuer, K., Zhang, Z., Shen, Q., Chang, K.-O., Saif, L.J., 2012. The effects of simvastatin or interferon-*α* on infectivity of human norovirus using a gnotobiotic pig model for the study of antivirals. *PloS One* 7, e41619.

Kabsch, W., 1988. Automatic indexing of rotation diffraction patterns. *Journal of Applied Crystallography* 21, 67-72.

Kabsch, W., 2010. *Xds*. *Acta Crystallographica. Section D, Structural Biology* 66, 125-132.

Kankanamalage, A.C.G., Kim, Y., Weerawarna, P.M., Uy, R.A.Z., Damalanka, V.C., Mandadapu, S.R., Alliston, K.R., Mehzabeen, N., Battaile, K.P., Lovell, S., 2015. Structure-guided design and optimization of dipeptidyl inhibitors of norovirus 3CL protease. Structure-activity relationships and biochemical, X-ray crystallographic, cell-based, and in vivo studies. *Journal of Medicinal Chemistry* 58, 3144.

Kannangara, C.G., Andersen, R.V., Pontoppidan, B., Willows, R., von Wettstein, D., 1994. Enzymic and mechanistic studies on the conversion of glutamate to 5-aminolaevulinate. *Ciba Foundation Symposium* 180, 3-20; discussion 21-25.

Kannangara, C.G., Gough, S.P., Bruyant, P., Hoober, J.K., Kahn, A., von Wettstein, D., 1988. tRNA(Glu) as a cofactor in delta-aminolevulinate biosynthesis: steps that regulate chlorophyll synthesis. *Trends in Biochemical Sciences* 13, 139-143.

Kantardjieff, K.A., Rupp, B., 2003. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Science : a publication of the Protein Society* 12, 1865-1871.

Kapikian, A.Z., Wyatt, R.G., Dolin, R., Thornhill, T.S., Kalica, A.R., Chanock, R.M., 1972. Visualization by immune electron microscopy of a 27-nm particle associated with acute infectious nonbacterial gastroenteritis. *Journal of Virology* 10, 1075-1081.

Karle, J.t., Hauptman, H., 1956. A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22. *Acta Crystallographica* 9, 635-651.

Karplus, P.A., Diederichs, K., 2012. Linking crystallographic model and data quality. *Science* 336, 1030-1033.

Kawatkar, S., Gagnon, M., Hoesch, V., Tiong-Yip, C., Johnson, K., Ek, M., Nilsson, E., Lister, T., Olsson, L., Patel, J., 2016. Design and structure–activity relationships of novel inhibitors of human rhinovirus 3C protease. *Bioorganic & Medicinal Chemistry Letters* 26, 3248-3252.

Keegan, R., Lebedev, A., Erskine, P., Guo, J., Wood, S.P., Hopper, D.J., Rigby, S.E., Cooper, J.B., 2014. Structure of the 2,4'-dihydroxyacetophenone dioxygenase from Alcaligenes sp. 4HAP. *Acta Crystallographica. Section D, Structural Biology* 70, 2444-2454.

Keegan, R., Waterman, D.G., Hopper, D.J., Coates, L., Taylor, G., Guo, J., Coker, A.R., Erskine, P.T., Erskine, S., Cooper, J.B., 2016. The 1.1 Å resolution structure of a periplasmic phosphate-binding protein from *Stenotrophomonas maltophilia*: a crystallization contaminant identified by molecular replacement using the entire Protein Data Bank. *Acta Crystallographica. Section D, Structural Biology* 72.

Keegan, R.M., Winn, M.D., 2008. *MrBUMP*: an automated pipeline for molecular replacement. *Acta Crystallographica. Section D, Structural Biology* 64, 119-124.

Keilova, H., Tomášek, V., 1976a. Further characteristics of cathepsin D inhibitor from potatoes. *Collection of Czechoslovak Chemical Communications* 41, 2440-2447.

Keilova, H., Tomášek, V., 1976b. Isolation and some properties of cathepsin D inhibitor from potatoes. *Collection of Czechoslovak Chemical Communications* 41, 489-497.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R., Wyckoff, H., Phillips, D.C., 1958. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* 181, 662-666.

Khamrui, S., Dasgupta, J., Dattagupta, J.K., Sen, U., 2005. Single mutation at P1 of a chymotrypsin inhibitor changes it to a trypsin inhibitor: X-ray structural (2.15 Å) and biochemical basis. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1752, 65-72.

Kidd, J.G., 1953. Regression of transplanted lymphomas induced in vivo by means of normal guinea pig serum. I. Course of transplanted cancers of various kinds in mice and rats given guinea pig serum, horse serum, or rabbit serum. *The Journal of Experimental Medicine* 98, 565-582.

Killelea, T., Ghosh, S., Tan, S.S., Heslop, P., Firbank, S.J., Kool, E.T., Connolly, B.A., 2010. Probing the interaction of archaeal DNA polymerases with deaminated bases using X-ray crystallography and non-hydrogen bonding isosteric base analogues. *Biochemistry* 49, 5772-5781.

Kim, C.-H., Kim, D.-S., Taniguchi, H., Maruyama, Y., 1990. Purification of a amylase -pullulanase bifunctional enzyme by high-performance size-exclusion and hydrophobic-interaction chromatography. *Journal of Chromatography A* 512, 131-137.

Kim, J.J., Culley, C.M., Mohammad, R.A., 2012a. Telaprevir: An oral protease inhibitor for hepatitis C virus infection. *American Journal of Health-System Pharmacy* 69.

Kim, J.S., Cha, S.S., Kim, H.J., Kim, T.J., Ha, N.C., Oh, S.T., Cho, H.S., Cho, M.J., Kim, M.J., Lee, H.S., Kim, J.W., Choi, K.Y., Park, K.H., Oh, B.H., 1999. Crystal structure of a maltogenic amylase provides insights into a catalytic versatility. *The Journal of Biological Chemistry* 274, 26279-26286.

Kim, Y., Lovell, S., Tiew, K.-C., Mandadapu, S.R., Alliston, K.R., Battaile, K.P., Groutas, W.C., Chang, K.-O., 2012b. Broad-spectrum antivirals against 3C or 3C-like proteases of picornaviruses, noroviruses, and coronaviruses. *Journal of Virology* 86, 11754-11762.

Kiriyama, Y., Kubota, M., Takimoto, T., Kitoh, T., Tanizawa, A., Akiyama, Y., Mikawa, H., 1989. Biochemical characterization of U937 cells resistant to L-asparaginase: the role of asparagine synthetase. *Leukemia* 3, 294-297.

Kitagawa, M., Hayakawa, T., 2007. Antiproteases in the treatment of acute pancreatitis. *JOP : Journal of the Pancreas* 8, 518-525.

Knight, M.J., Ruaux, A., Mikolajek, H., Erskine, P.T., Gill, R., Wood, S.P., Wood, M., Cooper, J.B., 2006. Crystallization and preliminary X-ray diffraction analysis of BipD, a virulence factor from *Burkholderia pseudomallei*. *Acta Crystallographica. Section F, Structural Biology Communications* 62, 761-764.

Kohl, N.E., Emini, E.A., Schleif, W.A., Davis, L.J., Heimbach, J.C., Dixon, R., Scolnick, E.M., Sigal, I.S., 1988. Active human immunodeficiency virus protease is required for viral infectivity. *Proceedings of the National Academy of Sciences* 85, 4686-4690.

Koonin, E.V., Mushegian, A.R., Galperin, M.Y., Walker, D.R., 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Molecular Microbiology* 25, 619-637.

Koopman, J.S., Eckert, E.A., Greenberg, H.B., Strohm, B.C., Isaacson, R.E., Monto, A.S., 1982. Norwalk virus enteric illness acquired by swimming exposure. *American Journal of Epidemiology* 115, 173-177.

Kornfeld, R., Kornfeld, S., 1985. Assembly of asparagine-linked oligosaccharides. *Annual Review of Biochemistry* 54, 631-664.

Koshland, D.E., 1953. Stereochemistry and the mechanism of enzymatic reactions. *Biological Reviews* 28, 416-436.

Krauchenco, S., Pando, S.C., Marangoni, S., Polikarpov, I., 2003. Crystal structure of the Kunitz (STI)-type inhibitor from *Delonix regia* seeds. *Biochemical and Biophysical Research Communications* 312, 1303-1308.

Krissinel, E., Henrick, K., 2007. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* 372, 774-797.

Krojer, T., Talon, R., Pearce, N., Collins, P., Douangamath, A., Brandao-Neto, J., Dias, A., Marsden, B., von Delft, F., 2017. The *XChemExplorer* graphical workflow tool for routine or large-scale protein-ligand structure determination. *Acta Crystallographica. Section D, Structural Biology* 73, 267-278.

Kumar, K., Kataria, M., Verma, N., 2013. Plant asparaginase-based asparagine biosensor for leukemia. *Artificial Cells, Nanomedicine, and Biotechnology* 41, 184-188.

Kumar, S., Tsai, C.-J., Nussinov, R., 2000a. Factors enhancing protein thermostability. *Protein Engineering* 13, 179-191.

Kumar, S., Ma, B., Tsai, C.J., Nussinov, R., 2000b. Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers. *Proteins: Structure, Function, and Bioinformatics* 38, 368-383.

Kuriki, T., Okada, S., Imanaka, T., 1988. New type of pullulanase from *Bacillus stearothermophilus* and molecular cloning and expression of the gene in *Bacillus subtilis*. *Journal of Bacteriology* 170, 1554-1559.

Kuroita, T., Matsumura, H., Yokota, N., Kitabayashi, M., Hashimoto, H., Inoue, T., Imanaka, T., Kai, Y., 2005. Structural mechanism for coordination of proofreading and polymerase activities in archaeal DNA polymerases. *Journal of Molecular Biology* 351, 291-298.

Lambden, P.R., Caul, E.O., Ashley, C.R., Clarke, I.N., 1993. Sequence and genome organization of a human small round-structured (Norwalk-like) virus. *Science* 259, 516-519.

Lambert, R., Brownlie, P.D., Woodcock, S.C., Louie, G.V., Cooper, J.C., Warren, M.J., Jordan, P.M., Blundell, T.L., Wood, S.P., 1994. Structural studies on porphobilinogen deaminase. *Ciba Foundation Symposium* 180, 97-104; discussion 105-110.

Lander, M., Pitt, A.R., Alefounder, P.R., Bardy, D., Abell, C., Battersby, A.R., 1991. Studies on the mechanism of hydroxymethylbilane synthase concerning the role of arginine residues in substrate binding. *The Biochemical Journal* 275 (Pt 2), 447-452.

Laskowski, M., Kato, I., Leary, T., Schrode, J., Sealock, R., 1974. Evolution of specificity of protein proteinase inhibitors, p. 597-611, Proteinase Inhibitors, *Springer*.

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

Lathrop, J.T., Timko, M.P., 1993. Regulation by heme of mitochondrial protein transport through a conserved amino acid motif. *Science* 259, 522-525.

Le Guyader, F.S., Atmar, R.L., Le Pendu, J., 2012. Transmission of viruses through shellfish: when specific ligands come into play. *Current Opinion in Virology* 2, 103-110.

Le Pendu, J., Ruvoën-Clouet, N., Kindberg, E., Svensson, L. 2006. Mendelian resistance to human norovirus infections, pp. 375-386 Seminars in immunology, Vol. 18. *Elsevier*.

Leathers, T.D., 2003. Biotechnological production and applications of pullulan. *Applied Microbiology and Biotechnology* 62, 468-473.

Lee, A.Y., Gulnik, S.V., Erickson, J.W., 1998. Conformational switching in an aspartic proteinase. *Nature Structural Biology* 5, 866-871.

Lee, H.S., Kim, M.S., Cho, H.S., Kim, J.I., Kim, T.J., Choi, J.H., Park, C., Lee, H.S., Oh, B.H., Park, K.H., 2002. Cyclomaltodextrinase, neopullulanase, and maltogenic amylase are nearly indistinguishable from each other. *The Journal of Biological Chemistry* 277, 21891-21897.

Leen, E.N., Baeza, G., Curry, S., 2012. Structure of a murine norovirus NS6 protease-product complex revealed by adventitious crystallisation. *PloS One* 7, e38723.

Leeper, F.J., 1985. The biosynthesis of porphyrins, chlorophylls, and vitamin B12. *Natural Product Reports* 2, 561-580.

Leong, L., Walker, P., Porter, A., 1993a. Human rhinovirus-14 protease 3C (3Cpro) binds specifically to the 5'-noncoding region of the viral RNA. Evidence that 3Cpro has different domains for the RNA binding and proteolytic activities. *Journal of Biological Chemistry* 268, 25735-25739.

Leong, L.E., Walker, P.A., Porter, A.G., 1993b. Human rhinovirus-14 protease 3C (3Cpro) binds specifically to the 5'-noncoding region of the viral RNA. Evidence that 3Cpro has different domains for the RNA binding and proteolytic activities. *The Journal of Biological Chemistry* 268, 25735-25739.

Leslie, A.G., 2006. The integration of macromolecular diffraction data. *Acta Crystallographica. Section D, Structural Biology* 62, 48-57.

Lew, J.F., Kapikian, A.Z., Valdesuso, J., Green, K.Y., 1994a. Molecular characterization of Hawaii virus and other Norwalk-like viruses: evidence for genetic polymorphism among human caliciviruses. *Journal of Infectious Diseases* 170, 535-542.

Lew, J.F., Kapikian, A.Z., Jiang, X., Estes, M.K., Green, K.Y., 1994b. Molecular characterization and expression of the capsid protein of a Norwalk-like virus recovered from a Desert Shield troop with gastroenteritis. *Virology* 200, 319-325.

Li, J.M., Brathwaite, O., Cosloy, S.D., Russell, C.S., 1989. 5-Aminolevulinic acid synthesis in *Escherichia coli*. *Journal of Bacteriology* 171, 2547-2552.

Li, M., Phylip, L.H., Lees, W.E., Winther, J.R., Dunn, B.M., Wlodawer, A., Kay, J., Gustchina, A., 2000a. The aspartic proteinase from *Saccharomyces cerevisiae* folds its own inhibitor into a helix. *Nature Structural Biology* 7, 113-117.

Li, M., Phylip, L.H., Lees, W.E., Winther, J.R., Dunn, B.M., Wlodawer, A., Kay, J., Gustchina, A., 2000b. The aspartic proteinase from *Saccharomyces cerevisiae* folds its own inhibitor into a helix. *Nature Structural & Molecular Biology* 7, 113-117.

Li, S., Zhang, Q.-R., Xu, W.-H., Schooley, D.A., 2005. Juvenile hormone diol kinase, a calcium-binding protein with kinase activity, from the silkworm, *Bombyx mori*. *Insect Biochemistry and Molecular Biology* 35, 1235-1248.

Li, W., Wang, B.-E., Moran, P., Lipari, T., Ganesan, R., Corpuz, R., Ludlam, M.J., Gogineni, A., Koeppen, H., Bunting, S., 2009. Pegylated kunitz domain inhibitor suppresses hepsin-mediated invasive tumor growth and metastasis. *Cancer Research* 69, 8395-8402.

Lindesmith, L., Moe, C., Marionneau, S., Ruvoen, N., Jiang, X., Lindblad, L., Stewart, P., LePendu, J., Baric, R., 2003. Human susceptibility and resistance to Norwalk virus infection. *Nature Medicine* 9, 548-553.

Liu, L., Johnson, H.L., Cousens, S., Perin, J., Scott, S., Lawn, J., Rudan, I., Campbell, H., Cibulskis, R., Li, M., 2012. Child Health Epidemiology Reference Group of WHO and UNICEF Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* 379, 2151-2161.

Lochridge, V.P., Hardy, M.E., 2003. Snow Mountain virus genome sequence and virus-like particle assembly. *Virus Genes* 26, 71-82.

Long, F., Vagin, A.A., Young, P., Murshudov, G.N., 2008. *BALBES*: a molecular-replacement pipeline. *Acta Crystallographica. Section D, Structural Biology* 64, 125-132.

Longley, M.J., Clark, S., Man, C.Y.W., Hudson, G., Durham, S.E., Taylor, R.W., Nightingale, S., Turnbull, D.M., Copeland, W.C., Chinnery, P.F., 2006. Mutant POLG2 disrupts DNA polymerase γ subunits and causes progressive external ophthalmoplegia. *The American Journal of Human Genetics* 78, 1026-1034.

Louie, G.V., 1993. Porphobilinogen deaminase and its structural similarity to the bidomain binding proteins. *Current Opinion in Structural Biology* 3, 401-408.

Louie, G.V., Brownlie, P.D., Lambert, R., Cooper, J.B., Blundell, T.L., Wood, S.P., Warren, M.J., Woodcock, S.C., Jordan, P.M., 1992. Structure of porphobilinogen deaminase reveals a flexible multidomain polymerase with a single catalytic site. *Nature* 359, 33-39.

Louie, G.V., Brownlie, P.D., Lambert, R., Cooper, J.B., Blundell, T.L., Wood, S.P., Malashkevich, V.N., Hadener, A., Warren, M.J., Shoolingin-Jordan, P.M., 1996. The three-dimensional structure of *Escherichia coli* porphobilinogen deaminase at 1.76-A resolution. *Proteins* 25, 48-78.

MacGregor, E.A., Janeček, Š., Svensson, B., 2001. Relationship of sequence and structure to specificity in the *α*-amylase family of enzymes. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* 1546, 1-20.

Machida, Y., Fukui, F., Komoto, T., 1986. Use of oligosaccharides for promoting the proliferation of bidifidobacteria. *European Patent* 242459.

Maillard, J.-Y., 2001. Virus susceptibility to biocides: an understanding. *Reviews in Medical Microbiology* 12, 63-74.

Mandadapu, S.R., Weerawarna, P.M., Gunnam, M.R., Alliston, K.R., Lushington, G.H., Kim, Y., Chang, K.-O., Groutas, W.C., 2012. Potent inhibition of norovirus 3CL protease by peptidyl *α*-ketoamides and *α*-ketoheterocycles. *Bioorganic & Medicinal Chemistry Letters* 22, 4820-4826.

Mandadapu, S.R., Gunnam, M.R., Kankanamalage, A.C.G., Uy, R.A.Z., Alliston, K.R., Lushington, G.H., Kim, Y., Chang, K.-O., Groutas, W.C., 2013a. Potent inhibition of norovirus by dipeptidyl *α*-hydroxyphosphonate transition state mimics. *Bioorganic & Medicinal Chemistry Letters* 23, 5941-5944.

Mandadapu, S.R., Weerawarna, P.M., Prior, A.M., Uy, R.A.Z., Aravapalli, S., Alliston, K.R., Lushington, G.H., Kim, Y., Hua, D.H., Chang, K.-O., 2013b. Macrocyclic inhibitors of 3C and 3C-like proteases of picornavirus, norovirus, and coronavirus. *Bioorganic & Medicinal Chemistry Letters* 23, 3709-3712.

Mares, M., Meloun, B., Pavlik, M., Kostka, V., Baudys, M., 1989. Primary structure of cathepsin D inhibitor from potatoes and its structure relationship to soybean trypsin inhibitor family. *FEBS Letters* 251, 94-98.

Marionneau, S., Cailleau-Thomas, A., Rocher, J., Le Moullac-Vaidye, B., Ruvoën, N., Clément, M., Le Pendu, J., 2001. ABH and Lewis *histo*-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world. *Biochimie* 83, 565-573.

Marsault, E., Peterson, M.L., 2011. Macrocycles are great cycles: applications, opportunities, and challenges of synthetic macrocycles in drug discovery. *Journal of Medicinal Chemistry* 54, 1961-2004.

Martin, J.A., 1992. Recent advances in the design of HIV proteinase inhibitors. *Antiviral Research* 17, 265-278.

Masutani, C., Kusumoto, R., Yamada, A., Dohmae, N., Yokoi, M., Yuasa, M., Araki, M., Iwai, S., Takio, K., Hanaoka, F., 1999. The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase eta. *Nature* 399, 700-704.

Matthews, B.W., 1968. Solvent content of protein crystals. *Journal of Molecular Biology* 33, 491-497.

Matthews, D., Dragovich, P., Webber, S., Fuhrman, S., Patick, A., Zalman, L., Hendrickson, T., Love, R., Prins, T., Marakovits, J., 1999. Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C

protease with potent antiviral activity against multiple rhinovirus serotypes. *Proceedings of the National Academy of Sciences* 96, 11000-11007.

Mauzerall, D., Granick, S., 1956. The occurrence and determination of delta-amino-levulinic acid and porphobilinogen in urine. *The Journal of Biological Chemistry* 219, 435-446.

Maxwell, R.A., Welch, W.H., Schooley, D.A., 2002a. Juvenile hormone diol kinase I. Purification, characterization, and substrate specificity of juvenile hormone-selective diol kinase from *Manduca sexta*. *Journal of Biological Chemistry* 277, 21874-21881.

Maxwell, R.A., Welch, W.H., Horodyski, F.M., Schegg, K.M., Schooley, D.A., 2002b. Juvenile hormone diol kinase II. Sequencing, cloning, and molecular modeling of juvenile hormone-selective diol kinase from *Manduca sexta*. *Journal of Biological Chemistry* 277, 21882-21890.

McCoy, A.J., Storoni, L.C., Read, R.J., 2004a. Simple algorithm for a maximum-likelihood SAD function. *Acta Crystallographica. Section D, Biological Crystallography* 60, 1220-1228.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., Read, R.J., 2007. Phaser crystallographic software. *Journal of Applied Crystallography* 40, 658-674.

McFadden, N., Bailey, D., Carrara, G., Benson, A., Chaudhry, Y., Shortland, A., Heeney, J., Yarovinsky, F., Simmonds, P., Macdonald, A., Goodfellow, I., 2011.

Norovirus regulation of the innate immune response and apoptosis occurs via the product of the alternative open reading frame 4. *PLoS Pathogens* 7, e1002413.

McPherson, A., Gavira, J.A., 2014. Introduction to protein crystallization. *Acta crystallographica. Section F, Structural Biology Communications* 70, 2-20.

McPherson, A., Jr., 1976. Crystallization of proteins from polyethylene glycol. *The Journal of Biological Chemistry* 251, 6300-6303.

McRee, D.E., 1999. Practical protein crystallography *Academic Press*.

Meeroff, J.C., Schreiber, D.S., Trier, J.S., Blacklow, N.R., 1980. Abnormal gastric motor function in viral gastroenteritis. *Annals of Internal Medicine* 92, 370-373.

Messina, J.P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G.S., Pybus, O.G., Barnes, E., 2015. Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology* 61, 77-87.

Meulenbroek, E.M., Thomassen, E.A., Pouvreau, L., Abrahams, J.P., Gruppen, H., Pannu, N.S., 2012. Structure of a post-translationally processed heterodimeric double-headed Kunitz-type serine protease inhibitor from potato. *Acta crystallographica. Section D, Structural Biology* 68, 794-799.

Michalska, K., Tan, K., Chang, C., Li, H., Hatzos-Skintges, C., Molitsky, M., Alkire, R., Joachimiak, A., 2015. *In situ* X-ray data collection and structure phasing of protein crystals at Structural Biology Center 19-ID. *Journal of Synchrotron Radiation* 22, 1386-1395.

Migliolo, L., de Oliveira, A.S., Santos, E.A., Franco, O.L., Maurício, P., 2010. Structural and mechanistic insights into a novel non-competitive Kunitz trypsin inhibitor from *Adenanthera pavonina* L. seeds with double activity toward serine- and cysteine-proteinases. *Journal of Molecular Graphics and Modelling* 29, 148-156.

Moore, M.R., McColl, K.E., 1989. Therapy of the acute porphyrias. *Clinical Biochemistry* 22, 181-188.

Möricke, A., Reiter, A., Zimmermann, M., Gadner, H., Stanulla, M., Dördelmann, M., Löning, L., Beier, R., Ludwig, W.-D., Ratei, R., 2008. Risk-adjusted therapy of acute lymphoblastic leukemia can decrease treatment burden and improve survival: treatment results of 2169 unselected pediatric and adolescent patients enrolled in the trial ALL-BFM 95. *Blood* 111, 4477-4489.

Mosimann, S.C., Cherney, M.M., Sia, S., Plotch, S., James, M.N., 1997. Refined X-ray crystallographic structure of the poliovirus 3C gene product. *Journal of Molecular Biology* 273, 1032-1047.

Mueller-Dieckmann, C., Panjikar, S., Schmidt, A., Mueller, S., Kuper, J., Geerlof, A., Wilmanns, M., Singh, R.K., Tucker, P.A., Weiss, M.S., 2007. On the routine use of soft X-rays in macromolecular crystallography. Part IV. Efficient determination of anomalous substructures in biomacromolecules using longer X-ray wavelengths. *Acta Crystallographica. Section D, Structural Biology* 63, 366-380.

Murshudov, G.N., Vagin, A.A., Dodson, E.J., 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica. Section D, Structural Biology* 53, 240-255.

Murshudov, G.N., Skubak, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F., Vagin, A.A., 2011. *REFMAC*5 for the refinement of macromolecular crystal structures. *Acta crystallographica. Section D, Structural Biology* 67, 355-367.

Nakamura, K., Someya, Y., Kumasaka, T., Ueno, G., Yamamoto, M., Sato, T., Takeda, N., Miyamura, T., Tanaka, N., 2005. A norovirus protease structure provides insights into active and substrate binding site integrity. *Journal of Virology* 79, 13685-13693.

Nakayama, S., Kretsinger, R.H., 1994. Evolution of the EF-hand family of proteins. *Annual Review of Biophysics and Biomolecular Structure* 23, 473-507.

Navaza, J., 1994. *AMoRe*: an automated package for molecular replacement. *Acta Crystallographica. Section A, Foundations of Crystallography* 50, 157-163.

Navaza, J., 2001. Implementation of molecular replacement in *AMoRe*. *Acta Crystallographica. Section D, Structural Biology* 57, 1367-1372.

Nayak, A., Goodfellow, I.G., Woolaway, K.E., Birtley, J., Curry, S., Belsham, G.J., 2006. Role of RNA structure and RNA binding activity of foot-and-mouth disease virus 3C protein in VPg uridylylation and virus replication. *Journal of Virology* 80, 9865-9875.

Ng, J.T., Dekker, C., Kroemer, M., Osborne, M., von Delft, F., 2014. Using textons to rank crystallization droplets by the likely presence of crystals. *Acta Crystallographica. Section D,Structural Biology* 70, 2702-2718.

Nguyen, H.A., Su, Y., Lavie, A., 2016. Structural insight into substrate selectivity of *Erwinia chrysanthemi* L-asparaginase. *Biochemistry* 55, 1246-1253.

Niehaus, F., Peters, A., Groudieva, T., Antranikian, G., 2000. Cloning, expression and biochemical characterisation of a unique thermostable pullulan-hydrolysing enzyme from the hyperthermophilic archaeon *Thermococcus aggregans*. *FEMS Microbiology Letters* 190, 223-229.

Nisha, M., Satyanarayana, T., 2013. Recombinant bacterial amylopullulanases: developments and perspectives. *Bioengineered* 4, 388-400.

Nisha, M., Satyanarayana, T., 2016. Characteristics, protein engineering and applications of microbial thermostable pullulanases and pullulan hydrolases. *Applied Microbiology and Biotechnology* 100, 5661-5679.

Ohtaki, A., Mizuno, M., Yoshida, H., Tonozuka, T., Sakano, Y., Kamitori, S., 2006. Structure of a complex of *Thermoactinomyces vulgaris* R-47 alpha-amylase 2 with maltohexaose demonstrates the important role of aromatic residues at the reducing end of the substrate binding cleft. *Carbohydrate Research* 341, 1041-1046.

Nisha, M., Satyanarayana, T., 2017. Characteristics and applications of recombinant thermostable amylopullulanase of Geobacillus thermoleovorans

secreted by Pichia pastoris. *Applied Microbiology and Biotechnology* 101, 2357-2369.

Oliva, M.L., Silva, M.C., Sallai, R.C., Brito, M.V., Sampaio, M.U., 2010. A novel subclassification for Kunitz proteinase inhibitors from leguminous seeds. *Biochimie* 92, 1667-1673.

Onesti, S., Brick, P., Blow, D.M., 1991. Crystal structure of a Kunitz-type trypsin inhibitor from *Erythrina caffra* seeds. *Journal of Molecular Biology* 217, 153-176.

Park, H., Rangarajan, E.S., Sygusch, J., Izard, T., 2010. Dramatic improvement of crystal quality for low-temperature-grown rabbit muscle aldolase. *Acta Crystallographica. Section F, Structural Biology Communications* 66, 595-600.

Park, Y., Choi, B.H., Kwak, J.S., Kang, C.W., Lim, H.T., Cheong, H.S., Hahm, K.S., 2005. Kunitz-type serine protease inhibitor from potato (*Solanum tuberosum* L. cv. Jopung). *Journal of Agricultural and Food Chemistry* 53, 6491-6496.

Parrino, T.A., Schreiber, D.S., Trier, J.S., Kapikian, A.Z., Blacklow, N.R., 1977. Clinical immunity in acute gastroenteritis caused by Norwalk agent. *The New England Journal of Medicine* 297, 86-89.

Pearce, N., Bradley, A.R., Collins, P., Krojer, T., Nowak, R., Talon, R., Marsden, B.D., Kelm, S., Shi, J., Deane, C., 2016. A Multi-Crystal Method for Extracting Obscured Signal from Crystallographic Electron Density. *bioRxiv*, 073411.

Phillips, G., Tam, C.C., Conti, S., Rodrigues, L.C., Brown, D., Iturriza-Gomara, M., Gray, J., Lopman, B., 2010. Community incidence of norovirus-associated infectious intestinal disease in England: improved estimates using viral load for norovirus diagnosis. *American Journal of Epidemiology*, kwq021.

Plant, A.R., Morgan, H.W., Daniel, R.M., 1986. A highly stable pullulanase from *Thermus aquaticus* YT-1. *Enzyme and Microbial Technology* 8, 668-672.

Pluscec, J., Bogorad, L., 1970. A dipyrrylmethane intermediate in the enzymatic synthesis of uroporphyrinogen. *Biochemistry* 9, 4736-4743.

Poliakoff, M., Licence, P., 2007. Sustainable technology: green chemistry. *Nature* 450, 810-812.

Ponting, C.P., Russell, R.B., 2000. Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all *β*-trefoil proteins. *Journal of Molecular Biology* 302, 1041-1047.

Pouvreau, L., Gruppen, H., Piersma, S.R., van den Broek, L.A., van Koningsveld, G.A., Voragen, A.G., 2001. Relative abundance and inhibitory distribution of protease inhibitors in potato juice from cv. *Elkana*. *Journal of Agricultural and Food Chemistry* 49, 2864-2874.

Powell, H.R., Johnson, O., Leslie, A.G., 2013. Autoindexing diffraction images with *iMosflm*. *Acta Crystallographica. Section D, Structural Biology* 69, 1195-1203.

Prasad, B.V., Matson, D.O., Smith, A.W., 1994. Three-dimensional structure of calicivirus. *Journal of Molecular Biology* 240, 256-264.

Prasad, B.V., Hardy, M.E., Dokland, T., Bella, J., Rossmann, M.G., Estes, M.K., 1999. X-ray crystallographic structure of the Norwalk virus capsid. *Science* 286, 287-290.

Pui, C.-H., Campana, D., Pei, D., Bowman, W.P., Sandlund, J.T., Kaste, S.C., Ribeiro, R.C., Rubnitz, J.E., Raimondi, S.C., Onciu, M., 2009. Treating childhood acute lymphoblastic leukemia without cranial irradiation. *New England Journal of Medicine* 360, 2730-2741.

Rakashanda, S., Ishaq, M., Masood, A., Amin, S., 2012. Antibacterial activity of a trypsin-chymotrypsin-elastase inhibitor isolated from *Lavatera cashmeriana camb*. seeds. *The Journal of Animal & Plant Sciences* 22, 983-986.

Ravichandran, S., Dasgupta, J., Chakrabarti, C., Ghosh, S., Singh, M., Dattagupta, J., 2001. The role of Asn14 in the stability and conformation of the reactive-site loop of winged bean chymotrypsin inhibitor: crystal structures of two point mutants Asn14→ Lys and Asn14→ Asp. *Protein engineering* 14, 349-357.

Read, R., 1990. Structure-factor probabilities for related structures. *Acta Crystallographica. Section A, Foundations of Crystallography* 46, 900-912.

Rendleman, J.A., 1997. Enhancement of cyclodextrin production through use of debranching enzymes. *Biotechnology and Applied Biochemistry* 26, 51-61.

Rhodes, G., 2010. Crystallography made crystal clear: a guide for users of macromolecular models *Academic Press*.

Richardson, M., 1977. The proteinase inhibitors of plants and micro-organisms. *Phytochemistry* 16, 159-169.

Richter, C., Tanaka, T., Koseki, T., Yada, R.Y., 1999. Contribution of a prosegment lysine residue to the function and structure of porcine pepsinogen A and its active form pepsin A. *European Journal of Biochemistry* 261, 746-752.

Riddiford, L.M., Hiruma, K., Zhou, X., Nelson, C.A., 2003. Insights into the molecular basis of the hormonal control of molting and metamorphosis from *Manduca sexta* and *Drosophila melanogaster*. *Insect Biochemistry and Molecular Biology* 33, 1327-1338.

Ritonja, A., Križaj, I., Meško, P., Kopitar, M., Lučovnik, P., Štrukelj, B., Pungerčar, J., Buttle, D.J., Barrett, A.J., Turk, V., 1990. The amino acid sequence of a novel inhibitor of cathepsin D from potato. *FEBS Letters* 267, 13-15.

Robert, X., Gouet, P., 2014. Deciphering key features in protein structures with the new *ENDscript* server. *Nucleic Acids Research* 42, W320-324.

Roberts, A., Gill, R., Hussey, R.J., Mikolajek, H., Erskine, P.T., Cooper, J.B., Wood, S.P., Chrystal, E.J., Shoolingin-Jordan, P.M., 2013. Insights into the mechanism of pyrrole polymerization catalysed by porphobilinogen deaminase: high-resolution X-ray studies of the *Arabidopsis thaliana* enzyme. *Acta Crystallographica. Section D, Structural Biology* 69, 471-485.

Rodgers, K.R., 1999. Heme-based sensors in biological systems. *Current Opinion in Chemical Biology* 3, 158-167.

Rodriguez, A.C., Park, H.-W., Mao, C., Beese, L.S., 2000. Crystal structure of a pol *α* family DNA polymerase from the hyperthermophilic archaeon *Thermococcus* sp. 9 N-7. *Journal of Molecular Biology* 299, 447-462.

Rodriguez, D.D., Grosse, C., Himmel, S., Gonzalez, C., de Ilarduya, I.M., Becker, S., Sheldrick, G.M., Uson, I., 2009. Crystallographic ab initio protein structure solution below atomic resolution. *Nature methods* 6, 651-653.

Rossmann, M.G., Blow, D.M., 1962. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica* 15, 24-31.

Russell, R.C., Yuan, H.-X., Guan, K.-L., 2014. Autophagy regulation by nutrient signaling. *Cell Research* 24, 42-57.

Russell, R.J., Ferguson, J.M., Hough, D.W., Danson, M.J., Taylor, G.L., 1997. The crystal structure of citrate synthase from the hyperthermophilic archaeon *Pyrococcus furiosus* at 1.9 Å resolution. *Biochemistry* 36, 9983-9994.

Salminen, T., Teplyakov, A., Kankare, J., Cooperman, B.S., Lahti, R., Goldman, A., 1996. An unusual route to thermostability disclosed by the comparison of *Thermus thermophilus* and *Escherichia coli* inorganic pyrophosphatases. *Protein Science* 5, 1014-1025.

Satheesh, L.S., Murugan, K., 2011. Antimicrobial activity of protease inhibitor from leaves of *Coccinia grandis* (*L*.) *Voigt*.

Savino, C., Federici, L., Johnson, K.A., Vallone, B., Nastopoulos, V., Rossi, M., Pisani, F.M., Tsernoglou, D., 2004. Insights into DNA replication: the crystal structure of DNA polymerase B1 from the archaeon *Sulfolobus solfataricus*. *Structure* 12, 2001-2008.

Scallan, E., Hoekstra, R.M., Angulo, F.J., Tauxe, R.V., Widdowson, M.A., Roy, S.L., Jones, J.L., Griffin, P.M., 2011. Foodborne illness acquired in the United States--major pathogens. *Emerging Infectious Diseases* 17, 7-15.

Scapin, G., 2013. Molecular replacement then and now. *Acta Crystallographica. Section D, Structural Biology* 69, 2266-2275.

Schluter, U., Benchabane, M., Munger, A., Kiggundu, A., Vorster, J., Goulet, M.C., Cloutier, C., Michaud, D., 2010. Recombinant protease inhibitors for herbivore pest control: a multitrophic perspective. *Journal of Experimental Botany* 61, 4169-4183.

Schmitt, M.P., 1999. Identification of a two-component signal transduction system from *Corynebacterium diphtheriae* that activates gene expression in response to the presence of heme and hemoglobin. *Journal of Bacteriology* 181, 5330-5340.

Schwartz, J.H., Reeves, J.Y., Broome, J.D., 1966. Two L-asparaginases from *E. coli* and their action against tumors. *Proceedings of the National Academy of Sciences* 56, 1516-1519.

Scott, A.I., Clemens, K.R., Stolowich, N.J., Santander, P.J., Gonzalez, M.D., Roessner, C.A., 1989. Reconstitution of apo-porphobilinogen deaminase: structural changes induced by cofactor binding. *FEBS Letters* 242, 319-324.

Share, M.R., Roe, R.M., 1988. A partition assay for the simultaneous determination of insect juvenile hormone esterase and epoxide hydrolase activity. *Analytical Biochemistry* 169, 81-88.

Sheldrick, G.M., 2008. A short history of *SHELX*. *Acta Crystallographica. Section A, Fundation and Advances* 64, 112-122.

Sheldrick, G.M., 2010. Experimental phasing with *SHELXC*/*D*/*E*: combining chain tracing with density modification. *Acta Crystallographica. Section D, Structural Biology* 66, 479-485.

Sherwood, D., Cooper, J., 2011. Crystals, X-rays and Proteins: Comprehensive Protein Crystallography, Oxford University Press.

Shingel, K.I., 2004. Current knowledge on biosynthesis, biological activity, and chemical modification of the exopolysaccharide, pullulan. *Carbohydrate Research* 339, 447-460.

Shirato, H., Ogawa, S., Ito, H., Sato, T., Kameyama, A., Narimatsu, H., Xiaofan, Z., Miyamura, T., Wakita, T., Ishii, K., 2008. Noroviruses distinguish between type 1 and type 2 *histo*-blood group antigens for binding. *Journal of Virology* 82, 10756-10767.

Shoolingin-Jordan, P., 1998. Structure and mechanism of enzymes involved in the assembly of the tetrapyrrole macrocycle. *Biochemical Society Transactions* 26, 326-336.

Shoolingin-Jordan, P.M., Warren, M.J., Awan, S.J., 1997. Dipyrromethane cofactor assembly of porphobilinogen deaminase: formation of apoenzyme and preparation of holoenzyme. *Methods in Enzymology* 281, 317-327.

Shrivastava, A., Khan, A.A., Khurshid, M., Kalam, M.A., Jain, S.K., Singhal, P.K., 2016. Recent developments in L-asparaginase discovery and its potential as anticancer agent. *Critical Reviews in Oncology/Hematology* 100, 1-10.

Siddiq, D.M., Koo, H.L., Adachi, J.A., Viola, G.M., 2011. Norovirus gastroenteritis successfully treated with nitazoxanide. *Journal of Infection* 63, 394-397.

Siebenga, J.J., Beersma, M.F., Vennema, H., van Biezen, P., Hartwig, N.J., Koopmans, M., 2008. High prevalence of prolonged norovirus shedding and illness among hospitalized patients: a model for in vivo molecular evolution. *Journal of Infectious Diseases* 198, 994-1001.

Siebenga, J.J., Vennema, H., Zheng, D.-P., Vinjé, J., Lee, B.E., Pang, X.-L., Ho, E.C., Lim, W., Choudekar, A., Broor, S., 2009. Norovirus illness is a global problem: emergence and spread of norovirus GII. 4 variants, 2001–2007. *Journal of Infectious Diseases* 200, 802-812.

Sieciechowicz, K.A., Joy, K.W., Ireland, R.J., 1988. The metabolism of asparagine in plants. *Phytochemistry* 27, 663-671.

Silverman, L.B., Gelber, R.D., Dalton, V.K., Asselin, B.L., Barr, R.D., Clavell, L.A., Hurwitz, C.A., Moghrabi, A., Samson, Y., Schorin, M.A., 2001. Improved outcome for children with acute lymphoblastic leukemia: results of Dana-Farber Consortium Protocol 91-01. *Blood* 97, 1211-1218.

Sim, G., 1959. The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures. *Acta Crystallographica* 12, 813-815.

Singh, R.S., Saini, G.K., Kennedy, J.F., 2008. Pullulan: microbial sources, production and applications. *Carbohydrate Polymers* 73, 515-531.

Skubák, P., Pannu, N.S., 2013a. Automatic protein structure solution from weak X-ray data. *Nature Communications* 4, 2777.

Skubák, P., Pannu, N.S., 2013b. Automatic protein structure solution from weak X-ray data. *Nature Communications* 4.

Sledz, P., Zheng, H., Murzyn, K., Chruszcz, M., Zimmerman, M.D., Chordia, M.D., Joachimiak, A., Minor, W., 2010. New surface contacts formed upon reductive lysine methylation: improving the probability of protein crystallization. *Protein Science* 19, 1395-1404.

Soares, A.L., Guimarães, G.M., Polakiewicz, B., de Moraes Pitombo, R.N., Abrahão-Neto, J., 2002. Effects of polyethylene glycol attachment on physicochemical and biological stability of *E. coli* L-asparaginase. *International Journal of Pharmaceutics* 237, 163-170.

Soding, J., Biegert, A., Lupas, A.N., 2005. The *HHpred* interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* 33, W244-248.

Someya, Y., Takeda, N., Miyamura, T., 2002. Identification of active-site amino acid residues in the Chiba virus 3C-like protease. *Journal of Virology* 76, 5949-5958.

Song, H.K., Suh, S.W., 1998. Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from *Erythrina caffra* and tissue-type plasminogen activator. *Journal of Molecular Biology* 275, 347-363.

Song, P., Ye, L., Fan, J., Li, Y., Zeng, X., Wang, Z., Wang, S., Zhang, G., Yang, P., Cao, Z., 2015. Asparaginase induces apoptosis and cytoprotective autophagy in chronic myeloid leukemia cells. *Oncotarget* 6, 3861.

Sosnovtsev, S.V., Belliot, G., Chang, K.-O., Onwudiwe, O., Green, K.Y., 2005. Feline calicivirus VP2 is essential for the production of infectious virions. *Journal of Virology* 79, 4012-4024.

Srinivasan, A., Giri, A.P., Harsulkar, A.M., Gatehouse, J.A., Gupta, V.S., 2005. A Kunitz trypsin inhibitor from chickpea (*Cicer arietinum* L.) that exerts anti-metabolic effect on podborer (*Helicoverpa armigera*) larvae. *Plant Molecular Biology* 57, 359-374.

Srinivasan, R., Ramachandran, G., 1965. Probability distribution connected with structure amplitudes of two related crystals. V. The effect of errors in the atomic coordinates on the distribution of observed and calculated structure factors. *Acta Crystallographica* 19, 1008-1014.

Stams, W.A., den Boer, M.L., Beverloo, H.B., Meijerink, J.P., Stigter, R.L., van Wering, E.R., Janka-Schaub, G.E., Slater, R., Pieters, R., 2003. Sensitivity to L-asparaginase is not associated with expression levels of asparagine synthetase in t (12; 21)+ pediatric ALL. *Blood* 101, 2743-2747.

Starcevic, D., Dalal, S., Sweasy, J.B., 2004. Is there a link between DNA polymerase beta and cancer? *Cell Cycle* 3, 998-1001.

Subba-Reddy, C.V., Goodfellow, I., Kao, C.C., 2011. VPg-primed RNA synthesis of norovirus RNA-dependent RNA polymerases by using a novel cell-based assay. *Journal of Virology* 85, 13027-13037.

Sweeney, T.R., Roqué-Rosell, N., Birtley, J.R., Leatherbarrow, R.J., Curry, S., 2007. Structural and mutagenic analysis of foot-and-mouth disease virus 3C protease reveals the role of the *β*-ribbon in proteolysis. *Journal of Virology* 81, 115-124.

Swinkels, J., 1985. Sources of starch, its chemistry and physics.

Takeda, Y., Shibahara, S., Hanashiro, I., 2003. Examination of the structure of amylopectin molecules by fluorescent labeling. *Carbohydrate Research* 338, 471-475.

Taylor, D., Cawley, G., Hayward, S., 2013. Classification of domain movements in proteins using dynamic contact graphs. *PloS One* 8, e81224.

Taylor, D., Cawley, G., Hayward, S., 2014. Quantitative method for the assignment of hinge and shear mechanism in protein domain movements. *Bioinformatics* 30, 3189-3196.

Terwilliger, T.C., 2013. Finding non-crystallographic symmetry in density maps of macromolecular structures. *Journal of Structural and Functional Genomics* 14, 91-95.

Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Adams, P.D., Read, R.J., Zwart, P.H., Hung, L.-W., 2008a. Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias. *Acta Crystallographica. Section D, Structural Biology* 64, 515-524.

Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Zwart, P.H., Hung, L.W., Read, R.J., Adams, P.D., 2008b. Iterative model building, structure refinement and density modification with the *PHENIX AutoBuild* wizard. *Acta Crystallographica. Section D, Structural Biology* 64, 61-69.

Terwilliger, T.C., Adams, P.D., Read, R.J., McCoy, A.J., Moriarty, N.W., Grosse-Kunstleve, R.W., Afonine, P.V., Zwart, P.H., Hung, L.W., 2009. Decision-making in structure solution using Bayesian estimates of map quality: the *PHENIX AutoSol* wizard. *Acta Crystallographica. Section D, Structural Biology* 65, 582-601.

Teunis, P.F., Moe, C.L., Liu, P., E Miller, S., Lindesmith, L., Baric, R.S., Le Pendu, J., Calderon, R.L., 2008. Norwalk virus: how infectious is it? *Journal of Medical Virology* 80, 1468-1476.

Thorven, M., Grahn, A., Hedlund, K.-O., Johansson, H., Wahlfrid, C., Larson, G., Svensson, L., 2005. A homozygous nonsense mutation (428G→ A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *Journal of Virology* 79, 15351-15355.

Tickle, I.J., Driessen, H.P., 1996. Molecular replacement using known structural information. *Crystallographic Methods and Protocols*, 173-203.

Tiew, K.-C., He, G., Aravapalli, S., Mandadapu, S.R., Gunnam, M.R., Alliston, K.R., Lushington, G.H., Kim, Y., Chang, K.-O., Groutas, W.C., 2011. Design, synthesis, and evaluation of inhibitors of Norwalk virus 3C protease. *Bioorganic & Medicinal Chemistry Letters* 21, 5315-5319.

Tomar, R., Sharma, P., Srivastava, A., Bansal, S., Kundu, B., 2014. Structural and functional insights into an archaeal L-asparaginase obtained through the linker-less assembly of constituent domains. *Acta Crystallographica. Section D, Structural Biology* 70, 3187-3197.

Trivedi, T.K., Desai, R., Hall, A.J., Patel, M., Parashar, U.D., Lopman, B.A., 2013. Clinical characteristics of norovirus-associated deaths: a systematic literature review. *American Journal of Infection Control* 41, 654-657.

Troeger, H., Loddenkemper, C., Schneider, T., Schreier, E., Epple, H.J., Zeitz, M., Fromm, M., Schulzke, J.D., 2009. Structural and functional changes of the duodenum in human norovirus infection. *Gut* 58, 1070-1077.

Uitdehaag, J.C., Mosi, R., Kalk, K.H., van der Veen, B.A., Dijkhuizen, L., Withers, S.G., Dijkstra, B.W., 1999. X-ray structures along the reaction pathway of cyclodextrin glycosyltransferase elucidate catalysis in the *α*-amylase family. *Nature Structural & Molecular Biology* 6, 432-436.

Vagin, A., Teplyakov, A., 2010. Molecular replacement with *MOLREP*. *Acta Crystallographica. Section D, Structural Biology* 66, 22-25.

Van Der Maarel, M.J., Van Der Veen, B., Uitdehaag, J.C., Leemhuis, H., Dijkhuizen, L., 2002. Properties and applications of starch-converting enzymes of the *α*-amylase family. *Journal of Biotechnology* 94, 137-155.

Vandavasi, V.G., Weiss, K.L., Cooper, J.B., Erskine, P.T., Tomanicek, S.J., Ostermann, A., Schrader, T.E., Ginell, S.L., Coates, L., 2016. Exploring the Mechanism of beta-Lactam Ring Protonation in the Class A beta-lactamase Acylation Mechanism Using Neutron and X-ray Crystallography. *Journal of Medicinal Chemistry* 59, 474-479.

Vega, E., Barclay, L., Gregoricus, N., Shirley, S.H., Lee, D., Vinje, J., 2014. Genotypic and epidemiologic trends of norovirus outbreaks in the United States, 2009 to 2013. *Journal of Clinical Microbiology* 52, 147-155.

Verma, N., Kumar, K., Kaur, G., Anand, S., 2007. L-asparaginase: a promising chemotherapeutic agent. *Critical Reviews in Biotechnology* 27, 45-62.

Villa, P., Corada, M., Bartosek, I., 1986. L-asparaginase effects on inhibition of protein synthesis and lowering of the glutamine content in cultured rat hepatocytes. *Toxicology Letters* 32, 235-241.

Vinjé, J., 2015. Advances in laboratory methods for detection and typing of norovirus. *Journal of Clinical Microbiology* 53, 373-381.

Viswanathan, P., May, J., Uhm, S., Yon, C., Korba, B., 2013. RNA binding by human Norovirus 3C-like proteases inhibits protease activity. *Virology* 438, 20-27.

Vogt, G., Argos, P., 1997. Protein thermal stability: hydrogen bonds or internal packing? *Folding and Design* 2, S40-S46.

Vogt, G., Woell, S., Argos, P., 1997. Protein thermal stability, hydrogen bonds, and ion pairs. *Journal of Molecular Biology* 269, 631-643.

Vongpunsawad, S., Prasad, B.V., Estes, M.K., 2013. Norwalk virus minor capsid protein VP2 associates within the VP1 shell domain. *Journal of Virology* 87, 4818-4825.

Wallace, A.C., Laskowski, R.A., Thornton, J.M., 1995. *LIGPLOT*: a program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering* 8, 127-134.

Wang, J., Sattar, A.K., Wang, C.C., Karam, J.D., Konigsberg, W.H., Steitz, T.A., 1997. Crystal structure of a pol alpha family replication DNA polymerase from bacteriophage RB69. *Cell* 89, 1087-1099.

Wang, L., Elliott, M., Elliott, T., 1999. Conditional stability of the HemA protein (glutamyl-tRNA reductase) regulates heme biosynthesis in *Salmonella typhimurium*. *Journal of Bacteriology* 181, 1211-1219.

Warren, G.L., Petsko, G.A., 1995. Composition analysis of *α*-helices in thermophilic organisms. *Protein Engineering, Design and Selection* 8, 905-913.

Warren, M.J., Jordan, P.M., 1988. Investigation into the nature of substrate binding to the dipyrromethane cofactor of *Escherichia coli* porphobilinogen deaminase. *Biochemistry* 27, 9020-9030.

Warren, M.J., Smith, A.J., 2009. Tetrapyrroles: Birth, Life and Death *Austin: Landes Biosciences*.

Warren, M.J., Gul, S., Aplin, R.T., Scott, A.I., Roessner, C.A., O'Grady, P., Shoolingin-Jordan, P.M., 1995. Evidence for conformational changes in *Escherichia coli* porphobilinogen deaminase during stepwise pyrrole chain elongation monitored by increased reactivity of cysteine-134 to alkylation by N-ethylmaleimide. *Biochemistry* 34, 11288-11295.

Watanabe, K., Hata, Y., Kizaki, H., Katsube, Y., Suzuki, Y., 1997. The refined crystal structure of Bacillus cereus oligo-1, 6-glucosidase at 2.0 Å resolution:

structural characterization of proline-substitution sites for protein thermostabilization. *Journal of Molecular Biology* 269, 142-153.

Waterman, D., Winter, G., Parkhurst, J., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N., Evans, G., 2013. The *DIALS* framework for integration software. *CCP4 Newslett. Protein Crystallography* 49, 13-15.

Wehner, A., Harms, E., Jennings, M.P., Beacham, I.R., Derst, C., Bast, P., Röhm, K.H., 1992. Site‐specific mutagenesis of *Escherichia coli* asparaginase II. *The FEBS Journal* 208, 475-480.

Werner, R., Guitton, M.C., Mühlbach, H.P., 1993. Nucleotide sequence of a cathepsin D inhibitor protein from tomato. *Plant Physiology* 103, 1473.

Widdowson, M.-A., 2005. Norovirus and Foodborne Disease, United States, 1991–2000-Volume 11, Number 1—January 2005-*Emerging Infectious Disease journal*-CDC.

Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R.F., Sykes, B.D., Wishart, D.S., 2003. *VADAR*: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Research* 31, 3316-3319.

Williams, P., Coates, L., Mohammed, F., Gill, R., Erskine, P., Bourgeois, D., Wood, S.P., Anthony, C., Cooper, J.B., 2006. The 1.6 Å X-ray structure of the unusual c-type cytochrome, cytochrome c_L, from the methylotrophic bacterium *Methylobacterium extorquens*. *Journal of Molecular Biology* 357, 151-162.

Wilson, K., Yeates, D., 1979. On the treatment of protein data measured on the oscillation camera. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 35, 146-157.

Winn, M.D., Isupov, M.N., Murshudov, G.N., 2001. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallographica. Section D, Structural Biology* 57, 122-133.

Winn, M.D., Murshudov, G.N., Papiz, M.Z., 2003. Macromolecular TLS refinement in *REFMAC* at moderate resolutions. *Methods in Enzymology* 374, 300-321.

Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, A., Wilson, K.S., 2011. Overview of the CCP4 suite and current developments. *Acta Crystallographica. Section D, Structural Biology* 67, 235-242.

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

Wood, S., Lambert, R., Jordan, P.M., 1995. Molecular basis of acute intermittent porphyria. *Molecular Medicine Today* 1, 232-239.

Woodcock, S.C., Jordan, P.M., 1994. Evidence for participation of aspartate-84 as a catalytic group at the active site of porphobilinogen deaminase obtained by

site-directed mutagenesis of the hemC gene from *Escherichia coli*. *Biochemistry* 33, 2688-2695.

Woodward, J.M., Gkrania-Klotsas, E., Cordero-Ng, A.Y., Aravinthan, A., Bandoh, B.N., Liu, H., Davies, S., Zhang, H., Stevenson, P., Curran, M.D., 2015. The role of chronic norovirus infection in the enteropathy associated with common variable immunodeficiency. *The American Journal of Gastroenterology* 110, 320-327.

Wyatt, G.R., Davey, K.G., 1996. Cellular and molecular actions of juvenile hormone. II. Roles of juvenile hormone in adult insects. *Advances in Insect Physiology* 26, 1-155.

Wyatt, R.G., Dolin, R., Blacklow, N.R., DuPont, H.L., Buscho, R.F., Thornhill, T.S., Kapikian, A.Z., Chanock, R.M., 1974. Comparison of three agents of acute infectious nonbacterial gastroenteritis by cross-challenge in volunteers. *The Journal of infectious diseases* 129, 709-714.

Yao, M., Yasutake, Y., Morita, H., Tanaka, I., 2005. Structure of the type I L-asparaginase from the hyperthermophilic archaeon *Pyrococcus horikoshii* at 2.16 angstroms resolution. *Acta Crystallographica. Section D, Structural Biology* 61, 294-301.

Yip, K., Stillman, T., Britton, K., Artymiuk, P., Baker, P., Sedelnikova, S., Engel, P., Pasquo, A., Chiaraluce, R., Consalvi, V., 1995. The structure of *Pyrococcus furiosus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure* 3, 1147-1158.

Yip, K.S., Britton, K.L., Stillman, T.J., Lebbink, J., de Vos, W.M., Robb, F.T., Vetriani, C., Maeder, D., Rice, D.W., 1998. Insights into the molecular basis of thermal stability from the analysis of ion‐pair networks in the glutamate dehydrogenase family. *The FEBS Journal* 255, 336-346.

Yun, M.-K., Nourse, A., White, S.W., Rock, C.O., Heath, R.J., 2007. Crystal Structure and Allosteric Regulation of the Cytoplasmic *Escherichia coli* L-Asparaginase I. *Journal of Molecular Biology* 369, 794-811.

Zahorsky, J., 1929. Hyperemesis hiemis or the winter vomiting disease. *Archives de Pédiatrie* 46, 391-395.

Zeitler, C.E., Estes, M.K., Prasad, B.V., 2006. X-ray crystallographic structure of the Norwalk virus protease at 1.5-Å resolution. *Journal of Virology* 80, 5050-5058.

Zeng, B.J., Lv, Y., Zhang, L.L., Huang, L.H., Feng, Q.L., 2015. Cloning and structural characterization of juvenile hormone diol kinase in *Spodoptera litura*. *Insect Science*.

Zhang, H., Jin, Z., 2011. Preparation of resistant starch by hydrolysis of maize starch with pullulanase. *Carbohydrate Polymers* 83, 865-867.

Zuber, H., 1988. Temperature adaptation of lactate dehydrogenase Structural, functional and genetic aspects. *Biophysical Chemistry* 29, 171-179.

Zwart, P., Grosse-Kunstleve, R., Adams, P., 2005. *Xtriage* and *Fest*: automatic assessment of X-ray data and substructure structure factor estimation. *CCP4 Newsletter* 43, 27-35.

Appendices

Methods

i DNA transformation

Plasmid DNA (50 ng) is added to 20-50 µl of competent cells. The cells are then stored on ice for 30 min followed by a heat shock at 42 °C for 30 s in a water bath before placing back on ice for another 5 min. 200 µl of Luria Broth (LB, Sigma-Aldrich, Dorset, UK) is then added to each 20 µl cells and the cells are grown at 37 °C, 300 rpm for 1 h in a ThermoMixer (Eppendorf, Stevenage, UK). The culture is then spread on an LB-agar plate containing appropriate antibiotics. The agar plate is incubated at 37 °C overnight.

ii Expression

Expression is undertaken in 2 L flasks containing 500 ml LB and appropriate antibiotics. A 500 µl culture from a 10 ml overnight starter culture is inoculated into each flask and grown at 37 °C, 150 rpm to mid-log stage with an OD_{600} of 0.4 to 0.6 in an Innova 43 shaking incubator (New Brunswick Scientific, New Jersey, US). For a heat shock method, the culture is then grown at 42 °C for 25 min followed by cooling for 5 min in ice water. The cells are induced by 0.5 mM or 1 mM IPTG and the bacterial growth is continued for 40 h at 16 °C, 150 rpm. For a standard method, the cells are induced by 0.5 mM or 1 mM IPTG followed by shaking at 37 °C, 150 rpm overnight. The culture is centrifuged at 14,000 g, 4 °C for 20 min with a Beckman Coulter Avanti J-26 XP ultracentrifuge (Beckman Coulter, Buckinghamshire, UK) and the cell-pellet obtained is then re-suspended and sonicated in a buffer containing 50 mM Tris, 100 mM NaCl, pH 7.3 using a Soniprep 150 sonicator (MSE, London, UK) to lyse the cells. The lysate is then

centrifuged at 55,000 g, 4 °C for 40 min and the supernatant is stored on ice before purification. The expression is checked by SDS-PAGE using the non-induced culture as a control.

iii Nickel affinity chromatography

Each supernatant solution which has been confirmed to contain the desired protein by SDS-PAGE is passed through a 1 ml His-trap FF nickel affinity column (GE Healthcare, Buckinghamshire, UK) to purify the His-tagged protein. The purification is achieved using the following buffers:

Binding buffer: 50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH 8.0
Wash buffer: 50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, pH 8.0
Elution buffer: 50 mM NaH_2PO_4, 300 mM NaCl, 500 mM imidazole, pH 8.0

For a manual run, buffers and supernatant are loaded on the column through a 0.45 μm filter using a 10 ml syringe with a flow rate of approximately 1 ml/min. The column is equilibrated with 10 ml of the binding buffer followed by sequential introduction of the supernatant and 10 ml of the wash buffer to bind the His-tagged protein and to remove the non-specifically bound substances. Finally, 5 ml of the elution buffer is used to elute the desired protein from the column. For an automatic run, pre-filtered degassed buffers and supernatant are loaded by using an ÄKTA Start FPLC system (GE Healthcare, Buckinghamshire, UK) and the process is monitored by UV absorbance at 280 nm. All the flow-throughs are collected for SDS-PAGE analysis. The protein concentration in the eluted sample is determined using a Nanodrop spectrophotometer (Labtech International Ltd., East Sussex, UK) at 280 nm with the calibration of the extinction coefficient. The His-tag is removed by the addition of an appropriate protease, e.g. thrombin

(Merck, Darmstadt, Germany) followed by overnight dialysis in a GeBaflex-tube dialysis kit (Gene Bio-Application Ltd., Slough, UK) in 50 mM Tris, 100 mM NaCl pH 7.3 at room temperature/4 °C to remove the cleaved tag and imidazole.

iv HiTrap Benzamidine affinity chromatography

Thrombin is removed by passing the protein solution through a 1 ml HiTrap Benzamidine FF column (GE Healthcare, Buckinghamshire, UK). The purification is carried out using the following buffers:

Binding buffer: 20 mM Na_3PO_4, 150 mM NaCl, pH 7.5
High salt wash buffer: 20 mM Na_3PO_4, 1.0 M NaCl, pH 7.5
Benzamidine elution buffer: 20 mM *p*-aminobenzamidine in binding buffer

Buffers and the protein sample are loaded on the column through a 0.45 μm filter using a 10 ml syringe with a flow rate of approximately 1 ml/min. The column is equilibrated with 10 ml of the binding buffer followed by sequential introduction of the protein sample and 10 ml of the binding buffer to bind the thrombin and to remove the non-specifically bound substances. 5 ml of the high salt wash buffer is applied to elute the protein from the column because it forms ionic interaction with the column. Finally, 5 ml of the elution buffer is used to elute the thrombin from the column. All the flow-throughs are collected for SDS-PAGE analysis.

v Gel-filtration chromatography

The eluted sample is concentrated in a 20 ml, Vivaspin centrifugal concentrator (GE Healthcare, Buckinghamshire, UK) by use of an Allegra X-12R centrifuge (Beckman Coulter, California, USA) at 3250 g and 4 °C. The column (e.g. Superdex 75) is equilibrated firstly with 3 bed-volume of degassed appropriate

buffer with a flow rate of 1 ml/min. The sample is then loaded onto the column followed by running with the same buffer and flow rate. The separation of the protein is monitored by UV absorbance at 280 nm. The fractions are analysed by SDS-PAGE.

Figures

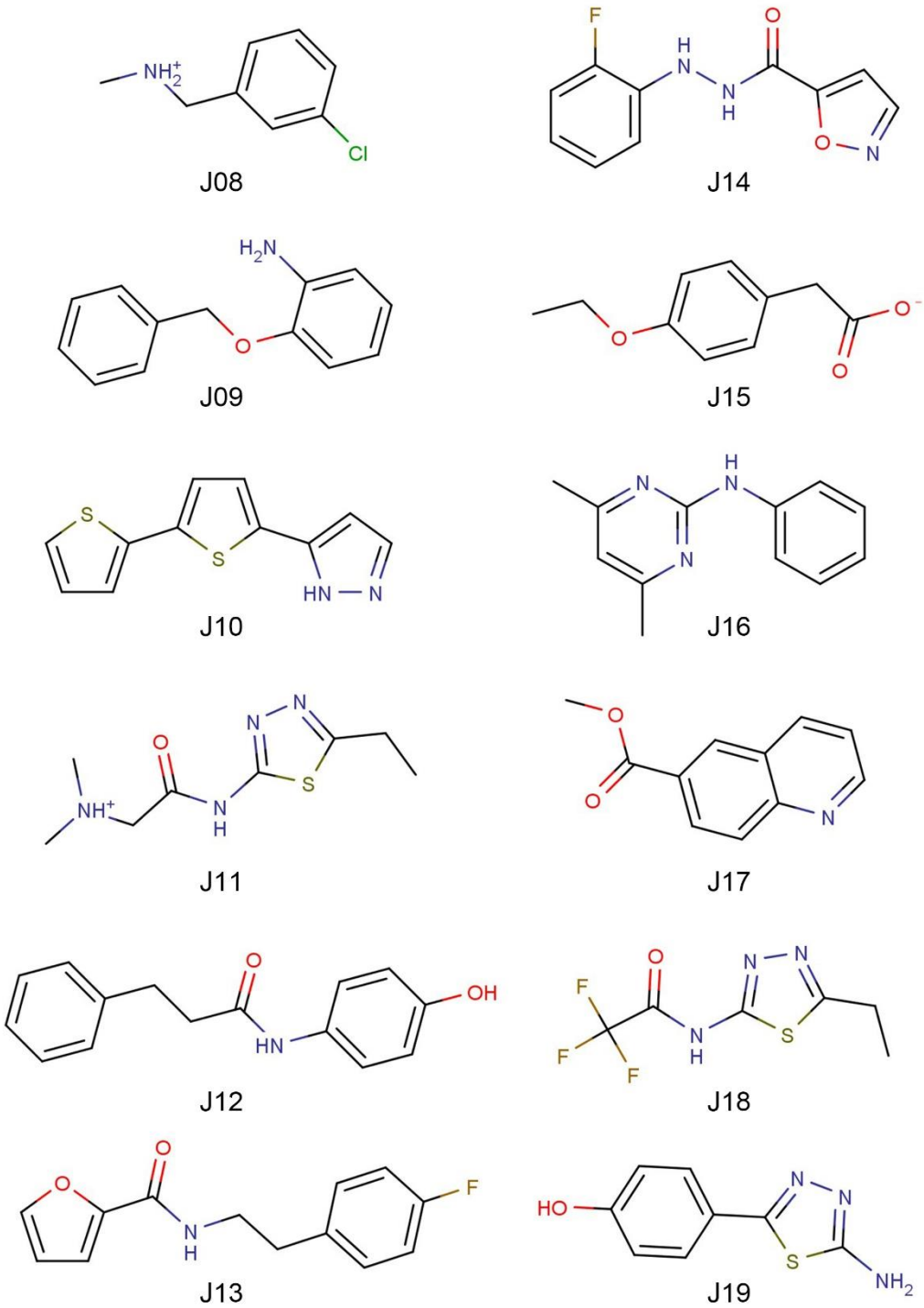


Figure A 2D structure of fragments J08 to J19.

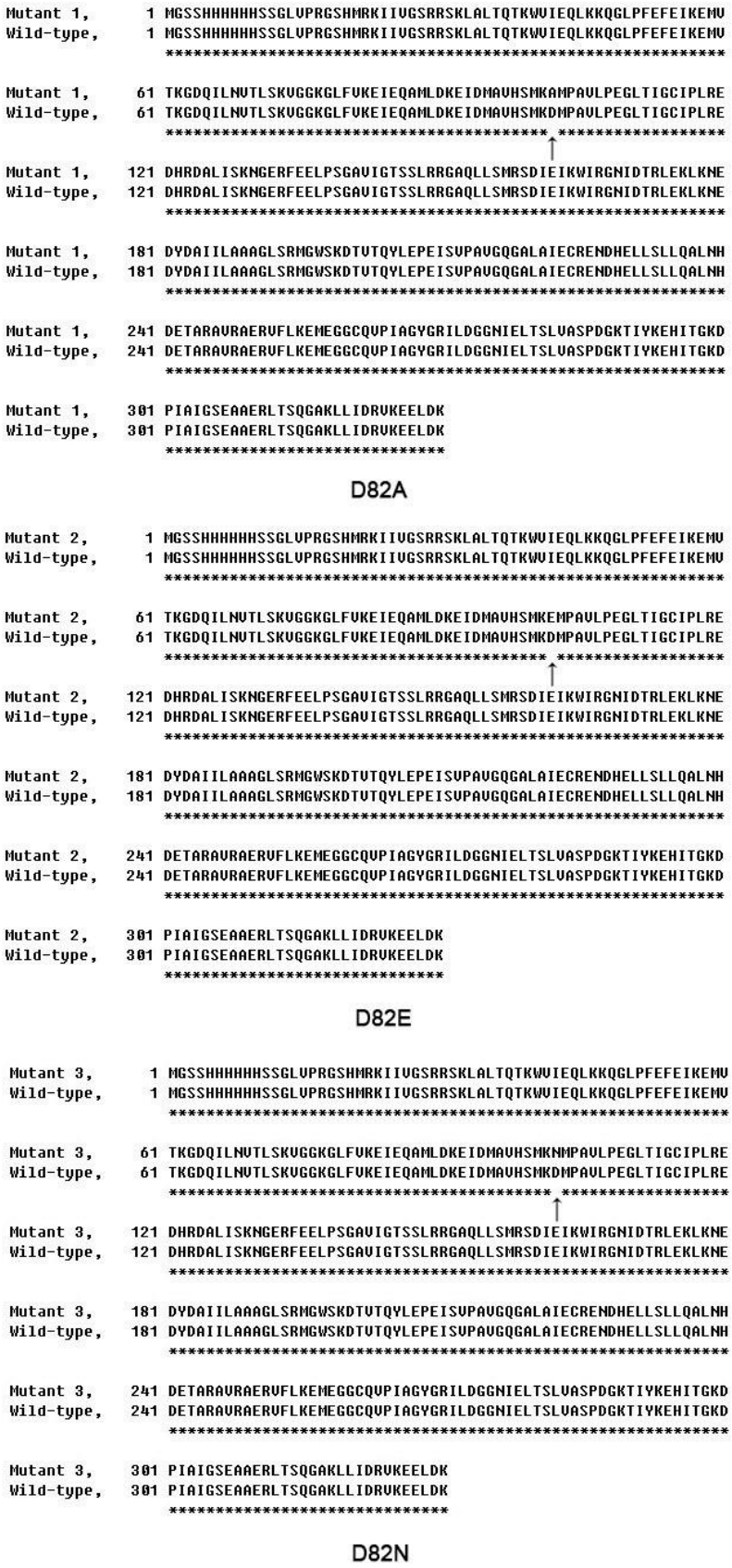


Figure B Comparison of amino acid sequences translated from DNA sequencing result between the WT and mutant BPBGDs. The mutated amino acids are indicated by the arrows.